UNIVERSIDAD CARLOS III DE MADRID

PROTECCIÓN DE LA PRIVACIDAD EN LA GESTIÓN DE BIG DATA

UNA VISIÓN MULTIDIMENSIONAL: TECNOLÓGICA Y NORMATIVA

PROYECTO FIN DE CARRERA

ALBERTO QUIÑONES LAYOS

INGENIERÍA TÉCNICA DE TELECOMUNICACIÓN: Especialidad SISTEMAS DE TELECOMUNICACIÓN

Versión 1.0 – 15 de octubre 2015

PÁGINA INTENCIONADAMENTE EN BLANCO

Proyecto Fin de Carrera

Título: **Protección de la Privacidad en la Gestión de Big Data. Una visión multidimensional: tecnológica y normativa**

Autor: **Alberto Quiñones Layos**

Tutor: **Harold Molina-Bulla. Departamento de Teoría de la Señal y Comunicaciones**

Titulación: **Ingeniería Técnica de Telecomunicación. Especialidad Sistemas de Telecomunicación**

SINOPSIS, TÉRMINOS DE REFERENCIA Y ESTRUCTURA DEL PROYECTO

Sinopsis

El presente proyecto presenta un análisis de la situación actual en relación con la gestión de la privacidad y la protección de datos de carácter personal asociada al empleo de Big Data. Se cubre de esta manera uno de los retos fundamentales a abordar para asegurar que los Big Data permitan a las distintas organizaciones la prestación de servicios más eficientes, sostenibles y personalizados que redunden en un mayor beneficio para la sociedad.

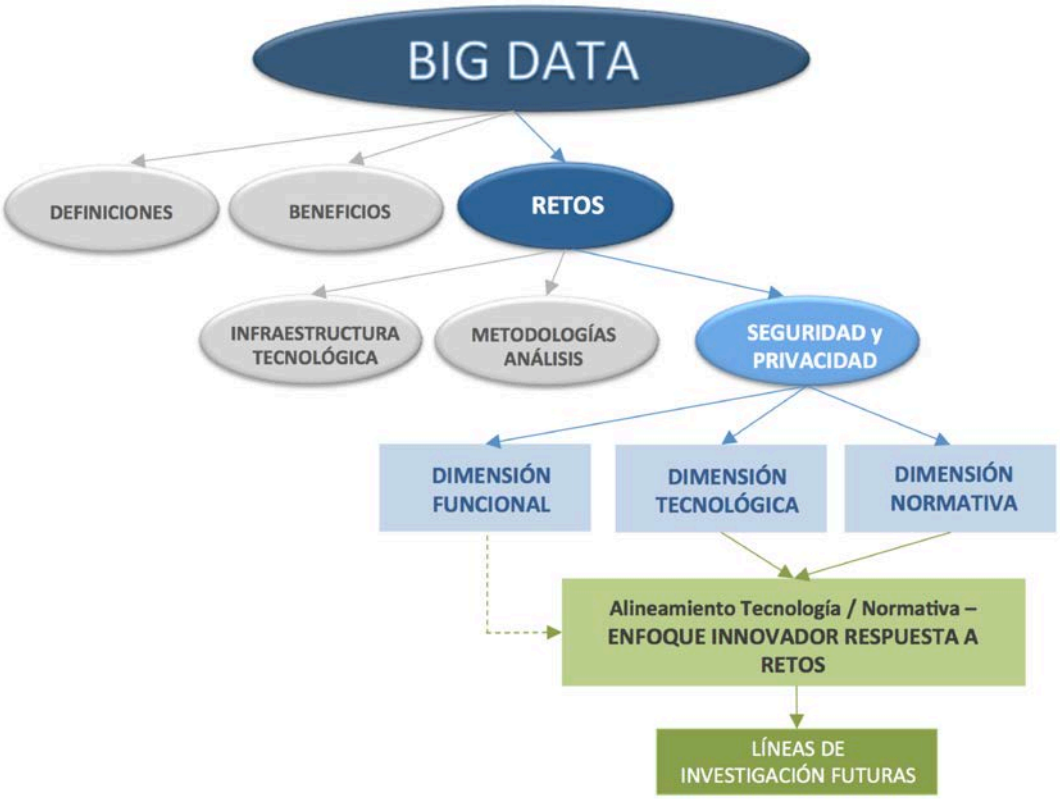
Se propone para ello un enfoque holístico que, partiendo del análisis de la situación actual en relación con las soluciones tecnológicas y metodológicas que para reforzar la seguridad de la información y la privacidad se emplean en la gestión de Big Data, y de la normativa en vigor o en desarrollo en relación con la protección de datos de carácter personal, combine los elementos de ambas en aras de una mayor protección de los datos de carácter personal de los ciudadanos al mismo tiempo que se asegura la máxima eficiencia y entrega de valor por parte de los análisis de Big Data.

Términos de referencia

Estructura del proyecto

El proyecto se estructura de la siguiente manera:

- En el *primer capítulo* se incluye una descripción de qué son los Big Data, cuál es el motivo de su relevancia en el momento actual, así como de cuáles son los retos principales que se derivan de su empleo, haciendo especial énfasis en los retos derivados de la protección de la privacidad.
- En el *segundo capítulo* se presenta el estado del arte en relación con la privacidad y la seguridad de la información en Big Data desde un punto de vista tecnológico.
- El *tercer capítulo* pone el foco en la dimensión normativa de la protección de datos de carácter personal, indicándose cómo se ven afectadas las disposiciones vigentes por la proliferación de los Big Data, y qué respuestas se dan a los nuevos retos que suponen éstos.
- Una vez analizado el estado actual, se presentan en el *cuarto capítulo* posibles soluciones y enfoques que podrían permitir, por un lado superar las carencias que las tecnologías y métodos que se emplean actualmente pueden tener, y por otro alinear convenientemente tecnología y normativa para un eficiente explotación de los Big Data y las posibilidades que ofrecen junto a una protección de los datos personales.
- El proyecto se cierra finalmente con un resumen de las *conclusiones* obtenidas y una serie de líneas generales sobre las que puede asentarse la investigación futura en esta materia.



ÍNDICE GLOBAL

ÍNDICE DE IMÁGENES

INTRODUCCIÓN

OBJETIVO DEL PROYECTO

1. Objetivo Principal del Proyecto

Presentar un marco de referencia global para la Protección de la Información de Carácter Personal en el ámbito de la Gestión de los Big Data, a partir de la cual **maximizar el valor que aportan los Big Data al mismo tiempo que se asegura un adecuado tratamiento de la Privacidad** (alineado con los Objetivos Estratégicos y que cumpla con la Normativa vigente), combinando elementos de la dimensión tecnológica y normativa.

Se articula de este modo la Protección de la Privacidad como una ventaja competitiva de la Analítica y la Gestión de Big Data

2. Objetivos Parciales del Proyecto

Para la consecución de este **Objetivo Principal**, se abordan los siguientes **Objetivos Parciales**:

1. Presentar una aproximación a los Big Data, el porqué de su relevancia, los beneficios que pueden aportar y los más importantes retos asociados con su gestión. En cuanto a éstos últimos la atención se centra en los relativos a la protección de la privacidad.

2. Analizar el estado del arte en relación con la Protección de la Información en las Soluciones de Big Data existentes en el mercado, desde un punto de vista tecnológico (mecanismos de autenticación y control de acceso y técnicas potenciadoras de la privacidad como la anonimización). Para cada caso se abordan sus beneficios así como los puntos de mejora en relación con ellos.

3. Conocer la situación existente en relación a la Normativa que en distintos países y Organizaciones supranacionales regula la Protección de Datos de Carácter Personal, centrándose en la manera en que la misma aborda los retos y nuevas casuísticas que plantea la gestión de Big Data (bien directa o indirectamente).

4. Presentar y evaluar una propuesta de solución que podría permitir una más óptima protección de la privacidad, al tiempo que se mantiene en la mayor medida posible las ventajas y posibilidades que ofrecen los Big Data, combinando elementos de las dimensiones tecnológica y normativa, e intentando afrontar los retos existentes y las carencias identificadas.

5. Identificar líneas de investigación futuras que, en línea con la evolución de la tecnología, permitan una aproximación proactiva (en lugar de reactiva) y permanente respecto a la Protección de la Privacidad.

CAPÍTULO I

LOS BIG DATA Y LA PRIVACIDAD COMO RETO FUNDAMENTAL PARA ASEGURAR EL VALOR DERIVADO DE SU GESTIÓN

1. Introducción y Objeto del Capítulo

En este capítulo se presenta una primera aproximación a los Big Data: ¿qué son?, ¿qué oportunidades suponen?, ¿qué riegos se derivan de su empleo?, etc. Este término se ha convertido en uno de los más comunes hoy en día en relación con la evolución de las Tecnologías de la Información y las Comunicaciones.

Los datos son la materia prima más relevante para asentar el desarrollo económico y comercial por parte de múltiples empresas y organizaciones privadas. Por su parte las Administraciones Públicas también trabajan con los datos para ofrecer servicios más eficientes, más personalizados y de mayor calidad. Por este motivo el análisis y gestión de Big Data es una tendencia fundamental para estas empresas, organizaciones y Administraciones Públicas.

Además de una descripción de qué son los Big Data y el porqué de su relevancia, el apartado se centra en presentar los principales **retos asociados con su gestión** y en especial la **protección de la privacidad**.

1.1. Esquema de estructura del Capítulo

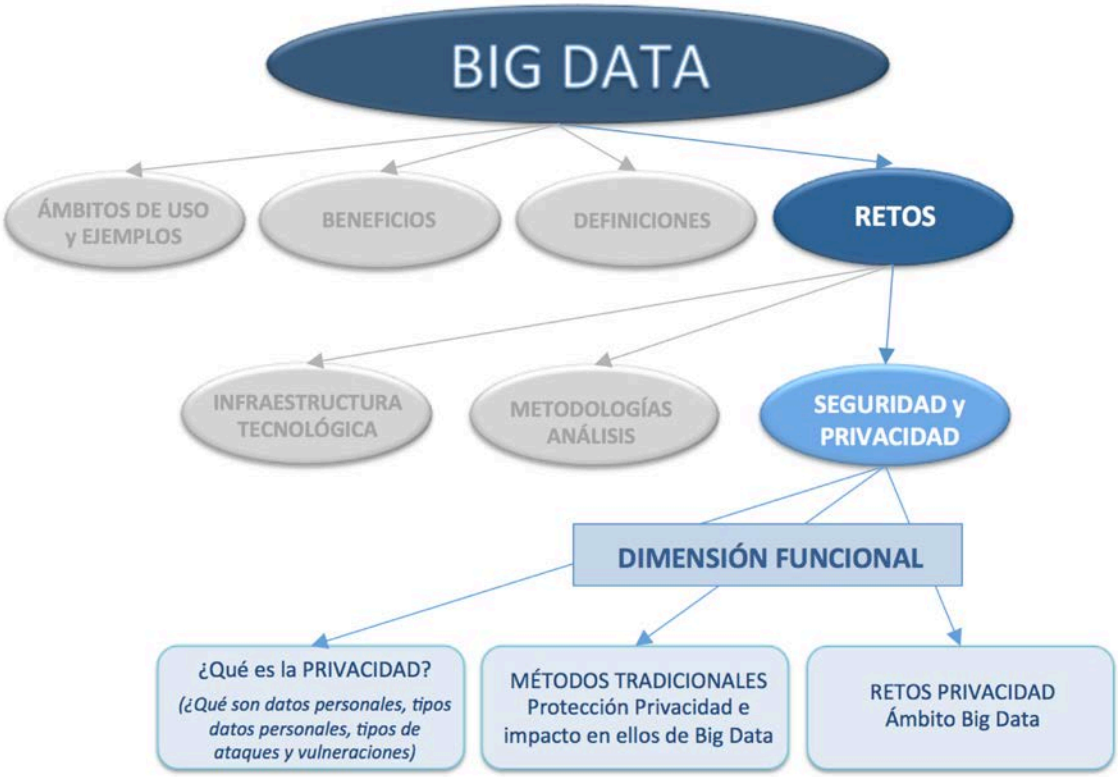


Imagen I-1 – Esquema Estructura Capítulo I

2. ¿Qué son los Big Data y por qué son tan relevantes?

2.1. Ámbitos de uso, ejemplos y beneficios

Los Big Data detrás del éxito de la Selección de Fútbol de Alemania en el pasado Mundial

Durante el pasado Mundial de Fútbol de Brasil 2014, titulares como el anterior pudieron leerse en la prensa. Aludía al hecho de que la selección de Alemania, campeona del torneo, empleó durante su preparación una solución de gestión de Big Data de la empresa SAP para el análisis de diferentes datos y variantes del juego, y hacer así más efectivo el entrenamiento y el rendimiento durante los partidos.

De este modo los preparadores germanos pudieron conocer en tiempo real todo tipo de datos en relación con la posesión del balón, pases, tiros a puerta, aceleración o velocidad de carrera, obtenidos gracias a sensores que los jugadores llevaban colocados en su cuerpo. Se estima que en un entrenamiento y en unos 10 minutos, 10 jugadores con tres balones pueden generar del orden de unos 7 millones de datos, lo que da muestra de la dimensión de información que puede obtenerse y de la necesidad de una tecnología de análisis, gestión y presentación de resultados muy potente.

(Fuente:[53])

Hoy en día noticias como éstas son cada vez más comunes, y el término Big Data se ha instalado dentro del conocimiento colectivo como una expresión popular que promete cambiar o al menos afectar a la forma en que hacemos prácticamente cualquier actividad de la vida cotidiana de personas, empresas y organizaciones, dado su potencial para analizar enormes volúmenes de datos para identificar tendencias y tomar decisiones en tiempo real.

…pero, ¿qué son exactamente los Big Data?

Se trata de una tendencia tecnológica en la ciencia, la industria, los negocios o la gestión pública, que afecta a la mayor parte de los aspectos de la actividad humana, combinándose con otras tecnologías como Cloud Computing, el Internet of Things o la Analítica Predictiva. Se basa básicamente en el enorme incremento en la generación de datos en el mundo y permite aplicar nuevas formas de gestión específicas y especializadas (captura de los mismos, almacenamiento, búsqueda, compartición, análisis y visualización) generando valor a partir de su análisis.

El incremento en el volumen de datos generados y en la velocidad y variedad en que los mismos se presentan, no responde únicamente al incremento de la actividad en Internet de las personas dentro del paradigma de la Web 2.0 y a la aparición de las redes sociales, en que los usuarios han dejado de ser meros consumidores de contenidos para pasar a ser generadores de los mismos, sino que además se ha visto incrementado exponencialmente gracias a la conexión a redes de comunicaciones de miles de dispositivos y sensores, que de manera autónoma generan una enorme cantidad de datos de su entorno (sirva como referencia de la dimensión de este volumen de datos que cada motor de un avión en un vuelo de Londres a Nueva York genera 10 Terabytes de datos cada 30 minutos (*fuente: [106]*)). Es el denominado Internet of Things, que permite que por ejemplo un smartphone a través de su localización GPS, pueda dar a conocer información relevante que implique datos personales, bien de manera individual o en combinación con otros datos que el usuario comparta de manera consciente.

En relación con el citado incremento en la generación de datos, el siguiente gráfico da una muestra de la dimensión del volumen de datos que se produce actualmente en Internet, y que constituyen la materia prima fundamental de los Big Data, así como algunas estimaciones de futuro (*fuente* [18], [32], [50], [82], [101]).

Imagen I-2 – Ejemplos volumen generación de datos

En lo que respecta a los ámbitos de actividad y campos de conocimiento donde la gestión de Big Data está aportando beneficios, el ejemplo con que se iniciaba este capítulo da muestra de la multiplicidad y variedad existente. A modo de resumen, los siguientes constituyen los ejemplos de mayor relevancia (tal y como se destacan en [64]):

- **Entender los comportamientos y preferencias de las personas, para la presentación de Servicios y Productos Personalizados.**

  Por ejemplo las compañías aseguradoras de vehículos pueden conocer con mayor exactitud la forma en que sus clientes actuales y potenciales conducen, y adaptar sus ofertas en consecuencia.

- **Entender y optimizar los procesos de negocio internos de las Empresas y Organizaciones.**

  Por ejemplo las cadenas distribuidoras pueden adaptar sus stocks sobre la base de las previsiones de demanda de los clientes, conocidas gracias a información publicada en redes sociales o a búsquedas en Internet.

- **Aplicaciones personales.**

  No sólo las empresas pueden hacer uso de las ventajas que ofrecen los Big Data, también las personas individualmente, por ejemplo a través de los datos generados por dispositivos

wearable que miden y monitorizan su actividad física, la analizan y recomiendan patrones de actividad, descanso, alimentación, etc., contribuyendo en definitiva a una mejora en la condición física y la salud.

- **Mejoras en los Servicios Públicos – Sanidad, Gestión del Tráfico, etc.**

  Anticipar epidemias, incrementos en la demanda de servicios sanitarios o en la congestión en carreteras, y poder adaptar la respuesta y la asignación de recursos a estos acontecimientos, son aspectos posibles gracias a los Big Data que redundan en un mejor aprovechamiento de los recursos disponibles y en un servicio más eficiente para los ciudadanos.

  Por ejemplo, en el ámbito de la Sanidad, cabe destacar que monitorizando y analizando cada latido y respiración de un bebé, es posible desarrollar algoritmos que pueden predecir infecciones antes de que cualquier síntoma aparezca (*Proyecto Artemis [77]*). En la gestión del tráfico en las ciudades puede reseñarse el caso de empleo de señales de tráfico inteligentes adaptativas para prevenir atascos, que se está empleando por ejemplo en el estado de Utah en EE.UU. para la gestión de un elevado porcentaje de los semáforos en dicho estado, con resultados positivos en términos de reducción de atascos (un 40% menos) y de accidentes de tráfico por choques en intersecciones (un 50% menos) (*[75]*).

- **Mejoras en la Ciencia y la Investigación.**

  Investigaciones científicas de gran relevancia que implican la recogida y análisis de enormes cantidades de datos en tiempo real, serían imposibles sin las técnicas de gestión de Big Data. Así por ejemplo el Gran Colisionador de Hadrones del CERN de Ginebra, cuenta con un centro de proceso de datos con 65.000 procesadores que son capaces de analizar hasta 30 Petabytes de datos.

  Los Big Data también permiten optimizar el funcionamiento de equipos y dispositivos, haciéndolo más eficientes y autónomos. Puede destacarse en este sentido algunas redes de distribución de energía que hacen un análisis de las condiciones del entorno mediante sensores inteligentes para adaptar la oferta a ellas, o también dispositivos con un funcionamiento autónomo como el prototipo de coche sin conductor de Google ([44]).

- **Mejora de la Seguridad y el cumplimento de la Ley.**

  Los Big Data, gracias a la identificación de patrones de actividad y el análisis y correlación de información, permiten contribuir a la prevención de ciberataques, atentados u otros delitos. Las compañías bancarias, por ejemplo, pueden identificar transacciones fraudulentas a partir del análisis y comparación de cada movimiento con los patrones conocidos del uso que habitualmente hace el titular de una cuenta o de una tarjeta de crédito.

2.2. Definiciones de Big Data

A continuación se presentan algunas definiciones del concepto Big Data de empresas fabricantes, consultoras y de análisis de mercado más importantes en el ámbito de las Tecnologías de la Información y las Comunicaciones.

- **Gartner**: Big Data se define en general como una serie de recursos de información de gran volumen, velocidad de generación / actualización y variedad, que requieren métodos y formas innovadoras y rentables de procesado para un conocimiento y toma de decisiones mejoradas *[41] y [88]*.

- **Forrester**: Big Data se establece en el límite de la habilidad de una Organización para almacenar, procesar y acceder a todos los datos que necesita para operar eficazmente, tomar decisiones, reducir riesgos y ofrecer servicios a sus clientes o usuarios *[47]*.
- **IDC**: Las tecnologías de Big Data describen una nueva generación de tecnologías y arquitecturas, diseñadas para de un modo económico extraer valor de volúmenes muy grandes de información de una gran variedad, al permitir una elevada velocidad de captura, descubrimiento y / o análisis *[40]*.
- **McKinsey Global Institute**: Big Data hace referencia a conjuntos de datos cuyo tamaño está por encima de las capacidades de las herramientas típicas de gestión de bases de datos de captura, almacenamiento y análisis. Esta definición es intencionadamente subjetiva y no concreta cómo de grande debe ser un conjunto de datos para ser considerado como Big Data; por ejemplo no se define Big Data en términos de ser mayor que una cantidad determinada de Terabytes. Se asume que, conforme la tecnología progrese a lo largo del tiempo el tamaño de los conjuntos de datos que se consideren Big Data también crecerá.

  Cabe mencionar también que la definición puede variar también según el sector que la vaya a emplear, dado que en cada uno de ellos puede ser diferente el tipo de herramientas de software disponibles de manera habitual y los volúmenes de datos que manejarán comúnmente. Con estas consideraciones, lo que se conoce como Big Data en muchos sectores todavía se mueve en un rango amplio que va desde unas pocas docenas de Terabytes hasta varios Petabytes *[63]*.
- **O'Reilly**: Big Data son aquellos datos que exceden la capacidad de procesamiento de los sistemas de gestión de bases de datos convencionales. Los datos son muy grandes, se generan muy rápido o no encajan en las estructuras de las arquitecturas de bases de datos disponibles habitualmente. Para poder obtener valor de estos datos, se debe optar por emplear formas de procesado alternativas *[34]*.
- **Microsoft**: Es el término que describe el proceso mediante el cual se aplica una potencia de procesado importante (teniendo en cuenta las capacidades que ofrecen actualmente el aprendizaje máquina y la inteligencia artificial) a conjuntos de información masivos y frecuentemente de gran complejidad *[66]*.
- **Oracle**: Los Big Data son aquellos datos caracterizados por 4 atributos clave: volumen, variedad, velocidad y valor *[74]*.

Si bien es cierto que no existe una definición oficial y consistente de qué son los Big Data, a la vista de las definiciones anteriores se puede concluir que hay un cierto consenso y convergencia sobre **cuáles son las características más importantes de los Big Data**. La siguiente definición (que toma como referencia la incluida en *[27]*) incluye estas características:

*"Las tecnologías y técnicas de Gestión de Big Data permiten el procesamiento de **grandes volúmenes** de datos, estructurados y no estructurados, generados a **gran velocidad** y de una **gran variedad**, para extraer **valor** de los mismos y asegurar una **elevada veracidad** entre la información obtenida y los datos originales, así como el mantenimiento de las diferentes dimensiones de la **Seguridad de la Información** relacionadas con dicha información (confidencialidad, integridad y disponibilidad). Para ello son necesarias **formas innovadoras y eficientes** en términos de consumo de recursos (económicos y de infraestructuras que emplean) **de análisis y procesado de datos e información** para una **mejora en los procesos de percepción, toma de decisiones y control**. Todo ello requiere y debe sustentarse en **nuevos modelos de datos** que den soporte a **todas las etapas del ciclo de***

vida de los datos *y* ***nuevas infraestructuras***, ***servicios***, ***herramientas*** *y* ***aplicaciones*** *que posibiliten la obtención y el procesado de datos de múltiples y variados orígenes."*

En definitiva, los Big Data y los métodos de análisis y gestión que conllevan, permiten, dentro del contexto ya referido de aumento en el volumen y velocidad en la generación de datos, apuntalar el avance en la **cadena de gestión del conocimiento**, transformando los datos progresivamente en información, conocimiento y sabiduría, con el consiguiente incremento en el **valor** que se aporta a la organización que acomete el análisis.

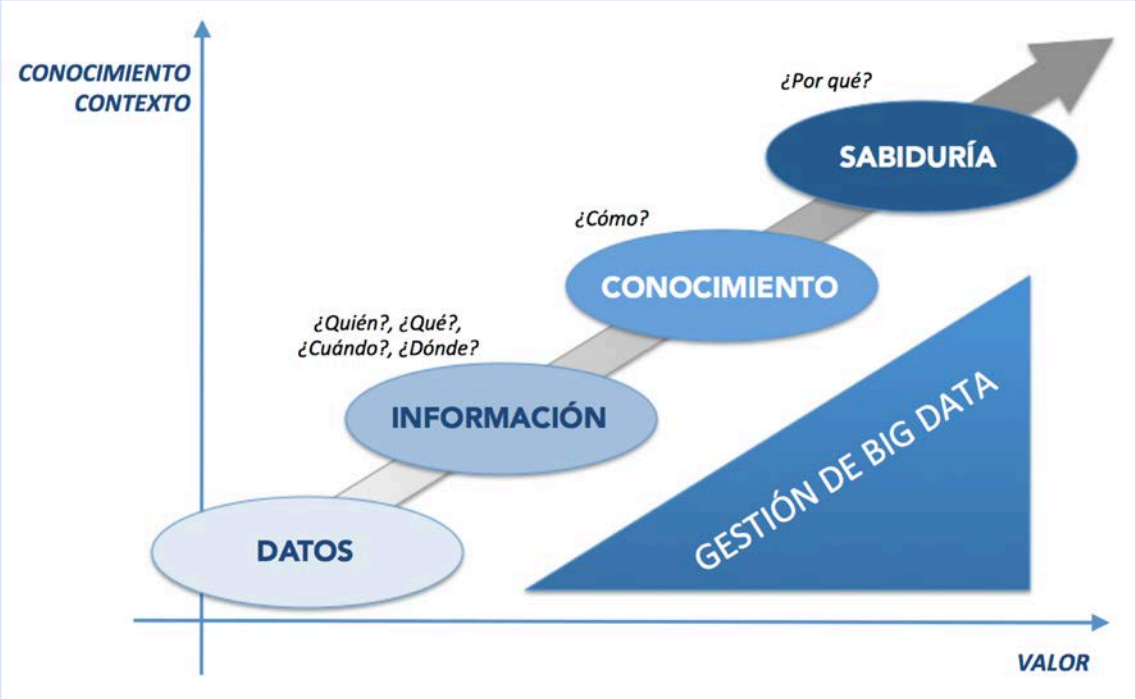


Imagen I-3 – Cadena de Gestión del Conocimiento
(Elaboración propia adaptada de la referencia [9])

Como conclusión, cabe destacar que los Big Data se están posicionando como uno de los **elementos fundamentales para la transformación de la sociedad y la creación de valor** en los próximos años, dada su importancia y su capacidad para influir en el incremento de la eficiencia de las Organizaciones (públicas y privadas) para el desarrollo de sus cometidos.

El análisis y la explotación de Big Data permite a las Organizaciones optimizar factores como:

- La toma de decisiones, al basarlas en un mayor número de fuentes de información y en la combinación de éstas.
- La posibilidad de ofrecer servicios más personalizados a los usuarios, lo que incrementa el valor percibido.

3. Los Retos asociados a la Gestión de Big Data

Lógicamente, la complejidad inherente al entorno de gestión de los Big Data implica una serie de retos que deben ser convenientemente abordados y tratados para permitir que las ventajas aludidas en el apartado anterior puedan ser aprovechadas en todas su extensión, y que los Big Data realmente puedan servir de herramienta fundamental para afianzar el crecimiento

económico, ofreciendo nuevos servicios de calidad, seguros, sostenibles y que protejan los intereses de los ciudadanos.

Google Flu Trends es una herramienta que tiene como objetivo la predicción de epidemias de gripe, antes incluso de que sean detectadas por los centros médicos de control y prevención (*[43]*). Para ello desarrolla una agregación de datos que asocia con las búsquedas en Internet relacionadas con la gripe con una futura e hipotética aparición de esta enfermedad. La aplicación práctica ha demostrado que los resultados que se obtenían no eran 100% conformes a la realidad. (*[57] y [58]*).

Este ejemplo da muestra de que a pesar de un punto de partida teórico correcto, no es suficiente con recopilar ingentes cantidades de datos (como pueden hacer empresas como Google derivadas de los miles de millones de búsquedas que gestionan a diario), sino que los análisis que desarrollan los Big Data deben contar con unas infraestructuras y unos criterios que permitan que el valor pretendido finalmente pueda proveerse (en el caso de Google Flu Trends, una mayor correlación entre búsquedas de síntomas y casos reales de gripe).

Los distintos factores que constituyen retos para asegurar que es posible obtener de los Big Data todo su potencial en unas condiciones de seguridad, protección de la información o sostenibilidad adecuadas, se pueden organizar según la siguiente estructura:

- **Infraestructura Tecnológica**: Necesidad de desarrollo e implantación de nuevos modelos de BBDD, de almacenamiento, etc.
- **Metodologías de Análisis**: Análisis en tiempo real y de correlación de eventos y datos.
- **Seguridad de la Información y Privacidad**: El resultado de un ataque que afecte a algún aspecto de la Seguridad de la Información tendría una dimensión mucho mayor que en un caso de empleo de datos tradicional, dado el volumen de información que podría verse comprometida, así como la posibilidad de emplear la información filtrada para obtener nueva información. Todo ello podría conllevar importantes repercusiones legales así como una disminución de la reputación de la organización que ha sufrido el ataque a su seguridad. Por todo ello es fundamental abordar una serie de retos en diferentes ámbitos:
    - Retos de Infraestructura de Seguridad: computación segura en entornos de programación distribuidos, buenas prácticas para bases de datos no relacionales, etc.
    - Retos de Gestión de Datos: almacenamiento seguro y control de transacciones (logs).
    - Privacidad: técnicas de autenticación y control de acceso, cifrado, analítica orientada al aseguramiento de la privacidad, anonimización, etc. Se trata, tal y como ya se ha apuntado, del reto fundamental que se abordará en este proyecto.

4. La privacidad como reto fundamental para asegurar el valor de los Big Data

A partir de este punto, el análisis se va a centrar en uno de los factores destacados en el apartado anterior: la Privacidad y la Protección de Datos de Carácter Personal, así como en algunos aspectos de la Seguridad de la Información que impactan sobre la privacidad.

4.1. ¿Qué es la Privacidad y por qué es crítica protegerla?

4.1.1. Introducción

Alan Westin, profesor de la Universidad de Columbia y uno de los pioneros y figuras más destacadas en el análisis de la Protección de Datos y la Privacidad, en su estudio *Privacy and Freedom* publicado en 1967, definió la Privacidad como "*la reivindicación de individuos, grupos o instituciones de determinar por sí mismas, cuándo, cómo y en qué medida, la información sobre ellos es comunicada a otros.*"

Los casos de revelación de datos de carácter personal gracias al empleo de técnicas habilitadas por los Big Data, ha sido una de las mayores fuentes de preocupación y uno de los mayores retos en relación con la gestión de datos en el escenario actual. Como algunos de los casos de mayor trascendencia cabe mencionar:

- **El caso Netflix**

  La aplicación de alquiler online de películas y series Netflix, incluye un registro con la valoración anónima que más de 500.000 de sus usuarios han hecho de diferentes películas. Se ha demostrado en *[71]* que con un conocimiento acerca de alguna característica personal de un usuario de Netflix, es posible identificar qué valoraciones de películas han sido escritos por él/ella.

  Así por ejemplo usando el Internet Movie Database (IMDb) como fuente de conocimiento adicional y público, es posible descubrir quién escribió una valoración, al existir una correlación entre las calificaciones públicas de IMDb y las privadas de Netflix. De este modo era posible revelar aspectos asociados a estas últimas, como preferencias políticas u otra información potencialmente sensible.

- **La reidentificación del Gobernador de Massachusetts**

  Se trata de uno de los casos más conocidos en relación con la protección de la privacidad.

  A mediados de los años 90 del siglo XX el estado de Massachusetts en EE.UU. decidió publicar listas de datos anonimizadas de los empleados estatales, incluyendo información sobre su salud (enfermedades, ingresos hospitalarios, etc.), con el objetivo de servir de ayuda a investigadores en sus estudios. La anonimización que se siguió eliminó los identificadores directos obvios como el nombre, la dirección y el número de la Seguridad Social, y el propio Gobernador de Massachussets, que en aquel momento era William Weld, aseguró que la privacidad de las personas incluidas en las listas estaba protegida al haberse eliminado estos identificadores directos.

  Con esta información, Latanya Sweeney, por entonces una estudiante (y después profesora en las Universidades Carnegie Mellon y Harvard, y autora de la teoría de la k-Anonimización como técnica de protección de la privacidad, que se tratará más adelante en este proyecto), comenzó a trabajar en sus investigaciones referentes a la reidentificación de datos previamente anonimizados, y eligió comenzar con la información del propio Gobernador Weld (también incluido en la lista como empleado público del Estado). Sabía que el Gobernador residía en la ciudad de Cambridge (con 54.000 habitantes y 7 códigos postales diferentes), y compró un registro de los votantes de esta ciudad (información también pública, al no asociarse las identidades con ninguna información sensible), que entre otros datos contenía el nombre, dirección, código postal, fecha de nacimiento y sexo de cada votante. Combinando los datos publicados por el Estado con información médica y el registro de votantes, fue capaz de identificar al Gobernador con facilidad (sólo 6 personas en Cambridge tenían la misma fecha de

nacimiento que él, sólo 3 eran hombres y sólo él vivía en su código postal). Latanya Sweeney remitió al propio Gobernador sus averiguaciones (incluyendo sus registros médicos, diagnósticos y prescripciones), y posteriormente demostró que un 87% de los estadounidenses podían ser inequívocamente identificados usando sólo su código postal, fecha de nacimiento y sexo.

- **El caso de los grandes almacenes Target y la adolescente embarazada**

  La cadena de grandes almacenes Target en EE.UU. implementó un sistema para predecir cuáles de sus clientes podrían estar embarazadas, sobre la base de sus patrones de búsqueda y compra de productos, y comparándolos con las búsquedas que habían hecho con anterioridad otras mujeres que estaban esperando un bebé.

  En una ocasión un hombre enfurecido acudió a uno de estos grandes almacenes para quejarse de que su hija, una adolescente que estaba aún en el instituto, había recibido en casa una serie de cupones para comprar ropa de bebé y cunas. Este hombre creía que se trataba de publicidad que de alguna manera podía animar a su hija a quedarse embarazada, pero sin embargo al poco tiempo descubrió tras hablar con ella, que efectivamente estaba ya embarazada.

Estos ejemplos son sólo una pequeña muestra de lo importante que es para cualquier solución de análisis de Big Data tener en cuenta la protección de la privacidad, dado que garantizar que su información personal se gestiona de manera segura y a salvo de revelaciones no autorizadas, es una de las cuestiones que los usuarios de las Tecnologías de la Información y Comunicaciones consideran más importantes y una de las mayores preocupaciones que muestran.

Por ejemplo, se han publicado estudios (como *[93]*) que destacan que el principal motivo por el que los usuarios dejan de usar determinadas redes sociales como Facebook, es la preocupación acerca de que su privacidad no esté siendo preservada convenientemente.

Por todo ello, podría llegar a darse la circunstancia de que no se pudieran aprovechar todas las ventajas que ofrecen las distintas tecnologías disponibles y las metodologías existentes de análisis de datos que se implementaran en relación con los Big Data, si al mismo tiempo no se asegura la Privacidad de los Datos de Carácter Personal empleados en los análisis y el consiguiente alineamiento con la normativa vigente en esta materia.

4.1.2. ¿Qué son y que no son Datos Personales? ¿Cómo los Big Data han modificado esta definición?

¿Qué son y qué no son datos personales? A priori esta parece una pregunta, que al menos hasta hace unos años podía parecer bastante fácil de responder.

De manera simplificada, por datos personales puede entenderse cualquier información que permite identificar a una persona.

Sin embargo, tendencias como el Internet of Things y las nuevas posibilidades que ofrece el análisis de Big Data hacen necesario que la definición de Dato Personal (Personally Identifiable Information –PII-) se concrete en mayor medida, para dar cabida a las posibilidades de identificación indirecta. Así por ejemplo, un metadato, en principio neutro en cuanto a posibilidades de identificación de la persona, en combinación otros datos, puede ayudar a inferir aspectos que si forman parte de la información personal.

Es decir, las posibilidades de análisis predictivo y establecimiento de correlaciones que posibilitan los Big Data, hacen que se incremente en una gran medida la posibilidad a la combinación de datos personales de diferente naturaleza para poder dar a conocer un nuevo

dato personal, que en determinadas ocasiones puede llegar incluso a representar una característica sensible.

La definición de dato personal que se incluye en la principal normativa en relación con la protección de la Privacidad, se encuentra alineada con la idea anterior.

Así, en la *Directiva Europea 95/46/CE de Protección de Datos*, establece que los datos personales son "*cualquier información relativa a una persona identificada o identificable […]; una persona identificable es alguien que puede ser identificado directa* ***o indirectamente****, en particular a través de una referencia a un número de identificación que le es propio, o a uno o más factores específicos de su identidad física, fisiológica, mental, económica, cultural o social*."

Por su parte, en la norma de Protección de la Privacidad en el ámbito de la Sanidad en Estados Unidos (*Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*), se establece también una definición de información individualmente identificable (y se circunscribe lógicamente al ámbito de la salud): *"Información sanitaria individualmente identificable es información que […]: 1) Identifica al individuo; o 2) Aquella con respecto a la cual existe una base razonable para pensar que* ***puede ser empleada para identificar a un individuo****."*

La privacidad, como habilidad de un individuo o grupo de individuos para aislar información acerca de él/ellos y sólo darla a conocer por cualesquiera medios de forma selectiva y voluntaria, se encarga de la protección de los datos personales. En ocasiones se trata de asimilar como sinónimos Privacidad y Anonimización y que es suficiente con proteger esta última a la hora de gestionar datos personales generados, almacenados o transmitidos a través de Tecnologías de la Información y Comunicaciones.

Sin embargo, como se expondrá más adelante, la privacidad tiene muchas otras implicaciones y por lo tanto va más allá de la simple anonimización. Las definiciones que se han presentado de datos personales refuerzan esta idea de que una mera anonimización sintáctica no puede asegurar por si sola la privacidad.

4.1.3. Tipos de Datos Personales

Según la forma en que los datos se obtienen, éstos se pueden categorizar como:

- **Compartidos** por su titular de manera voluntaria (por ejemplo en redes sociales).
- **Datos observados**: capturados por una organización a raíz de las acciones de las personas (por ejemplo datos de localización gracias al empleo de teléfonos móviles).
- **Datos inferidos / agregados**: obtenidos del análisis de otros datos voluntarios, observados o también inferidos (aquí radica la verdadera potencialidad de los Big Data).

Según su capacidad para suponer vulneraciones de la privacidad, la categorización puede ser la siguiente:

- **Datos identificativos**:

- Datos identificativos directos: Son aquéllos que permiten la identificación inequívoca de una persona. Por ejemplo se trata de números identificativos (número de documento de identidad, número de la seguridad social). En el caso del nombre, en ocasiones podrán existir varias personas que tengan el mismo nombre completo, si bien a efectos de este proyecto, se le dará la consideración de dato identificativo directo, como se expondrá en el *Capítulo II*.
- Datos biométricos.
- Datos de identificación débil: datos que potencialmente podrían identificar individuos, pero para lo cual deberán combinarse con otros y acompañarse de un trabajo de análisis. Por ejemplo direcciones IP, pseudónimos, etc.

- **Datos distintivos**:
  - Datos de comportamiento: información de localización, hábitos de compra, patrones de navegación en Internet, etc.
  - Datos de opiniones.
  - Información sensible: Se trata de la información que requiere una mayor protección como la relativa a tratamientos médicos o situación financiera.

4.1.4. Tipos de Ataques y Vulneraciones sobre la Privacidad

Con carácter general las vulneraciones sobre la privacidad se dividen en dos categorías: **revelación de Identidad** y **revelación de Atributos Personales**.

- La **revelación de Identidad** ocurre cuando la identidad de una persona se puede conocer a partir de los datos publicados (por ejemplo a partir de datos identificativos únicos de esa persona, como su número de identificación o de afiliación a la Seguridad Social).
- La **revelación de Atributos Personales** permite inferir información sensible (enfermedades, filiación política, condición sexual, salario, etc.) de una persona a partir de datos publicados.

Además, los ataques de Reidentificación de usuarios, se puede clasificar en tres subcategorías diferentes:

- **Ataques de Correlación**

  Este tipo de ataques se basan en establecer una relación entre los valores que se encuentran en una fuente de información y los de otra, permitiendo un análisis más preciso de las personas para las que la correlación se puede demostrar.

  Son ataques que se emplean en muchas ocasiones con un propósito comercial, en el que el atacante busca reidentificar correctamente cuantas más personas mejor, dado que pueden suponer con una alta probabilidad potenciales clientes para los productos o servicios que ofrece. Por este motivo no suele suponer un problema mayor el que algún individuo se reidentifique incorrectamente.

  Siempre y cuando el objetivo no sea la identificación de personas concretas, sino sólo el establecimiento de correlaciones entre eventos con carácter general (para objetivos de investigación por ejemplo), no se hablará de ataques, dado que no se produciría una violación de la privacidad de ninguna persona.

- **Ataques de Identificación Arbitrarios**

  En este caso, el objetivo es poder asociar, con un nivel de probabilidad suficientemente alto, al menos una entrada de un conjunto de datos con la identidad de una persona concreta, sin importar a priori quien sea ésta.

- **Ataques de Identificación Dirigidos**

  En este caso, el atacante busca encontrar los mayores detalles posibles acerca de una persona concreta. Se consideran los ataques más peligrosos, al tener el mayor impacto sobre la privacidad de los individuos, aunque se dan en un porcentaje más bajo que los ataques de identificación arbitrarios. En estos casos el atacante suele tener un conocimiento previo de la persona sobre la que dirige sus análisis, lo que le hace más factible el descubrimiento de nuevas características.

4.2. Métodos tradicionales de Protección de la Privacidad y cómo impacta en ellos los Big Data

La herramienta más importante con que se contaba hasta ahora para el aseguramiento de la Privacidad era la aplicación de los **Principios Fundamentales de Protección de la Privacidad** (**Fair Information Privacy Practices (FIPP)**):

1. *Limitación en la recopilación de datos / minimización de datos.*
2. *Garantía de la Calidad e Integridad de los datos.*
3. *Determinación explícita del propósito para el que se recopilan los datos.*
4. *Limitación de uso.*
5. *Seguridad.*
6. *Transparencia.*
7. *Participación Individual.*
8. *Responsabilidad y Auditoría.*

Los FIPP, si bien en esencia siguen siendo válidos y deben continuar constituyendo objetivos a alcanzar, deben evolucionar para adaptarse a las especificidades que el ámbito de la gestión y análisis de Big Data supone, al presentarse problemas específicos como:

- Maneras de abordar la reutilización de información para propósitos diferentes a aquéllos para los cuales inicialmente se recopilaron.
- Uso combinado de datos de distintas fuentes.
- Evaluación o segmentación de individuos por su perfil (para propósitos que van desde el marketing hasta decisiones de contratación).

Dada la relevancia de las FIPP, como referencia para la elaboración de la Normativa que rige la protección de datos de carácter personal, un estudio más pormenorizado de los FIPP, así como de sus necesidades de evolución se aborda en el *Capítulo III - Estado del Arte de la Protección de la Privacidad en la Gestión de Big Data – Dimensión Normativa*.

4.3. Retos para la Protección de la Privacidad en el ámbito de los Big Data

La propia naturaleza de los Big Data y el uso que se hace de ellos, conlleva una serie de implicaciones y retos en relación con la protección de la Privacidad:

- No se pueden presentar a priori los usos para los que se van a emplear los Datos Personales. Muchas veces combinándolos con otros se descubren aplicaciones que a priori era imposible determinar. Esto tiene un impacto decisivo en aspectos que hasta ahora constituían el fundamento normativo de los Big Data como el Principio de la Notificación y Consentimiento (Notice and Consent).
- Tal y como se ha expuesto al comienzo de este apartado, no se puede determinar con suficiente claridad qué son Datos Personales y que no. La frontera se ha difuminado con el empleo de las técnicas de análisis y las infraestructuras tecnológicas que posibilitan los Big Data. Por ejemplo la ubicación en un momento determinado de una persona no sería a priori un dato personal, pero combinado con otros aspectos, como un pago con una tarjeta de crédito, pueden llevar a concluir aspectos que den a conocer aspectos de la privacidad de alguien.
- Debe ponerse especial énfasis a la hora de diseñar soluciones para proteger la privacidad en los usuarios generadores de los Big Data relacionados con sus datos personales. Se trata de un aspecto fundamental para el caso del uso de las redes sociales y debe superar el simple consentimiento que se solicita actualmente, sustituyéndolo por soluciones los más protectoras posibles con la privacidad de los usuarios.

Estos retos conllevan una serie de actuaciones que se deben abordar en relación con las soluciones de Big Data:

- Por un lado, se debe dotar a la soluciones de Big Data de mecanismos que aseguren el cumplimiento de la Legislación en vigor, así como una sólida confianza ante los usuarios, crítica para la sostenibilidad de cualquier modelo de negocio.
- Por otro, se debe encontrar un **compromiso (solución intermedia) entre el grado de utilidad y el grado de protección de la información personal**. El nivel más alto de protección podría hacer que quedara sin contenido buena parte de las iniciativas sobre Big Data, ya que se haría mucho más difícil ofrecer por ejemplo servicios personalizados, y por lo tanto podrían verse afectadas buena parte de las ventajas que estas iniciativas presentan, y que se han puesto de manifiesto a lo largo del presente capítulo.
- No es suficiente con definir métodos de protección de la privacidad, sino que éstos deben permitir al mismo tiempo conservar en su nivel más alto posible las ventajas que los Big Data representan para la Organización.

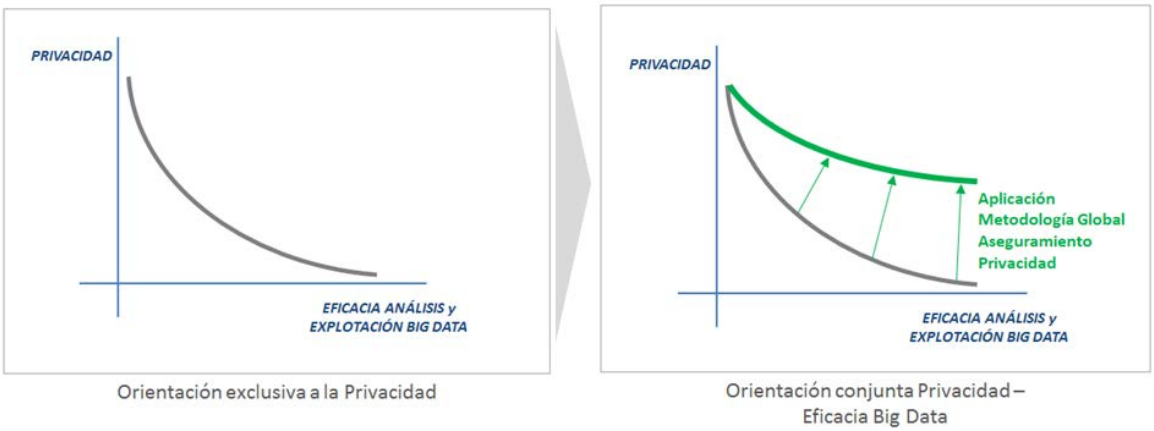


Imagen I-4 – Gráficas ilustrativas tendencia compromiso Utilidad – Protección Privacidad (Elaboración propia)

Es decir se debe buscar una aproximación a la privacidad que valore al mismo tiempo **maximizar el valor que aportan los Big Data al mismo tiempo que se asegura un adecuado tratamiento de la Privacidad** (alineado con los Objetivos Estratégicos de la organización y que cumpla así mismo con la Normativa vigente). Es más, el reto es conseguir que el aseguramiento de la privacidad sea una ventaja competitiva de la analítica y la gestión de Big Data más allá de una desventaja.

Así por ejemplo, en el ámbito del marketing el objetivo debe ser potenciar la posibilidad de las empresas de ofrecer anuncios y servicios personalizados a los clientes en función de las preferencias y necesidades de éstos. Se trataría de una relación Win-Win ya que por un lado la empresa consigue incrementar las posibilidades de venta y la efectividad de su marketing, y por otro el usuario recibe ofertas de servicios o productos que en principio están adaptados a sus requisitos. Se trata de un modelo que emplean ya empresas como Phorm (https://www.phorm.com), que ofrece servicios de anuncios personalizados basados en el historial de navegación de las personas.

No obstante para que este modelo sea realmente efectivo, es critico asegurar que no se hace uso de información basada en identificadores directos de un usuario (como su nombre o su dirección de correo electrónico) sin la autorización expresa de éste (ya que esto puede resultar contraproducente para los objetivos de la empresa u organización que ofrece sus servicios, aunque no se contravengan aspectos legales). Por el contrario, debe garantizarse que se elaboran anuncios que, empleando información anonimizada, pueden agregar datos y conseguir un grado significativo de incremento en la eficiencia de sus campañas al dirigirse a sectores potencialmente más interesados, respetando al mismo tiempo la privacidad de los ciudadanos.

Otro aspecto fundamental que debe ser tenido en cuenta (tal y como se ha destacado en *[52]*) es diferenciar entre resultados Probables y resultados Demostrables, de un análisis basado en Big Data.

Al establecer análisis que correlacionan la identidad de una persona a partir de una serie de datos, es importante tener en cuenta si estos últimos son:

- Datos confiables (o más o menos confiables como la dirección de correo electrónico, dado que siempre cabe la posibilidad de que un usuario pueda mentir en la dirección de correo electrónico que carga por ejemplo en un formulario o que más de una persona compartan dirección de correo), que darán lugar a resultados demostrables con una elevada verosimilitud.
- Datos con los que la correlación no es tan inmediata y/o sólida, como por ejemplo la dirección IP (es posible que una misma dirección IP sea empleada por dos personas completamente diferentes en el marco de un periodo de tiempo muy pequeño). Al usar estos datos menos confiables en términos de establecer en base a ellos la correlación, se debe hablar de resultados basados en la probabilidad, más que de resultados fehacientemente demostrables.

El usar una correlación basada en una probabilidad (que será el caso más usual, dado que en un contexto de gestión de Big Data se combinará el uso de datos confiables con otros que pueden serlo en menor medida), puede llevar a un resultado completamente erróneo, especialmente en el caso de que se encadenen solicitudes que se sustenten en una correlación probable, pero no demostrable.

Es por lo tanto crítico, tener en cuenta estas posibles desviaciones dados los efectos en términos de reidentificaciones negativas que pueden suponer, y las consecuencias negativas que se pueden derivar de ello (por ejemplo hacer creer que una persona cobra un salario

diferente al que en realidad cobra, o envío de información de marketing de los productos de una empresa a alguien que no debería ser objetivo de esa campaña).

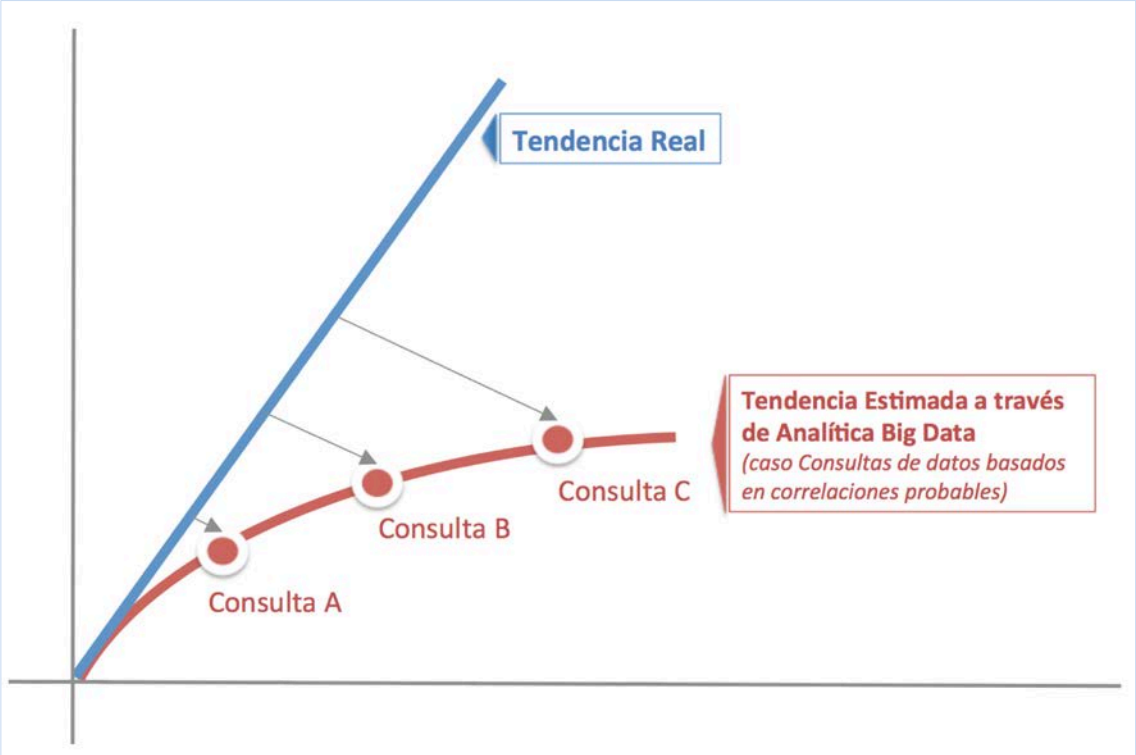


Imagen I-5 - Posible efecto de desviación de tendencias al emplear concatenación de correlaciones basadas en probabilidades

4.4. Enfoque para afrontar los Retos de Protección de la Privacidad

El presente proyecto, tal y como se ha destacado en su *Introducción* al establecer los Objetivos perseguidos, presenta una aproximación global a partir de la cual asentar el éxito de la gestión de Big Data en lo que a la Protección de la Información de Carácter Personal respecta. Para ello se busca la combinación de elementos de las siguientes dimensiones:

- Tecnológica
- Normativa

De nada serviría centrar los esfuerzos en una sola de las dimensiones, sin prestar la debida atención a la otra. En este proyecto se busca presentar una aproximación holística en la que se combinen ambas dimensiones permitiendo obtener unos resultados más óptimos que mediante su planteamiento de manera independiente.

CAPÍTULO II

Estado del Arte de la Protección de la Privacidad en la Gestión de Big Data – Dimensión Tecnológica

1. Introducción y Objeto del Capítulo.

En este capítulo se presenta un análisis del estado del arte en relación con la **Protección de la Información en las Soluciones de Big Data** existentes en el mercado, desde un **punto de vista tecnológico**. Se pondrá especial énfasis dado el tema principal del PFC en el **Aseguramiento de la Privacidad de** los **Datos** que se almacenan y gestionan en Plataformas Big Data.

- En primer lugar se presenta el **modelo de arquitectura general de los Sistemas de Big Data**, concretándolo en el caso de **Hadoop** como implementacion tecnológica más extendida en la actualidad y modelo paradigmático.
- Este análisis permite conocer cuales son los **retos** desde el punto de vista del aseguramiento de la privacidad, inherentes a una plataforma de esta naturaleza, y cuál es el modo actual de afrontar dichos retos.
- Como núcleo del capítulo se muestran los dos **enfoques principales para la protección de los datos de carácter personal** que se manejan en plataformas de Big Data; por un lado el implementar **mecanismos de autenticación y control de acceso** robustos y eficientes, y por el otro el emplear **técnicas potenciadoras de la privacidad** (como la anonimización).
- Por último se presenta un apartado relativo a las **métricas de control** que permiten comparar las soluciones disponibles para la protección de la privacidad en sistemas Big Data, evaluando lógicamente el grado de protección que se alcanza, pero además otros factores como el impacto sobre el rendimiento de la plataforma, la utilidad de los análisis que se pueden seguir haciendo sobre los datos protegidos o el coste.

1.1. Esquema de estructura del Capítulo

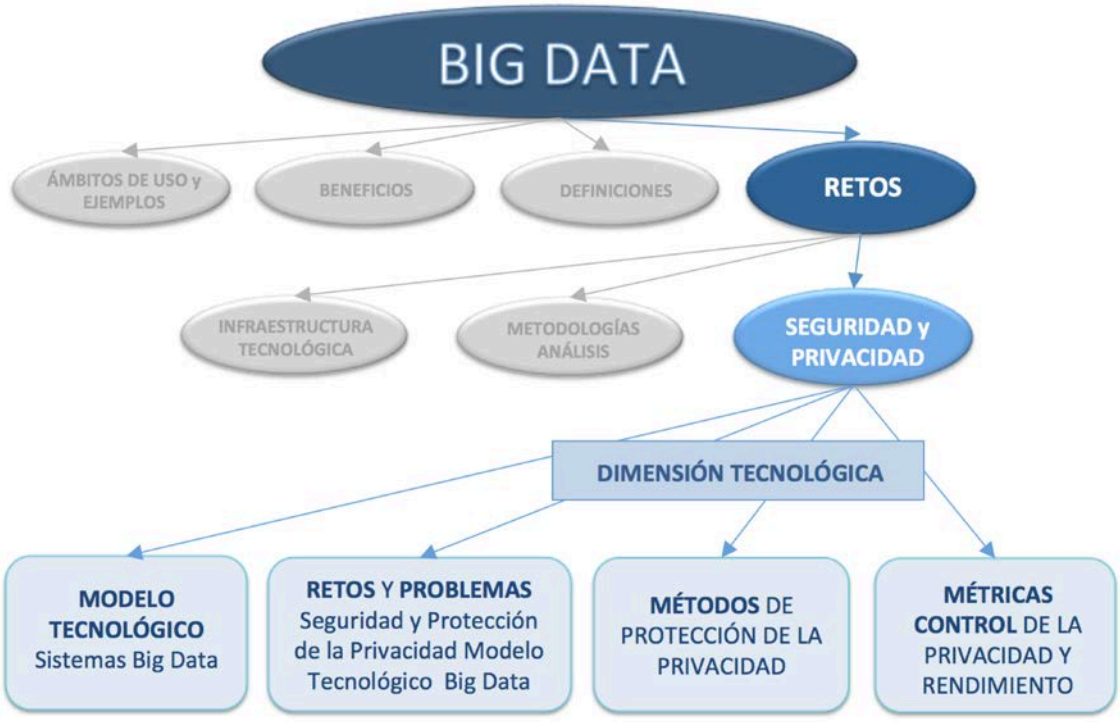


Imagen II-1 – Esquema Estructura Capítulo II

2. El Modelo Tecnológico de los Sistemas Big Data

2.1. Modelo de arquitectura general de Big Data

El elemento fundamental desde el punto de vista tecnológico sobre el que se asientan la arquitectura de los sistemas de Big Data es el concepto de **Procesamiento Distribuido**, a través del cual en lugar de una tarea de procesado de una dimensión muy grande se acometen múltiples procesados pequeños y manejables con equipos disponibles.

> ***Divide et vinces*** *(Divide y Vencerás)*
> *expresión latina*

Este concepto permite dar respuesta a las características de los Big Data ya presentadas en el capítulo anterior (Volumen de datos, Variedad de los mismos, Velocidad de generación y actualización, etc.), aprovechando la capacidad de procesamiento de múltiples equipos que desarrollan tareas en paralelo, coordinados por un equipo central que dirige el proceso.

Gracias a la **escalabilidad** que proporciona este sistema, es posible asegurar una adaptación a la dimensión de la tarea que en cada momento se requiere, reduciendo la infrautilización de plataformas, y haciendo por lo tanto un uso más eficiente de los recursos disponibles.

Además, el sistema de procesamiento distribuido habilita tasas de transferencia de datos rápidas y permite que el sistema pueda desarrollar sus operaciones de manera normal, aun cuando pueda haber un fallo en alguno de los nodos o equipos que se emplean, al ser posible la reordenación y redistribución de trabajos. Se reduce así el riesgo de una caída o fallo global del sistema.

De este modo se contaría (siguiendo el modelo destacado en *[49]*) con:

- Un **Sistema Maestro** (*Master System*). Es el responsable de recibir los datos de las fuentes que recopilan Big Data y los proveen para su análisis, así como las solicitudes de los distintos usuarios autorizados. Las funciones principales que desarrolla son:
  - Distribución de datos.
  - Distribución de tareas.
  - Recopilación de resultados y elaboración de las respuestas a las solicitudes autorizadas.
  - Además desarrollaría las funciones generales de coordinación y aseguramiento de cumplimiento de las políticas de seguridad y calidad que se hayan definido.
- Una serie de **Sistemas Cooperantes** (*Cooperated Systems*), designados por el Sistema Maestro, sobre la base de su capacidad de proceso y de la posibilidad que tengan para implementar las políticas de seguridad y calidad establecidas. Se trata normalmente de equipos de bajo coste, lo que garantiza la modularidad y adaptabilidad ágil del sistema para adecuarse a las necesidades cambiantes de procesamiento y almacenamiento.

Con ello se establecería el denominado **Big Data Cluster**, que conecta el Sistema Maestro con los Sistemas Cooperantes, en un patrón de interacción como el siguiente:

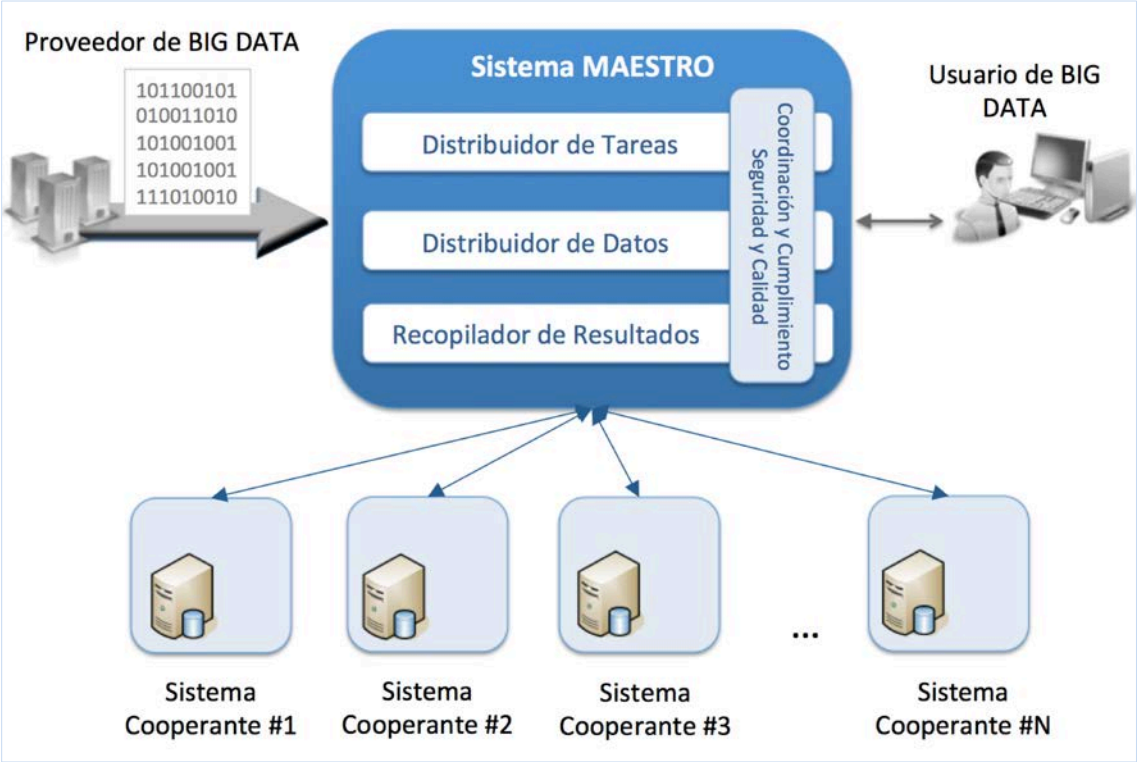


Imagen II-2 – Arquitectura General Big Data

2.2. El Sistema Hadoop

La Arquitectura General anterior se concreta como caso más representativo en **Hadoop**, que se ha constituido como un sistema distribuido para el almacenamiento de grandes cantidades de datos y su procesamiento en paralelo.

Hadoop, que está basado en Java, se ha desarrollado desde el año 2005 como un proyecto de fuentes abiertas de la Fundación Apache, y actualmente se emplea de manera extendida en diversos ámbitos funcionales [6].

Como sistema de procesamiento distribuido, Hadoop es escalable, eficiente en término de costes y empleo de recursos, flexible y tolerante a fallos, incluso en entornos de caída de varios sistemas cooperantes empleados.

2.2.1. Esquema general de Hadoop

Hadoop se articula alrededor de dos componentes fundamentales:

- **MapReduce,** que sería el equivalente al Distribuidor de Tareas de la Arquitectura General, articulándose como la capa de procesamiento y computación del sistema. MapReduce implementa el modelo de computación distribuida de Google (según se reseña en [26]), adaptándolo a las necesidades de procesamiento en paralelo con datos muy grandes (del orden de varios Petabytes como los que se requieren en Big Data), y se basa en el establecimiento de trabajos que a su vez consisten en dos pasos: **Map** y **Reduce**:
  - En el paso **Map**, los datos de entrada son preprocesados. Para ello el **Sistema Maestro** recibe los datos de entrada que deben ser procesados, los divide en partes y transfiere éstas a los Sistemas Cooperantes, o **Nodos Esclavos** como se denominan en Hadoop, que son los encargados de procesarlos.

- En el paso **Reduce**, los datos preprocesados en los nodos cooperantes son recopilados por el Sistema Maestro, que compone en base a ellos la solución requerida.

- El **Sistema de Ficheros de Hadoop (Hadoop Distributed File System –HDFS-)**, que equivaldría al Sistema de Distribución de Datos de la Arquitectura General. Es la capa de almacenamiento de Hadoop. Es un sistema distribuido, escalable, basado en Java, especialmente diseñado para el almacenamiento de grandes volúmenes de datos no estructurados.

  HDFS contiene un servidor de metadatos denominado **Name Node**, que gestiona los metadatos del sistema de ficheros, y una serie de **Data Nodes** que almacenan los bloques individuales de cada fichero.

- Además, existen otros componentes que ofrecen funcionalidades que complementan el funcionamiento de Hadoop, formando así mismo parte del denominado **Ecosistema Hadoop** (que se presentará más adelante).

El **Sistema Maestro** contará con un componente **Job Tracker** y otro **Name Node**, mientras que cada uno de los **Nodos Esclavos** tendrían también dos componentes: **Task Tracker** y **Data Node**. Job Tracker y Task Tracker se agruparán dentro de MapReduce y Name Node y Data Node, como ya se ha apuntado, permitirán el desarrollo de las funciones de HDFS.

La siguiente imagen muestra gráficamente este modelo [*[49]*]:

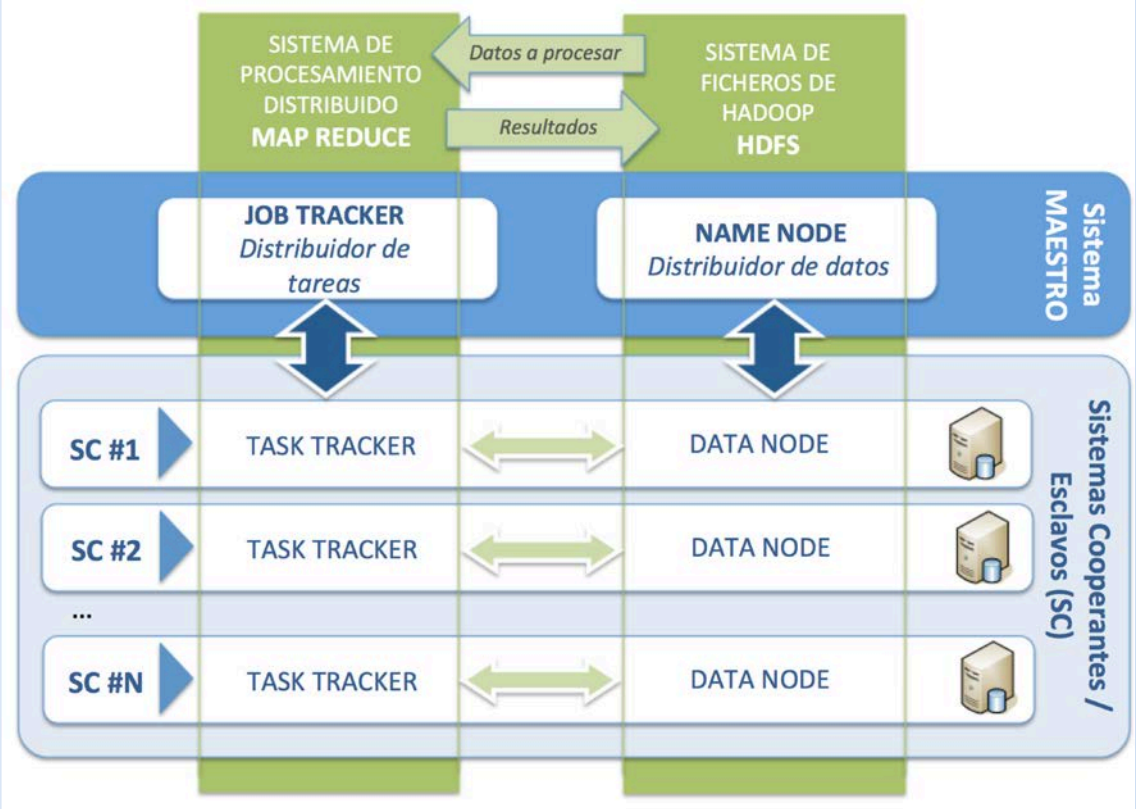


Imagen II-3 – Esquema General Hadoop

2.2.1.1 El esquema de funcionamiento de MapReduce en Hadoop

El componente Job Tracker del Sistema Maestro es el encargado de programar las diferentes tareas que componen los trabajos a ejecutar, para que puedan ser desarrolladas en los

Sistemas Esclavos. Además se encarga de monitorizar el progreso de los Sistemas Esclavos y de re-ejecutar las tareas que pudieran haber fallado.

Para ello, consulta previamente el componente Name Node para asignar las tareas a aquellos nodos que tienen disponibles los datos necesarios para el desarrollo de la tarea.

Por su parte, el componente Task Tracker de cada Sistema Esclavo ejecuta las tareas según las directrices establecidas por el Sistema Maestro.

Funcionamiento de MapReduce *[12] y [65]*:

El funcionamiento de MapReduce se basa en mapear los conjuntos de datos sobre los que se acometerán los distintos trabajos en pares de <Clave, Valor>.

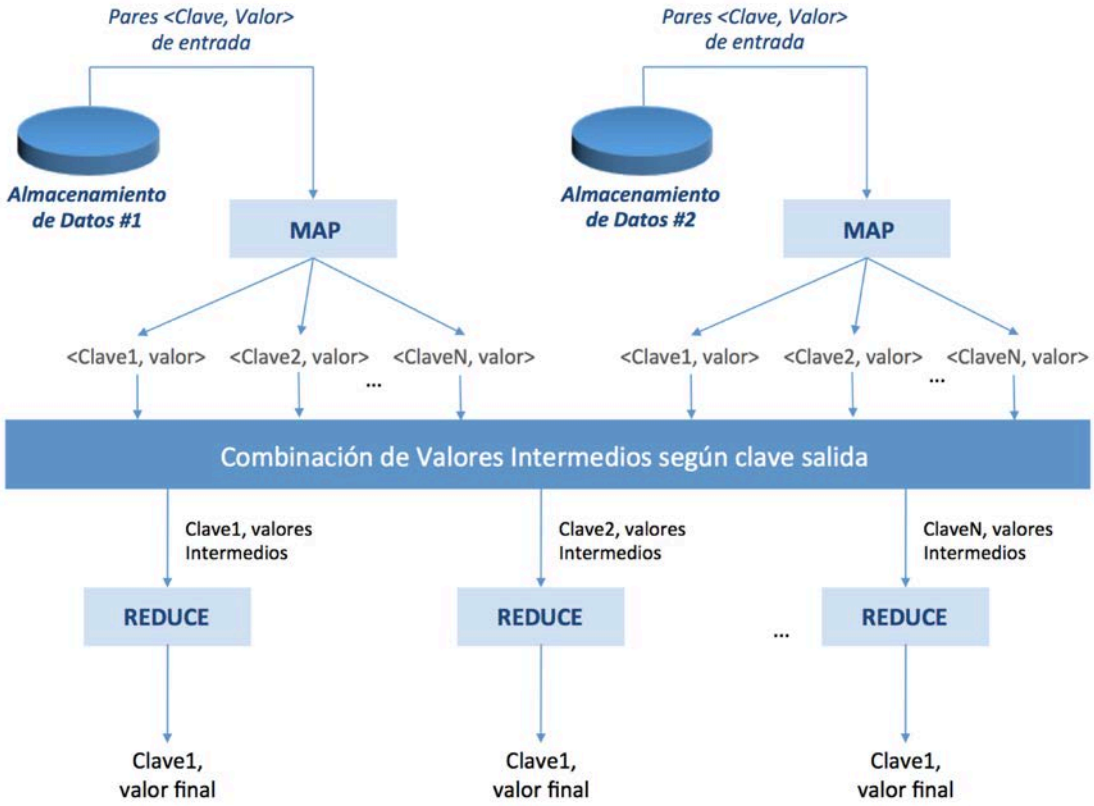


Imagen II-4 – Esquema Funcionamiento MapReduce

En la siguiente imagen se muestra el funcionamiento de modo esquemático con un ejemplo, en el que se busca cuantificar el número de veces que una palabra aparece en un texto (un ejemplo de este tipo podría servir, evidentemente en una dimensión mucho mayor, para conocer qué palabras se repiten con más frecuencia en búsquedas en Internet, y poder adaptar en consonancia una página Web (constituyendo de este modo un apoyo para la estrategia de Posicionamiento en Buscadores –SEO- asociada a dicha página)).

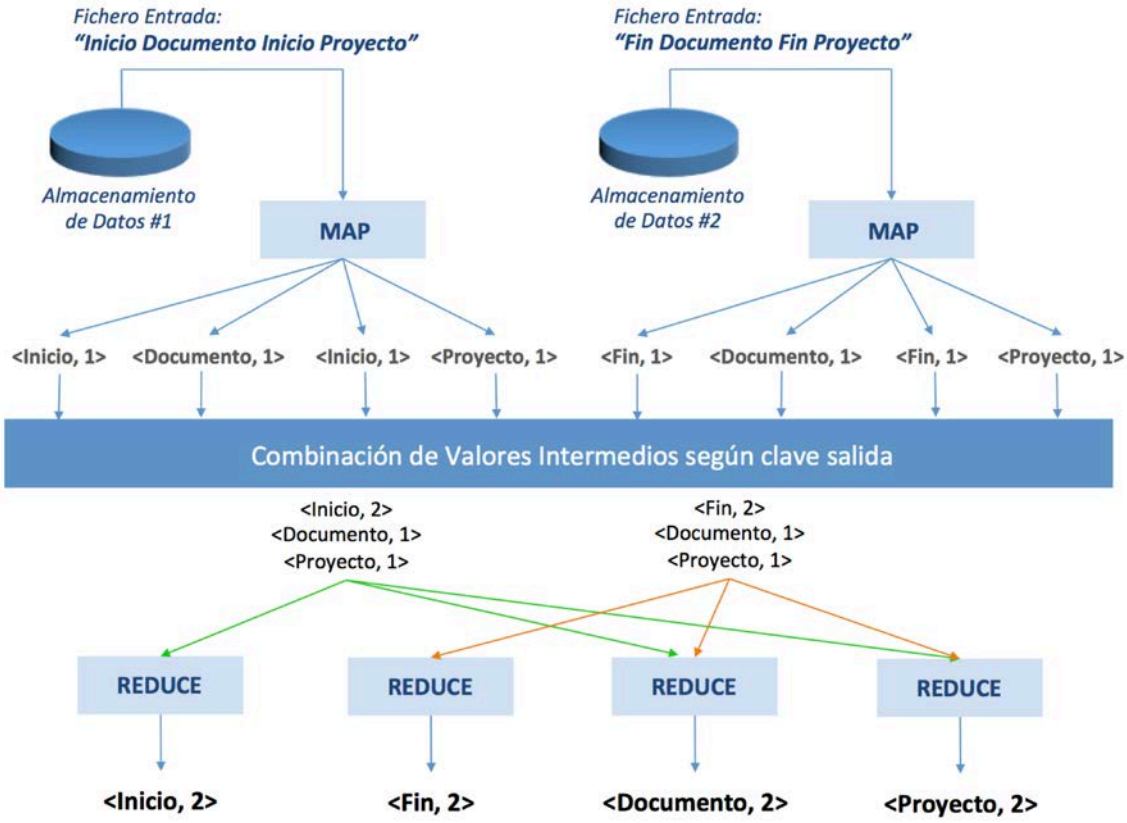


Imagen II-5 – Ejemplo Funcionamiento MapReduce

Programación:

Por defecto MapReduce emplea un esquema FIFO (-First In, First Out-, primer trabajo en entrar, primer trabajo en salir) para programar la ejecución de los distintos trabajos/tareas. No obstante se pueden emplear otras alternativas como la Programación Equitativa o por Capacidad.

En un esquema de *Programación Equitativa (Fair Scheduling)*, se asignan recursos a los diferentes trabajos de manera que todos los trabajos tengan disponible, en media, un número igual de recursos a lo largo de su tiempo de ejecución. En caso de que existiera un único trabajo, este ocuparía el cluster completo, y cuando se recibieran nuevos trabajos, éstos ocuparían los recursos que estuvieran libres.

Este esquema, al contrario de lo que ocurre con el funcionamiento FIFO por defecto, permite que los trabajos pequeños terminen en un tiempo razonable (al no tener que esperar en una cola a que finalicen otros posibles trabajos de mayor duración). Además la Programación Equitativa, permite tener en cuenta las prioridades que puedan haberse asignado a los trabajos a la hora de asignar el tiempo de procesado que cada trabajo tendrá.

El esquema de *Programación por Capacidad (Capacity Scheduling)* permite la compartición del cluster Big Data, garantizando una capacidad mínima a cada Organización / Usuario que lo emplee. La capacidad que una organización pueda no estar usando, pasará a estar disponible para otras, con lo que se consigue una flexibilidad que redunda en un funcionamiento eficiente en términos de tiempos de respuesta y empleo de los recursos.

Tolerancia a fallos:

El sistema Hadoop es tolerante a fallos, es decir puede seguir ejecutando los trabajos que le son requeridos, sin perdida de información incluso en caso de caída de alguno de los componentes del cluster. Esto se consigue gracias a la redundancia de datos que permite el esquema de HDFS.

En concreto si se detecta un fallo en un Data Node de un Sistema Esclavo, el Name Node del Maestro elimina dicho Data Node del cluster y redistribuye los datos que pudiera contener entre otros Data Nodes que no tuvieran problemas. Del mismo modo, si el problema se detecta en un Task Tracker, será el Job Tracker del Maestro el que decidirá como reprogramar la/s tarea/s que ese Task Tracker tuviera asignada/s.

En caso de fallo del Sistema Maestro, el sistema puede fallar, dado que contiene toda la información importante para la distribución de datos y su procesado. Para evitarlo se ha introducido el concepto de Sistema de Backup, que contendrá un Nodo de Backup que mantendrá actualizada la información del Name Node.

Limitaciones MapReduce:

- El esquema de funcionamiento de MapReduce hace que una vez lanzadas por el Sistema Maestro las diferentes tareas a ejecutar, no pueda controlarse el orden en que los sucesivos pasos de Map y Reduce son ejecutados.
- Además debe asegurarse que los pasos de Map y Reduce no dependan de los datos que puedan ser generados en el mismo trabajo MapReduce.
- Las operaciones de Reduce no ocurren hasta que todas las operaciones de Map se han completado.

2.2.2. El Ecosistema Hadoop

El ecosistema de Hadoop está formado por una serie de componentes, muchos de ellos desarrollados igualmente como proyectos en la Fundación de Software Apache, que ofrecen las distintas funcionalidades que garantizan una eficiente y coordinada gestión de los datos, permitiendo a los usuarios de la plataforma Hadoop obtener las ventajas que ofrecen los Big Data en unas condiciones de seguridad y del modo más eficiente posible.

Entre los componentes más relevantes de este ecosistema cabe mencionar los siguientes:

- **Hadoop Distributed File System (HDFS)**, ya presentado en el subapartado anterior.
- **MapReduce**, también presentado en el subapartado anterior.
- **YARN** (Yet Another Resource Negotiator): Es una tecnología para gestión de clusters y recursos que se articula como un MapReduce de nueva generación, dividiendo las funcionalidades principales de Job Tracker en dos: Gestión de Recursos y Monitorización y Planificación de trabajos.
- **HBase** (Hadoop DataBase): Base de datos no relacional organizada en columnas y distribuida, derivada de Google Big Table. Permite búsquedas rápidas con baja latencia en Hadoop.
- **Apache Pig**: Se trata de un compilador que genera una secuencia de programas MapReduce, junto con un lenguaje de programación propio (Pig Latin). Proporciona apoyo para el desarrollo de consultas tipo SQL (Structured Query Language) a las bases de datos distribuidas de Hadoop.

- **Hive**: Es una Infraestructura de Almacén de Datos (Data Warehouse) que permite escribir consultas en un lenguaje similar a SQL denominado HiveQL, y posteriormente convertirlas al léxico de MapReduce. Esto permite el uso del almacén por parte de programadores SQL sin experiencia en MapReduce, y hace más fácil la integración con herramientas de Inteligencia Empresarial y de Visualización.
- **Mahout:** Es una librería de DataMining y aprendizaje de máquina e incluye algoritmos clave como los de clasificación, modelado estadístico y filtrado por recomendaciones y colaborativo. Los algoritmos básicos son implementados con el paradigma MapReduce.
- **HCatalog**: Es un servicio centralizado de gestión y compartición de datos, que permite disponer de una vista unificada de todos los datos que existen en los clusters de Hadoop y posibilita que diferentes componentes como Hive o Pig procesen cualquier elemento de datos sin tener que saber en que parte del cluster se encuentra almacenado físicamente el dato en cuestión.
- **Avro:** Es un sistema de serialización de datos para codificar el esquema de los ficheros Hadoop.
- **Oozie:** Es un sistema de procesado de flujos de trabajos que permite a los usuarios definir una serie de tareas (escritas en distintos lenguajes como Hive, Pig o MapReduce) y conectarlas entre sí. Esto permite por ejemplo establecer que una solicitud determinada sólo debe iniciarse después de la finalización de un trabajo específico previo sobre cuyos datos resultado se basará, mejorando así la coordinación y la gestión de medios disponibles.
- **Zookeeper:** Su función principal es proporcionar servicios de sincronización distribuida para proveer servicios en alta disponibilidad.
- **Ambari:** Son una serie de herramientas Web para el despliegue, administración y monitorización de clusters Apache Hadoop.
- **Chukwa**: Se trata de un componente de monitorización de sistemas distribuidos para la gestión de Big Data.
- **Sqoop:** Es una herramienta de conectividad para mover datos de sistemas de almacenamiento No Hadoop, como por ejemplo Bases de Datos relacionales y almacenes de datos, a Hadoop. Permite a los usuarios especificar la localización destino dentro de Hadoop a la que se moverían los datos.
- **Flume:** Es un servicio distribuido y confiable para la recopilación, agregación y movimiento de volúmenes grandes de datos de registro (logs).

El siguiente esquema se presentan los componentes anteriores, estructurándolos según las funciones que desarrollan (gestión de datos, acceso a datos, almacenamiento de datos, procesado de datos, comunicación con el exterior,etc.)

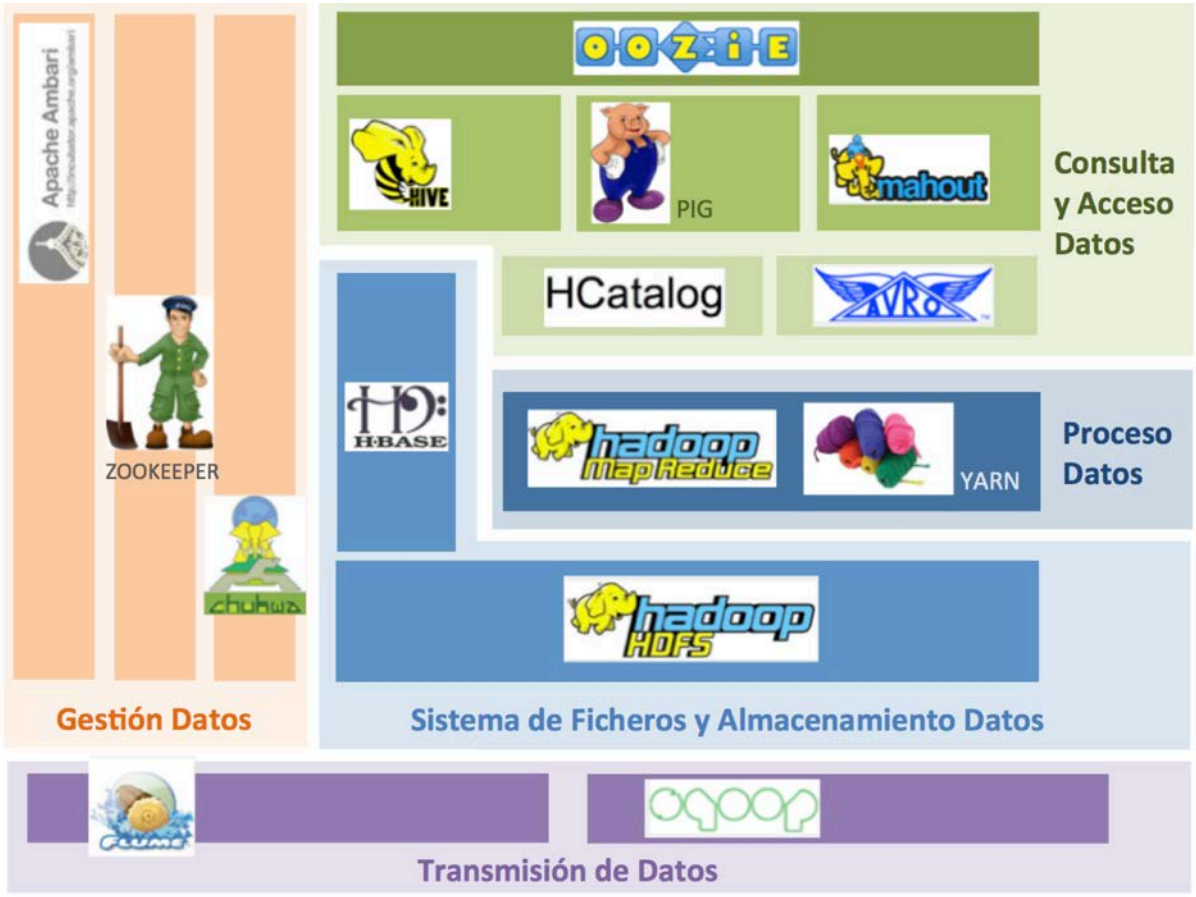


Imagen II-6 – El Ecosistema Hadoop

Este ecosistema se relaciona con productos y soluciones externas que requieren resultados derivados del análisis de Big Data, por ejemplo soluciones de Inteligencia Empresarial. Se trata de un ecosistema en constante evolución, respondiendo a los múltiples retos y necesidades que surgen de un empleo cada vez más extendido y exigente de los Big Data.

2.3. Spark

Apache Spark es un marco de referencia de fuentes abiertas, para el procesamiento de grandes volúmenes de datos en paralelo, que destaca por su carácter general (siendo de aplicación en múltiples ámbitos) y por su velocidad. Se trata de un complemento para Apache Hadoop, y se articula como una alternativa a MapReduce *[8]* y *[109]*.

Spark se ha convertido en uno de los proyectos más importantes de la Fundación Apache en relación con la gestión de Big Data, y uno de los que está concitando más interés dado el número de iniciativas y desarrollos que en relación con él se están llevando a cabo.

Spark permite hacer más fácil el desarrollo ágil de aplicaciones Big Data, combinando procesos batch, streaming y análisis interactivos de los datos. Así, uno de los objetivos fundamentales del proyecto es extender el modelo de MapReduce para dar un soporte más eficiente a dos tipos de análisis de datos:

- Algoritmos iterativos (como los empleados en aprendizaje máquina o en grafos (de especial utilidad por ejemplo para análisis de datos derivados de redes sociales)).
- Minería de Datos Interactiva.

Con los esquemas de procesamiento actuales, las aplicaciones recargan los datos del sistema de ficheros en cada solicitud de información, con el consiguiente incremento del tiempo medio de respuesta.

Estas características hacen al proyecto idóneo para aplicaciones de:

- Procesamiento en tiempo real
- Integración y procesado de datos

Se basa en el empleo de Series de Datos Distribuidas y Resilientes (Resilient Distributed Datasets –RDD-), las cuales son una serie de elementos tolerantes a fallos, sobre los que se puede operar en paralelo.

Como se ha indicado, con carácter general Spark destaca por:

- Su rapidez de procesado, alcanzando velocidades entre 10 y 100 veces más rápidas que por ejemplo MapReduce, en operaciones en memoria y en disco, lo que supone mejores decisiones y prestación de resultados a usuarios.
- Su potencia y versatilidad, al permitir escribir de manera ágil aplicaciones en paralelo en Java o Phyton sin tener que pensar únicamente en términos de operadores map y reduce.
- Su integración con otras aplicaciones y diversas distribuciones de Hadoop, siendo capaz de leer datos en HDFS.

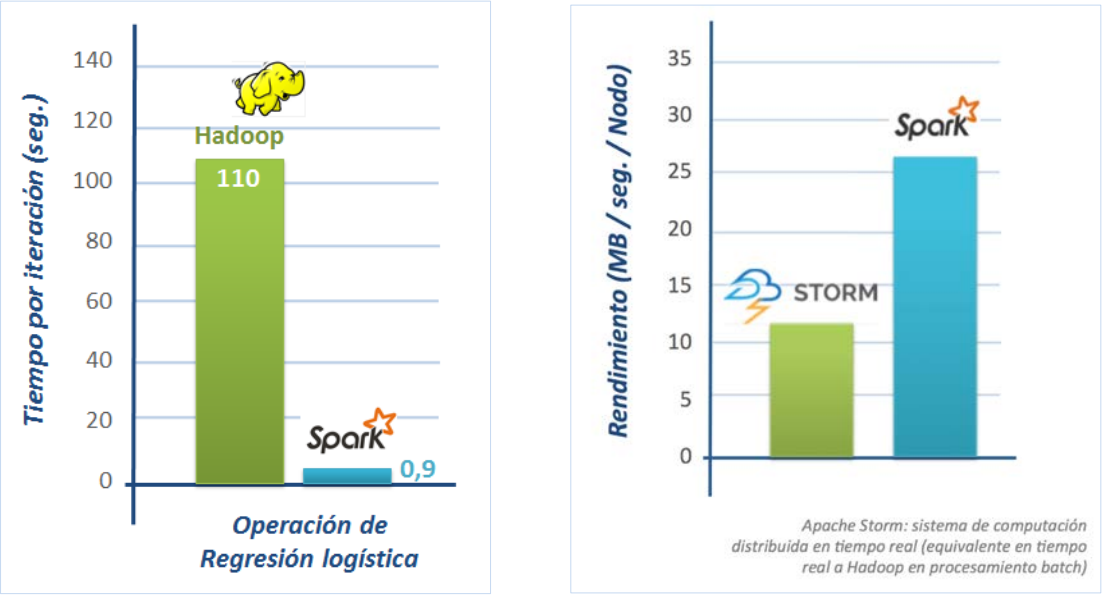


Imagen II-7 – Estadísticas comparativas Spark con otras aplicaciones de Big Data
(Fuente cloudera.com y spark.apache.org/)

2.3.1. Aplicaciones actuales de Spark

Spark no es simplemente un proyecto teórico, sino que su aplicación en casos reales de empresas de gran relevancia se está incrementando en los últimos tiempos.

De este modo cabe mencionar como el servicio de EMR (Elastic Map Reduce) de Amazon Web Services, se ha actualizado para manejar aplicaciones Spark.

También Yahoo! lo está empleando en un proyecto orientado a la personalización de páginas de noticias para usuarios y en otro para ejecutar análisis orientados a publicidad. En iniciativas que implican personalización es imprescindible la adaptación con rapidez a lo que hace el usuario y para ofrecerle soluciones prácticamente en tiempo real (lo que un usuario pudiera buscar hace unas horas, puede que ya no le sea de interés o que ya haya encontrado la solución que buscaba), por lo que el empleo de enfoques y capacidades como los que proporciona Spark suponen un avance fundamental.

Twitter también está utilizando las capacidades de Spark en un proyecto para clasificación de tweets como spam.

3. Retos y Problemas actuales relativos a la Seguridad de la Información y la Protección de la Privacidad en Big Data

3.1. Introducción

En lo que respecta a la Protección de la Seguridad de la Información y la Privacidad el reto, con carácter general y para cualquier Sistema de Información y Telecomunicaciones, es doble:

- Por un lado evitar que **atacantes externos** o usuarios con intenciones ilícitas **accedan al sistema**. Estos atacantes pueden buscar:
    - **Robar información**.
    - **Alterar el proceso correcto de ejecución de trabajos** para evitar que se obtengan resultados adecuados (sabotaje). Para ello, se pueden valer de la **remisión de solicitudes** con el objetivo de retrasar la ejecución de aquéllas que sí buscan un propósito lícito, o de la **inserción de datos aleatorios** o erróneos que modificarían el sentido del resultado que obtendrían otros usuarios (por ejemplo en un análisis predictivo de una tendencia de consumo de un producto y su relación con otra variable como el consumo de otro producto o el grado de actividad de los consumidores en redes sociales).
- Por otro lado se debe **evitar** que en el desarrollo habitual de los trabajos y ejecución de peticiones, los **usuarios autorizados puedan obtener información en relación con aspectos de la privacidad de los individuos** cuyos datos puedan estar almacenados en el sistema en cuestión.

Lógicamente se trata de aspectos que deben ser evaluados en cualquier sistema, pero que además en el caso de los Big Data, y dadas las especiales características de este tipo de datos (tal y como ya se ha destacado en el *Capítulo I*), requieren una aproximación particularizada y adaptada a los retos y especificidades existentes.

3.2. Las características inherentes de Big Data y su impacto sobre la Seguridad de la Información y la Privacidad

Si bien la Seguridad de la Información y la Protección de la Privacidad son aspectos fundamentales en la gestión de Big Data, se trata de ámbitos en los que, con carácter general y hasta ahora, no se ha producido un importante desarrollo.

Las propias **características inherentes a los sistemas de gestión de Big Data** (y entre ellos Hadoop) hacen muy difícil la implementación de los mecanismos eficientes tradicionales de gestión de la Seguridad de la Información y de Control de Acceso, dada la gran dificultad para

controlar todos los accesos y solicitudes susceptibles de producirse, sin con ello hacer inviable el desarrollo de la funcionalidad pretendida *[68]*.

Estas características son:

- **Computación Distribuida**. Los sistemas de gestión de Big Data se basan en sistemas de computación distribuida de carácter masivo y con procesamiento en paralelo (Massive Parallel Processing – MPP-) entre el Sistema Maestro y los Sistema Cooperantes o Esclavos, tal y como se ha expuesto en el *apartado 2.1.* anterior.

  Se trata de un entorno mucho más difícil de securizar, dado que las interacciones no siguen el patrón típico de relación Cliente-Servidor. Por este motivo:

  - Se requieren **múltiples puntos de autenticación** debido a que cada conjunto de datos se divide y distribuye a lo largo de múltiples Sistemas Cooperantes, que ejecutan miles de tareas concurrentes.
  - Un trabajo se ejecuta en **sistemas distintos a aquél en el que se autenticó el usuario** que requería ese trabajo (u otro de mayor dimensión que se ha dividido, entre otros, en ese trabajo en cuestión). Además la ejecución del trabajo se efectúa pasado un tiempo variable desde la autenticación, por lo que se hace necesario que se garanticen las condiciones de seguridad para el usuario en todo el sistema y a lo largo de todo el tiempo que dure la ejecución de la petición, así como el tratamiento que se hará de sus credenciales.

    En este sentido destaca la necesidad de asegurar la **propagación de credenciales a lo largo del sistema**, aspecto complejo y que puede implicar riesgos de violaciones de la seguridad, e incremento de las posibilidades de robo de dichas credenciales por un hipotético atacante.
  - Tareas de **diferentes usuarios se ejecutan en un mismo nodo de computación o sistema cooperante**.
  - Al poder ocurrir el procesamiento de los datos en cualquiera de los nodos que componen el sistema, resulta **muy complicado garantizar la seguridad del lugar exacto** donde un trabajo concreto se está desarrollando.
- **Datos Distribuidos.** Se manejan datos fragmentados y distribuidos, para que puedan ser empleados por la plataforma de computación distribuida. Además en muchas ocasiones los datos serán redundantes para garantizar la fiabilidad y la tolerancia a fallos propios del sistema.

  Todo ello hace más difícil garantizar dos de las dimensiones fundamentales de la Seguridad de la Información como son la **Integridad** (mantener los datos libres de modificaciones no autorizadas, garantizando que se conserva su contenido original) y la **Disponibilidad** (característica que garantiza que la información se encuentra a disposición de quien debe acceder a ella en el momento que así lo requiera).
- **Tamaño de los sistemas** (con múltiples nodos esclavos). Se trata de un reto en sí mismo, dado que los sistemas de autenticación actuales, como Kerberos, no pueden manejar un orden de magnitud de tantas tareas y usuarios autenticándose directa y simultáneamente.
- **Comunicación nodo a nodo**: normalmente Maestro y Esclavo se comunican a través de protocolos no seguros como RPC (Remote Procedure Call) sobre TCP/IP, en redes inalámbricas y cableadas, siendo posible que un atacante externo intercepte la comunicación entre nodos para acceder a la información transmitida.

- **Derechos de los Nodos Esclavos**. El hecho de que cualquier nodo pueda desarrollar una tarea y por lo tanto sea susceptible de procesar cualquier dato, es una ventaja en términos por ejemplo de mejora de la tolerancia a fallos e incremento en la velocidad de presentación de resultados (tal y como se ha expuesto hasta ahora). Sin embargo, al mismo tiempo puede suponer un riesgo en caso de que un atacante externo se haga con el control de algún nodo, para robar datos de usuarios o entorpecer el normal desarrollo de los trabajos.
- **Existencia de un número elevado de Sistemas Auxiliares**. Tal y como se comprueba del análisis del Ecosistema Hadoop, los usuarios pueden acceder o hacer uso de múltiples servicios auxiliares (como por ejemplo Oozie mediante el cual el usuario puede crear un flujo de trabajos conectados entre si). El modo en que las credenciales de usuario se compartirán entre estos servicios auxiliares también supone importantes retos, así como el desarrollo de la noción de confianza.

Debido a todo lo anterior, se comprueba como los enfoques tradicionales de protección de perímetro y de seguridad de la información no son adecuados para la protección de un Cluster de Big Data.

3.3. Seguridad y Privacidad en Hadoop – Situación actual

El presente apartado muestra un análisis del estado actual en relación con la protección de la Seguridad de la Información y de la Privacidad en Hadoop, destacando las principales amenazas a las que se enfrenta la plataforma, aspectos de securización física, las capacidades básicas implementadas (Kerberos) y las tendencias y buenas prácticas sobre las que se prevé que se asiente la evolución en este sentido *[25]*, *[68]* y *[79]*.

3.3.1. Introducción a la Seguridad en Hadoop - Capacidades Iniciales y Amenazas de Seguridad

Ante los retos anteriores, y si bien como ya se ha apuntado, la Seguridad de la Información no se ha considerado en un primer momento una de las prioridades en el desarrollo de Hadoop, sí **se han implementado unos mecanismos y capacidades básicas** que pueden servir de base sobre la que asentar una evolución posterior.

Desde un principio en la implementación de Hadoop se han venido empleando **controles sobre la autorización de acceso a datos y trabajos basados en permisos de acceso a ficheros**. Para incrementar la agilidad se busca no crear cuentas de usuario para cada uno de los usuarios de Hadoop, sino que en lugar de ello se emplean las cuentas de usuario y los identificadores locales existentes en la organización de la provengan o a la que pertenezcan los usuarios en cuestión. De este modo, **en principio, Hadoop depende de esta forma de credenciales de usuario externas**.

Además, el que todos los usuarios tengan a priori los mismos permisos también constituye un reto para la Seguridad de la Información en el Sistema, dado que permite que un atacante pueda modificar datos en otros cluster o priorizar sus tareas por delante de las de usuarios autorizados. Debe por lo tanto ser abordada una solución para este caso.

Inicialmente, en el modo de trabajo por defecto no existe cifrado de los datos entre Hadoop y el cliente y en HDFS todos los ficheros se almacenan sin cifrar, controlados por las capacidades que ofrece el Name Node del Sistema Maestro. Además como ya se ha indicado, la comunicación entre los sistemas Maestro y Esclavos tampoco está cifrada.

El hecho de que Hadoop **no cifre los datos** que almacena, procesa y transmite se debe a que se ha priorizado en principio la **mejora de la eficiencia en su manejo**, al necesitarse **menos recursos de computación** para procesar los datos no cifrados y los **tiempos de respuesta que se obtienen son más reducidos**.

Las circunstancias anteriores, unidas a las inherentes a un sistema de Big Data que se han presentado con anterioridad, permiten definir una serie de posibles violaciones de seguridad en Hadoop:

- Un usuario no autorizado puede acceder al HDFS.
- Un usuario no autorizado puede leer o escribir datos en el sistema de ficheros.
- Un usuario no autorizado puede cargar un trabajo para su ejecución, cambiar la prioridad de los trabajos o incluso borrar trabajos en espera de ser ejecutados.
- Una tarea en ejecución puede acceder a los datos de otra tarea.

A nivel conceptual, posibles soluciones de seguridad para las violaciones de seguridad anteriores pueden consistir en:

- Establecimiento de mecanismos de control de acceso a nivel del sistema de ficheros.
- Establecimiento de mecanismos de control de acceso antes de las operaciones de lectura / escritura.
- Establecimiento de mecanismos de autenticación segura de usuarios. Sin la adecuada autenticación (identificar fehacientemente la identidad del usuario) es imposible asegurar que la autorización (especificar los derechos de acceso de un usuario autenticado a determinados contenidos) es la apropiada.

Una autenticación basada simplemente en el uso de passwords no es eficiente frente a ataques de replicación (el atacante copia una parte del flujo de comunicaciones (que como se ha dicho no está cifrada) entre dos partes y lo reproduce para varias partes, en busca de autenticaciones positivas) o frente a robos de las passwords.

Para abordar los problemas anteriormente citados e implementar las soluciones que a nivel conceptual se han presentado, se han ido incorporando en Hadoop los siguientes mecanismos de seguridad:

- Soluciones de Seguridad de Perímetro Físico: Firewalls y Apache Knox.
- Sistemas de Autenticación Fuerte: Kerberos.
- Mecanismos de Autorización: Permisos HDFS, Listas de Control de Acceso de HDFS.

Con estos mecanismos, se puede hablar de una **configuración no segura** de Hadoop, en la que basta con que el usuario se autentique en el equipo / sistema desde el que se accede a Hadoop, y una **configuración segura**, en el que este sistema de login se complementa con una autenticación Kerberos y cualquier acceso subsiguiente a Hadoop lleva consigo las credenciales para autenticación.

3.3.2. Seguridad de Perímetro Física

Las capacidades definidas en la Política de Seguridad de Hadoop hacen que cada nodo de un cluster Hadoop sea **físicamente seguro** y tenga instalado el software Hadoop por los administradores del sistema; los usuarios no tienen acceso directo a los nodos y no pueden instalar ningún software en ellos.

Los usuarios no pueden ser superusuarios de ningún nodo del cluster. Más aún, un usuario no puede conectar un nodo que no pertenezca al cluster (por ejemplo una estación de trabajo del usuario) a la red cluster.

Además, como elemento específico de seguridad física, se cuenta como sistema de seguridad perimetral con:

- Firewalls de Seguridad de Red.
- Gateway de Apache Knox *[7]*.

  Es un sistema que proporciona un único punto de acceso y autenticación para los servicios de Apache Hadoop, accediendo al cluster de Hadoop sobre HTTP/HTTPS.

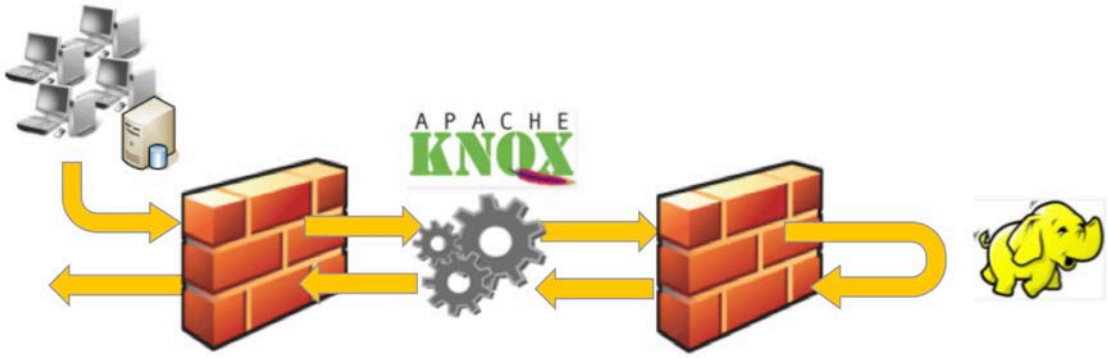


Imagen II-8 – Esquema Apache Knox

3.3.3. Capacidades de Protección Básicas – Sistema de Autenticación Kerberos

Dado que Hadoop se emplea para el almacenamiento y procesamiento de información sensible (como información personal o identificativa de personas) **es necesario asegurar una autenticación fuerte**.

Kerberos es el sistema a través del que se posibilita esta autenticación fuerte, y se caracteriza por:

- Emplea tecnología de cifrado, normalmente basado en criptografía de clave simétrica, en la que cada cliente comparte una clave secreta con Kerberos, conocida por ambos, y en la que el conocimiento de la clave permite al usuario probar su identidad. Se seleccionó esta opción dado que se obtienen resultados más rápidos que por ejemplo empleando operaciones de clave pública.
- Además se basa en el establecimiento de una tercera parte confiable, que actúa como mediador y que se denomina Centro de Distribución de Claves (Key Distribution Center – KDC-). El KDC se basa en un Servidor de Autenticación y un Servidor Emisor de Tickets.
- Permite una gestión de usuarios más sencilla, en la que la revocación de un usuario se puede hacer simplemente borrando a ese usuario del KDC, en lugar de tener que generar y distribuir a todos los usuarios una nueva lista de certificados revocados.
- Se basa en el uso de tickets criptográficos para evitar el envío de passwords, siendo de esta forma capaz de ofrecer autenticación segura en una red abierta.

Fue desarrollado en el Instituto Tecnológico de Massachussets (Massachussets Institute of Technology – MIT-) y se basa en el protocolo Needham-Schroeder.

De manera inicial se ha empleado Kerberos como mecanismo de autenticación y por lo tanto de seguridad; de manera complementaria se emplean otras soluciones como los tokens delegados, los tokens de acceso por capacidades o el desarrollo de la noción de confianza en servicios auxiliares.

La política de empleo de tickets busca reemplazar a los modelos de control de acceso y autorización basados en listas (Access Control Lists – ACL-) los cuales en muchos casos no son válidos toda vez que en muchas organizaciones se emplean políticas de Control de Acceso dinámicas basadas en los atributos de los usuarios y en los recursos y procesos de negocio. Por ello se pasa a sustentar modelos más avanzados como los basados en Atributos (Attribute Based Access Control – ABAC-) y los basados en el estándar XACML (eXtensible Access Control Markup Language).

No obstante Kerberos presenta varios problemas y puntos de mejora, fundamentalmente en lo relativo a la posibilidad de que un atacante externo puede definir un código a través del cual suplantar a los usuarios de servicios Hadoop, y con ello por ejemplo, registrar un falso componente Task Tracker, borrar contenido de HDFS, etc.

Los Data Nodes no establecen un sistema de Control de Acceso, por lo que un atacante puede leer bloques de datos de los Data Nodes o escribir contenido sin sentido que socavaría la integridad y por ende la utilidad de los datos a analizar. También podría incluso enviar una tarea al Job Tracker, que con su ejecución podría afectar al resto de tareas bien por perturbar los resultados que de ellas se obtendrían, o bien simplemente retrayendo recursos que las retrasarían.

3.3.4. Buenas prácticas de Seguridad a implementar en Hadoop

- El mapeo de máquinas virtuales a máquinas físicas debe ser desarrollado de manera muy segura. También la ubicación de recursos y la gestión de memoria.
- Las propias técnicas de Data Mining se pueden emplear para una eficiente detección de malware y ataques en entornos de Big Data y Cloud Computing.
- La seguridad de los datos no sólo afecta al cifrado de los datos, sino que se debe ver acompañada por políticas adecuadas de compartición de datos.
- Se debe tener en cuenta la importancia del los registros (logs). Es crítico para conocer quién y cuándo puede haber borrado o alterado datos de usuario, así como para saber qué nodos se están empleando en el Cluster (y desde cuándo), qué trabajos de MapReduce se están llevando a cabo, qué cambios se derivan de esos trabajos y quién es responsable de esos trabajos. Los logs se revisarán periódicamente de acuerdo con las Políticas de Seguridad de la Organización en cuestión, permitiendo así detectar operaciones maliciosas que buscan el robo o la modificación de información.
- Cualquier nodo que se una al cluster debe autenticarse (empleando técnicas como Kerberos).
- Presencia de Nodos Honeypot, para atraer atacantes, identificarlos e inhabilitarlos.
- Asegurar la publicación segura de datos por terceras partes.
- Potenciar el control de acceso (incluyendo en tiempo real).
- Cifrado de Datos, empleando distintas claves en diferentes máquinas, y teniendo en cuenta que cierta información clave debe ser almacenada y protegida centralizadamente,

para asegurar que incluso si un atacante puede acceder a ciertos datos, será muy difícil que pueda obtener de ellos información relevante.

- Cifrado de red: Toda las comunicaciones que se establezcan en la red debe cifrarse; así por ejemplo las comunicaciones basadas en RPC deben sustentarse en protocolo TLS / SSL (Transport Layer Security / Secure Socket Layer).

4. Métodos de Protección de la Privacidad

4.1. Introducción

En este apartado se presentan los dos enfoques principales para asegurar o al menos incrementar la protección de los datos con carácter general. Se trata de las dos aproximaciones que se están empleando a la hora de proteger los datos tanto de sistemas, empresas y personas en un esquema de gestión de Big Data:

- El primer enfoque se centra en el control de la visibilidad de los datos a través de su protección, limitando la visibilidad de los mismos a través **del Control de Acceso** a los sistemas en los que se manejan / procesan / almacenan.
- El segundo enfoque se basa en la protección del valor que puede obtenerse del análisis del dato, encapsulándolo mediante el **empleo de métodos criptográficos y técnicas de preservación de la Privacidad (Privacy-Enhancing Techniques).**

Tradicionalmente se han empleado los dos enfoques, si bien el primero ha sido el más habitual al ser más simple de implementar (normalmente combinado con comunicaciones protegidas criptográficamente).

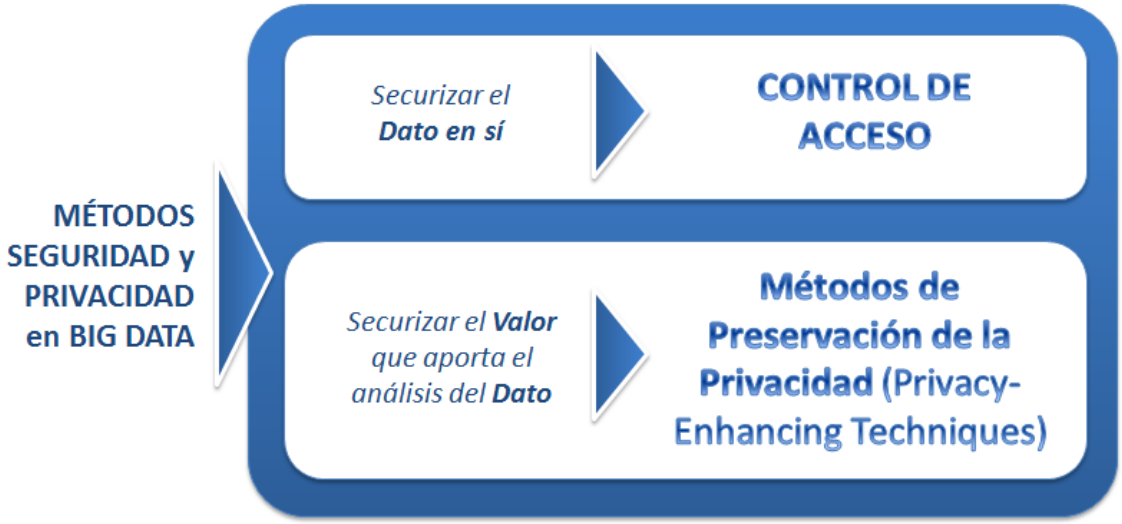


Imagen II-9 – Métodos Protección Seguridad y Privacidad en Big Data

Estos dos enfoques están asociados con el tipo de usuarios que pueden tener acceso a los datos que se gestionan en una plataforma de Big Data. Así, mientras el Control de Acceso busca asegurar que sólo tendrán acceso a los datos y podrán efectuar análisis sobre ellos (a través de solicitud de consultas / trabajos) aquellos usuarios debidamente autorizados para ello (se explicará en detalle este concepto en los siguientes apartados), los métodos de preservación de la privacidad buscan que los usuarios autorizados no puedan, mediante los análisis que pueden realizar y la combinación de los resultados que obtengan con otra

información previa de la que pudieran disponer, disponer de información que por su naturaleza e impacto sobre la privacidad de las personas titulares de los datos, deba ser protegida y no pueda ser revelada.

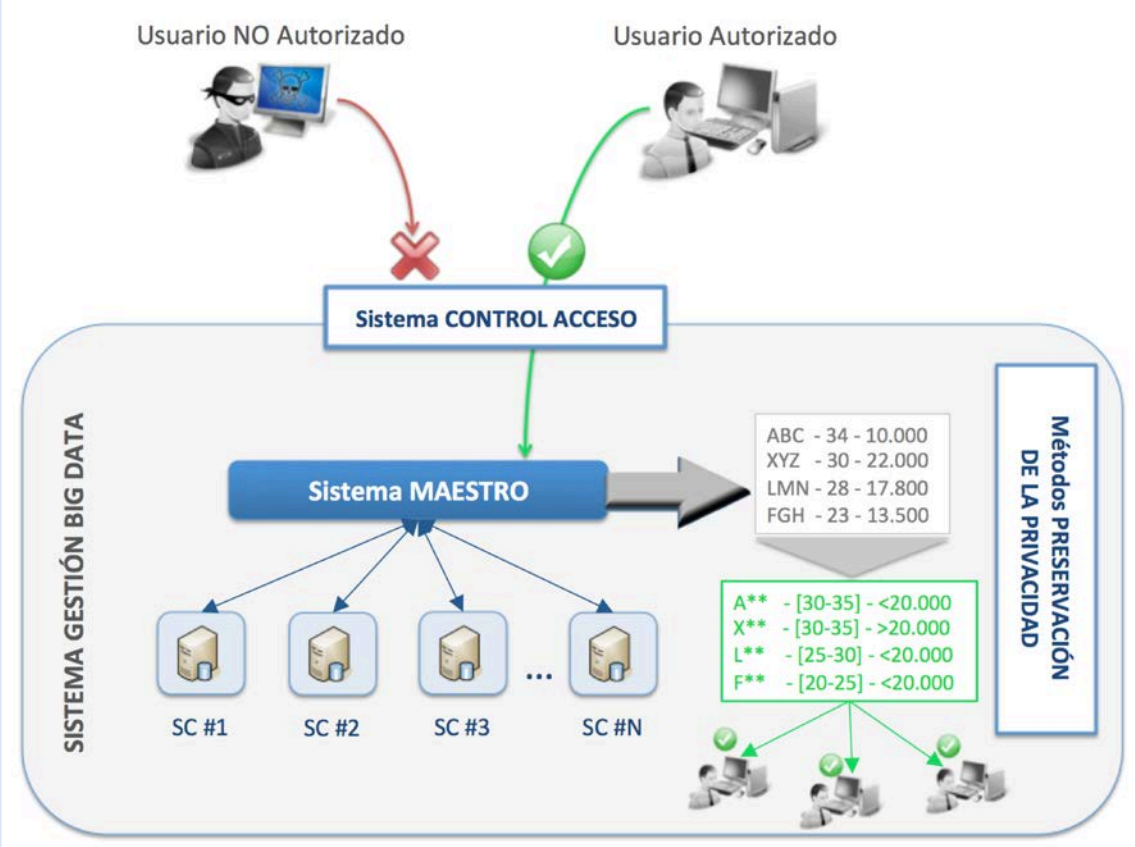


Imagen II-10 – Métodos Protección Seguridad y Privacidad en Big Data. Detalle aplicación en un Sistema de Gestión de Big Data

En los siguientes apartados se presentan el detalle de estos dos métodos y como se aplican a la casuística y los retos particulares que supone el tratamiento de Big Data.

4.2. El Control de Acceso en Big Data – Protección de la Confidencialidad, Integridad y Disponibilidad de los Datos

4.2.1. La relevancia de los sistemas de Control de Acceso y el caso particular de Big Data

Siguiendo a *[49]*, y tal y como ya se ha indicado a lo largo de este proyecto, garantizar la seguridad de la información se encuentra entre los retos más importantes a la hora de implantar los sistemas de gestión de Big Data, dado que los usuarios exigen que en los mismos se garantice un nivel de protección igual al existe en los sistemas de gestión de datos tradicionales.

Una de las técnicas de protección de la seguridad más importante es el Control de Acceso, que permite salvaguardar las distintas dimensiones de la seguridad de la información (Confidencialidad, Integridad y Disponibilidad) de posibles atacantes externos

Los sistemas de control de acceso están entre los más críticos de los componentes de seguridad de red. Es más probable que la privacidad o la seguridad de unos datos se vea comprometida debido a errores en la configuración de las políticas de control de acceso que a errores criptográficos. El refuerzo y gestión de las políticas de control de acceso es una de las técnicas de seguridad fundamentales.

Las políticas en práctica actualmente (no sólo en lo relativo a control de acceso, sino a otros medios de protección de la Seguridad) no están adaptadas al empleo de Big Data y, o bien introducen reglas demasiado estrictas que hacen inviable la obtención de las ventajas que supone el acceso a volúmenes ingentes de datos con nuevas técnicas de gestión y análisis, o representan un serio riesgo de pérdida de datos, debido a que permiten una compartición de datos demasiado poco restrictiva.

El problema es que el control de acceso debe hacerse de una manera lo más precisa posible, para no impedir obtener todo el potencial de los Big Data para personas, empresas u organizaciones que no tienen ningún tipo de intención maliciosa sobre los datos.

4.2.2. Diferenciación Autenticación – Autorización

La **autenticación** no está directamente relacionada con el contenido al que se desea acceder, mientras que en la **autorización** sí.

Por este motivo, y dado que como se ha indicado, el objetivo del control de acceso es proteger los datos que se manejan en los sistemas de Big Data, el foco se centrará en este caso en la Autorización. En ella es fundamental que exista una sincronización entre las credenciales de acceso para el Sistema Maestro y los Sistemas Cooperantes, siendo la misma más compleja que en entornos no-Big Data.

4.2.3. Problemas / Retos

Existen una serie de retos que deben ser abordados en relación con la definición de sistemas de control de acceso eficientes en sistemas de gestión de Big Data *[10]*:

- Se deben hacer coexistir las políticas de control de acceso de múltiples organizaciones, cuyos datos se integran en sistemas Big Data. Para cada uno de estos orígenes diferentes se deben cumplir sus propias políticas de control de acceso, incluso cuando los datos se integren con otros de un origen diferente, y que por tanto deben cumplir con una política distinta.
- En este punto surge el reto de evitar que una política que se puede haber definido de un modo excesiva e injustificadamente restrictivo, pueda afectar a la posibilidad de obtención de resultados, y por ende a la utilidad del sistema Big Data.
- El volumen de peticiones de acceso a datos y de ejecución de peticiones que se da en un sistema de Big Data hace imposible que se puedan autorizar de manera individualizada cada una de ellas.
- Los enfoques tradicionales basados en seguridad perimetral no son válidos por si solos para securizar un sistema Big Data (por ejemplo el mapeo de direcciones IP con credenciales de gestión de identidades, al requerir un diseño de red específico).
- Los enfoques de granularidad gruesa (coarse-grained) que se han venido empleando tradicionalmente para el control de acceso impiden que datos que en principio sí podrían ser accedidos / compartidos dejan de serlo al situarse en categorías más restrictivas en aras a una mayor protección de la privacidad. Por este motivo es necesario plantear

enfoques de sistemas de control de acceso más granulares, que den a los propietarios de los sistemas más precisión al compartir los datos sin comprometer la confidencialidad.

- En Big Data se manejan, tal y como ya se ha expuesto en el *Capítulo I*, datos muy diversos tanto en términos de tipos como de requisitos de seguridad involucrados.
- El cumplimiento de los múltiples requisitos legales existentes (Normativa, Acuerdos Corporativos, Políticas de Privacidad, etc.) hacen que cada vez se restrinja más el acceso a los datos y sean menos las personas que pueden hacer análisis, reduciéndose las posibilidades de obtener resultados y ventajas de los Big Data. Por todo ello se requieren enfoques de control de acceso más granulares.
- Se hace más complejo y caro el desarrollo de sistemas y aplicaciones que emplean Big Data, al tener que implementar mecanismos de control de acceso granulares.
- Se debe establecer una política de permisos para el sistema maestro y los sistemas cooperantes, para que únicamente puedan acceder a aquellos contenidos para los que hayan sido previamente autorizados.
- Las técnicas de control de acceso tradicionales se sustentan en el uso de Bases de Datos basadas en SQL que permiten control de acceso de solicitudes de datos asignando las columnas y filas con atributos de seguridad, cumpliendo de este modo con modelos de control de acceso basado en atributos (ABAC).
  - Control sobre acceso a columnas: Campos que se requiera proteger (por ejemplo derivados del cumplimiento de normativa)
  - Control sobre acceso a Filas: Entidades o usuarios cuya protección es imprescindible.
  - Control sobre acceso a celdas: El estadio final de máxima granularidad; en la mayor parte de los casos será imposible garantizarlo dada la inmensa cantidad de datos que existen y la carga computacional que requeriría (conllevaría problemas como retardos por encima de lo permitido/aceptado).

  Sin embargo en el entorno de Big Data se emplean bases de datos NoSQL para manejar datos no estructurados, sin un esquema predefinido.

4.2.4. Enfoque General para el Control de Acceso en Big Data

4.2.4.1. Esquema General de Control de Acceso en Sistemas Big Data

El control de acceso en Big Data no sólo debe tener en consideración los datos que salen del nodo maestro, sino también el acceso a los distintos sistemas cooperantes. De este modo el control de acceso al Sistema Maestro de un cluster de Big Data, debe complementarse con las consideraciones en relación con el control de acceso a los Sistemas Cooperantes.

Los siguientes componentes se consideran fundamentales a la hora de establecer un sistema de control de acceso en Big Data:

- **Acuerdos de Seguridad entre Sistemas Maestros y Proveedores de Big Data**.

  Permiten clasificar los Big Data puestos a disposición del sistema por el proveedor de Big Data en distintas clases de seguridad. En función de estas clases de seguridad, el Sistema Maestro junto con cada Sistema Cooperante podrá determinar qué nivel de seguridad estará habilitado para manejar cada Sistema Cooperante.

- **Listas de Sistemas Cooperantes de Confianza**.

  Están basadas en los acuerdos de Seguridad y en las categorías de Sistemas Cooperantes establecidos en función de su Clase de Seguridad por el tipo de datos que pueden manejar. Estas listas son gestionadas por el Sistema Maestro en base a su conocimiento de los Sistemas Cooperantes autorizados a trabajar en su Cluster.

- **Política de Control de Acceso del Sistema Maestro**.

  Esta política determina las reglas que deben cumplir los Sistemas Cooperantes para poder acceder a operar los datos, por ejemplo en función de que cumplan o no determinados atributos (en un enfoque de control de acceso basado en atributos –ABAC-).

- **Políticas de Control de Acceso de los Sistemas Cooperantes**.

  Definen cómo se realizará el acceso a los procesos y datos distribuidos teniendo en cuenta parámetros como la capacidad de proceso disponible (la carga del sistema en cada momento) y los requisitos de seguridad que debe cumplir cada Sistema Cooperante. Por ejemplo se establecerá que los Big Data distribuidos no podrán almacenarse en espacios de disco compartidos con otros usuarios locales del Sistema Cooperante que no estén involucrados en la gestión de los Big Data.

- **Definición de Atributos Federados**.

  Se trata de los atributos comunes empleados en el Sistema Maestro y los Cooperantes, etc., para que las Políticas de Control de Acceso del Sistema Maestro y de los Sistemas Cooperantes sean interoperables.

La siguiente imagen muestra la relación entre estos distintos componentes, con un ejemplo de cómo se aplicarían en un cluster de Big Data:

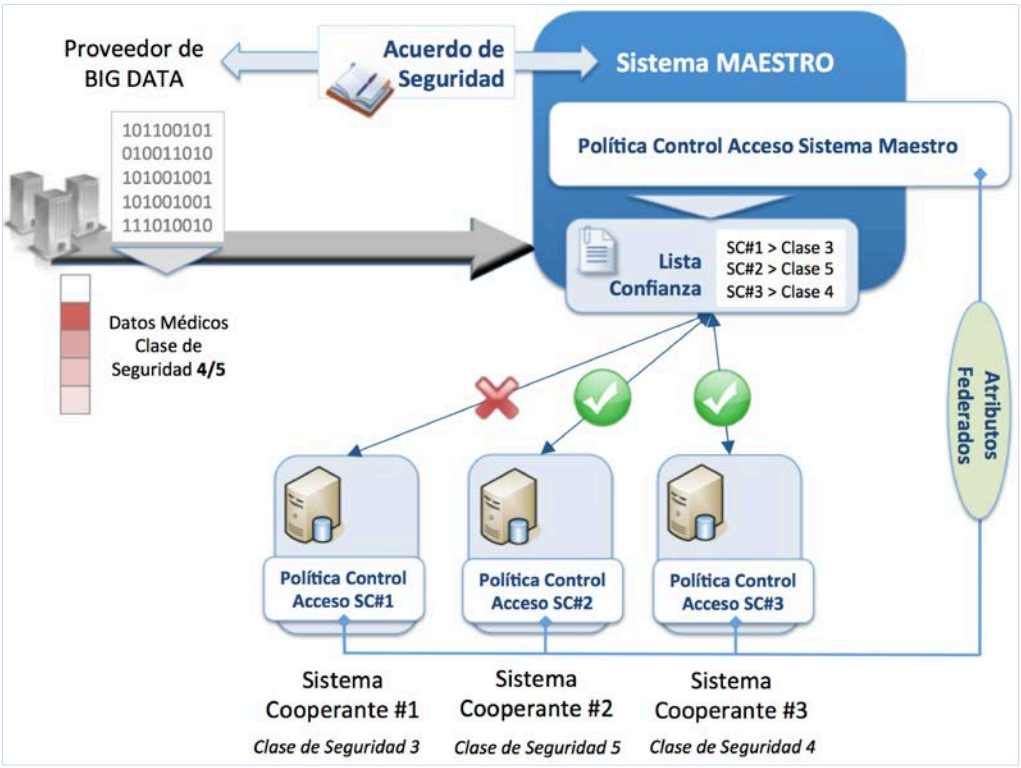


Imagen II-11 – Componentes Sistema de Control de Acceso en Big Data

Para poder establecer un esquema como el anterior es fundamental una estrecha coordinación entre los diferentes elementos implicados (Proveedor de Big Data, Sistema Maestro y Sistemas Cooperantes), así como que el Sistema Maestro pueda recopilar toda la información necesaria sobre los Sistemas Cooperantes (capacidades de seguridad y grado de confianza que se puede esperar de ellos). Esta información permitirá al Sistema Maestro, en coordinación con los proveedores de Big Data, establecer las clases de seguridad implicadas y las correspondientes Listas de Confianza.

La coordinación también permitirá la definición de los atributos federados que posibilitarán que el Sistema Maestro y los Cooperantes sean interoperables y por tanto las listas de confianza sean los más cercanas a la realidad, al poder evaluarse las capacidades de seguridad y protección sobre la base de un lenguaje común y comparable.

Además también se requiere que los Sistemas Cooperantes determinen la manera en que gestionarán los posibles conflictos entre su Política de Control de Acceso y la propia que pudieran tener definidos localmente cada uno de los Sistemas Cooperantes (no se debe olvidar que los sistemas cooperantes son equipos que forman parte de uno o varios clusters de Big Data y que además pueden operar en otros entornos independientes).

Además, en línea con las buenas prácticas de Seguridad de la Información, toda la actividad relativa a las autorizaciones que se concedan y el control de acceso, deberá ser controlada por las necesarias capacidades de auditoría.

4.2.4.2. Mecanismos Complementarios

Tal y como se ha indicado, en el *apartado 4.2.3.*, resulta imposible autorizar individualmente cada una de las solicitudes de acceso a datos o de ejecución de peticiones que se dan en una plataforma de Big Data.

Por ello se hace necesario analizar e implementar en su caso, las buenas prácticas que se están definiendo para incrementar la eficiencia y efectividad de los mecanismos de control de acceso *[19] y [68]*. Entre ellas destacan:

- **Mecanismos de administración automática de autorizaciones**, los cuales se pueden basar en aspectos como el **Control de Acceso Basado en Atributos** (Attribute-based Access Control –ABAC-):
  - La identidad del usuario.
  - Su perfil de usuario.
  - El contexto en que se produce la solicitud de acceso a datos.
- **Técnicas de Autenticación Continua**, que amplían, en combinación con el control de acceso basado en atributos, las posibilidades de autenticación de usuario para dar cabida a información como las pulsaciones de teclado o de ratón del usuario para verificar constantemente la identidad del usuario.
- **Políticas de Control de Acceso Basado en el propio contenido** de los datos a los que se desea acceder y los metadatos involucrados. Se trata de un mecanismo de especial relevancia a la hora de asegurar la adecuada protección de la privacidad de los usuarios cuyos datos de manejan en los sistemas de Big Data.

  Como reto asociado al control de acceso basado en contenidos, cabe destacar que requiere un profundo conocimiento de los contenidos a proteger, lo que puede llegar a ser complicado cuando se trabaja con contenidos no estructurados en forma de base de datos

relacionales (como por ejemplo datos derivados de redes sociales o contenidos multimedia).

Debido a la importancia de proteger los datos que requieren una protección especial, y en tanto en cuanto no se disponga de mecanismos eficientes de control de acceso basado en contenidos, como una primera medida se ha planteado la posibilidad de **separar los datos que requieren una protección especial en servidores independientes establecidos ad-hoc**, creando en definitiva un segundo cluster de Big Data para datos sensibles.

- **Políticas de Control de Acceso que se diseñen, evolucionen y se gestionen automáticamente**, en función de las circunstancias de cada caso. Se trata de un aspecto crítico para asegurar la confidencialidad, integridad y disponibilidad de los datos, a la hora de tratar con entornos dinámicos, donde los tipos de datos, usuarios y aplicaciones involucradas cambian constantemente.
- **Control de Acceso en Tiempo Real**, que constituirá una buena medida de seguridad en entornos de cloud computing.

4.3. Métodos de Preservación de la Privacidad - Las Privacy Enhancing Techniques y las más óptimas para el caso de Big Data.

4.3.1. Métodos Tradicionales de Preservación de la Privacidad: Cifrado frente a Anonimización

Los métodos más utilizados hasta ahora para la protección de la privacidad, se centran en el **empleo de la criptografía y en técnicas de anonimización de datos**. Sin embargo estas medidas tradicionales que se han venido empleando para la protección de la privacidad de los datos personales, pueden dejar de ser válidas en algunas circunstancias en el ámbito de la gestión de Big Data, al menos con los enfoques que se han venido utilizando hasta ahora.

Por este motivo es necesario reenfocar estas soluciones para adaptarlas a la casuística especial de los Big Data *[105]*.

4.3.1.1. La Criptografía como método para proteger la Privacidad.

La Criptografía se ha empleado tradicionalmente como un enfoque para la protección de datos. Sin embargo los enfoques de criptografía que se habían venido utilizando, no permiten por si solos asegurar la privacidad y la protección de datos en entornos de Big Data. Es una técnica robusta, pero el procesado eficiente de datos cifrados es una tarea exigente y que demanda un consumo de recursos que puede no ser siempre compatible con los tiempos de respuesta que se demandan en análisis asociados a la gestión de Big Data.

Es precisamente este bajo rendimiento computacional, una de las principales restricciones a la hora de plantear su empleo en enfoques de gestión de Big Data.

Además, tal y como se destaca en *[45]*, la criptografía hace inaccesibles los datos para todos aquellos usuarios o analistas que no disponen de la clave de descifrado, lo que puede hacer que muchos potenciales usuarios que pudieran obtener beneficios de la explotación de los Big Data no puedan hacerlo. Además se acrecientan los riesgos derivados de una posible pérdida / robo de la citada clave o de que se produzca una violación de los datos antes de que sean cifrados *[28]*.

En los últimos años, se están efectuando investigaciones para el desarrollo de protocolos y técnicas de cifrado protectoras de la privacidad. Entre ellas cabe mencionar:

- **Protocolo de Recuperación de Información Privada (Private Information Retrieval –PIR-)**. Permite a un usuario de un sistema de Gestión de Big Data o de Cloud Computing hacer peticiones en relación con sus datos, sin que puedan ser conocidas por el propio sistema.
- **Cifrado habilitador de búsquedas**. Permite a un usuario de un conjunto de documentos autorizar a una tercera parte el desarrollo de búsquedas en relación con una serie de palabras clave especificadas con carácter previo, sin que exista la posibilidad de revelación de ninguna información adicional.

No obstante, la técnica que ha suscitado el mayor desarrollo, dada su potencialidad ha sido la **Criptografía Homomórfica (Fully Homomorphic Encryption –FHE-)**.

Tal y como se destaca en *[108]*, la criptografía homomórfica permite la computación de datos cifrados. El cliente envía al sistema un dato cifrado, el sistema es capaz de analizarlo y procesarlo sin necesidad de descifrarlo y devolver un resultado que sólo tendrá sentido para el usuario inicial que aportó el dato, una vez que él/ella lo descifre convenientemente. Se trata por lo tanto de un sistema que en teoría garantiza la máxima protección de la privacidad, pero únicamente en un esquema de relación cliente-servidor, que en un contexto de gestión de Big Data se ve ampliamente superado.

Permite que las funciones mediante las cuales se desarrollan las operaciones a ejecutar, se computen sobre datos cifrados, sin necesidad de descifrarlos primero. De este modo, la aplicación de una función sobre un dato original sin cifrar, equivale a la aplicación de otra función (que no debe ser obligatoriamente la misma) sobre el mismo dato pero en su versión cifrada. De esta manera, si se aplican funciones sobre datos cifrados y posteriormente se descifra el resultado, se obtendría lo mismo que si se ejecutasen esas mismas funciones sobre los datos originales.

De manera más formal, considerando $E_k(m)$ la versión cifrada de un dato m mediante una clave k, un esquema de cifrado es homomórfico respecto a una función f, si existen una función f' tal que:

$$D_k\,(f'(E_k\,(m)))\;=\;f\,(m)$$

siendo D_k el algoritmo de descifrado mediante la misma clave k.

Desde la presentación inicial de la criptografía homomórfica completa por parte de Craig Gentry de la Universidad de Stanford en su tesis de 2009 *[42]*, se han desarrollado muchos otros trabajos enfocados en la aplicación práctica de estos sistemas. Si bien el FHE permite que se desarrolle una amplia variedad de operaciones sobre datos cifrados, dada la necesidad de emplear datos complementarios y cabeceras muy grandes, su aplicación a casos prácticos supone unos tiempos de ejecución muy altos que no son factibles, salvo para problemas muy pequeños. Evidentemente, este hecho hace que en la actualidad su empleo para la protección de la Privacidad en sistemas Big Data sea aún inviable.

Las siguientes gráficas muestran algunas métricas que dan muestra del tiempo que supone el uso de la Criptografía Homomórfica en el desarrollo de operaciones sobre grandes volúmenes de datos (fuente *[107]*):

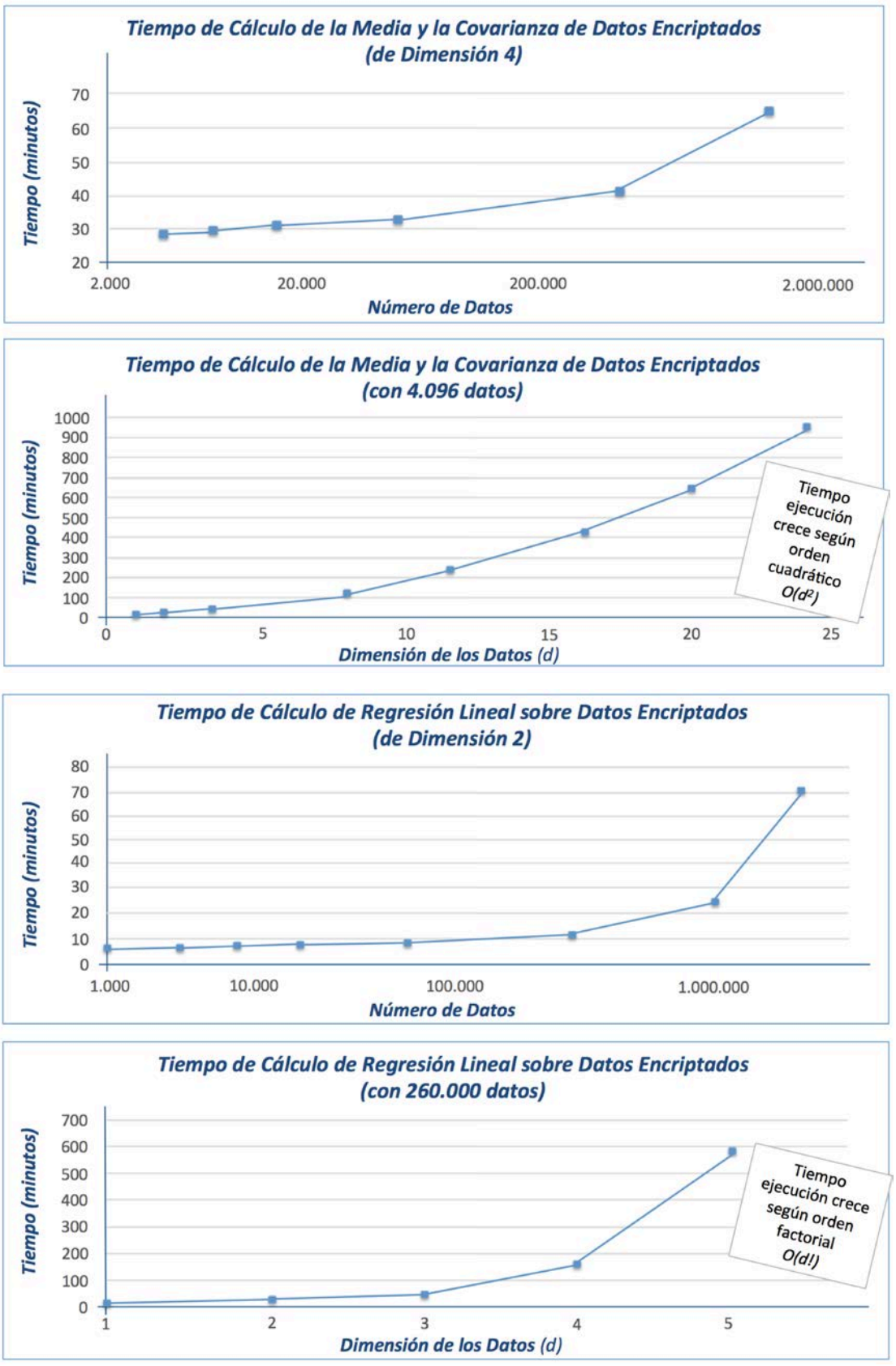


Imagen II-10 - Tiempo ejecución operaciones estadísticas sobre grandes volúmenes de datos empleando Criptografía Homomórfica

Un análisis más detallado de los fundamentos matemáticos de la FHE se puede encontrar por ejemplo en el ya mencionado *[42]*, o en *[86]*.

A la vista de los datos anteriores, cabe concluir que las últimas tendencias en cifrado se han demostrado eficientes para la protección de la Privacidad de un entorno de análisis y computación de los datos de un usuario individual, pero dejan de serlo en el momento en que se gestionan datos interrelacionados de diferentes usuarios, o casos en que se busca la obtención de estadísticas de datos de diferentes personas (que como se ha venido destacando ya, son los entornos más comunes en la gestión de Big Data, dado que en ellos se aprovecha realmente toda sus potencialidades. Incluso con herramientas tan potentes como la FHE, la criptografía no puede asegurar por si sola la protección de la privacidad que se demanda en entornos de gestión de Big Data y de Cloud Computing.

De este modo la criptografía, a día de hoy y por si sola, no puede considerarse como un método de protección de la privacidad en Big Data, sino simplemente un complemento para el desarrollo de otros métodos como la anonimización de datos.

No obstante, se están proponiendo diversos métodos y técnicas de computación que permiten, en determinados contextos, el desarrollo de análisis sobre datos cifrados cada vez más eficientes, usando así mismo peticiones cifradas, que afectan cada vez menos a la utilidad de los datos. Entre ellos cabe mencionar por ejemplo el presentado en *[67]*.

Evidentemente se trata de un campo de trabajo en el que los avances que se produzcan permitirán métodos cada vez más eficientes de protección de la privacidad, por lo que tal y como se destacará en mayor detalle en el *Capítulo V*, se trata de ámbitos de actividad sobre los que se debe seguir prestando atención y explorando nuevas posibilidades de avance.

4.3.1.2. Los métodos Estadísticos de Anonimización como método para proteger la Privacidad.

4.3.1.2.1. Introducción

Otro de los métodos tradicionales de protección de la Privacidad, son los métodos Estadísticos de Anonimización, como un punto intermedio que permite, en teoría, asegurar tanto la utilidad de los datos como la preservación de la privacidad *[45]*.

Por anonimización de datos se entiende el proceso para cambiar o eliminar datos que van a ser publicados, de un modo que previene la identificación de información sensible. Se emplea como sinónimo de deidentificación de datos. Gracias a la Anonimización es posible que las Organizaciones que gestionan los datos puedan publicarlos. Normalmente aplica a los identificadores directos (ver definición en el *Capítulo I*).

La anonimización es una técnica que se emplea para incrementar la seguridad de los datos, al mismo tiempo que se permite que dichos datos puedan seguir siendo empleados para análisis.

Tras ser anonimizado, en la mayoría de los casos, el dato continuará pareciendo real (a diferencia de lo que ocurre con el cifrado).

A la hora de trabajar con los Métodos de Anonimización (tanto en los métodos estadísticos presentados en el presente apartado, como en los métodos grupales que se expondrán más adelante), es necesario establecer una serie de definiciones de referencia, que complementan a las genéricas presentadas en el *Capítulo I* *[5]*, *[84]*, *[102]*.

- **Microdatos:** Son los datos que se publican, una vez que se han aplicado los métodos de anonimización sobre los valores originales.

- **Identificadores directos o explícitos**: Se trata de aquellas variables que describen una característica de una persona que es observable, que está registrada o en general que puede ser conocida, como por ejemplo nombres, direcciones o números de identificación. Permitirían la identificación directa de una persona pero no son necesarios para propósitos estadísticos o de investigación, por lo que pueden ser eliminados del conjunto de datos publicado.
- **Identificadores indirectos o cuasi-identificadores**: Son los atributos o combinaciones de atributos, que pueden ser compartidos por diferentes personas, y cuya combinación puede conducir a la reidentificación de una de ellas.
- **Atributos sensibles**: Son los atributos que contienen información relativa a una persona específica y por su naturaleza deben tener una protección especial (por ejemplo en España, en la *Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal*, se identifican en su *artículo 7* como "*datos especialmente protegidos*" aquellos relativos a religión, ideología, creencias, afiliación sindical o política, origen racial o étnico, salud o vida sexual). Estos atributos sensibles no deben ser revelados (al menos asociándolos a una persona en concreto) en los microdatos publicados.
- **Clase de Equivalencia**: Se trata de aquel conjunto de registros que no se diferencian entre ellos con respecto a sus cuasi-identificadores.

Anonimizar los datos consiste en determinar qué variables son identificadores potenciales y en modificar el nivel de precisión de esas variables para reducir el riesgo de reidentificación a un nivel aceptable. El objetivo, de este modo, es maximizar la seguridad mientras se minimiza la pérdida de información.

En todo caso, y tal y como se expondrá en mayor detalle en los siguientes apartados, en lo relativo a la anonimización existen una serie de consideraciones que deben ser tenidas en cuenta, al afectar de manera muy relevante a la efectividad de la protección de la privacidad:

- Posibilidad de acceso de un hipotético atacante a fuentes adicionales de información (públicas o privadas): Al aplicar las técnicas de anonimización, el responsable de la gestión de una base de datos puede no saber qué información adicional está disponible en otras bases de datos, y cómo ésta puede ser empleada para inferir datos sensibles, de la información anonimizada.
- Posibilidad de acceso de un hipotético atacante a nuevas técnicas de gestión de Big Data: Estas técnicas incrementan en un modo muy considerable, las posibilidades de búsqueda de correlaciones entre datos en principio no relacionados, pudiendo en determinados casos dar lugar a la reidentificación de un dato en principio anonimizado.

Se incluyen a continuación diferentes métodos estadísticos de anonimización que existen para proteger o al menos incrementar la privacidad. Para describirlos se tomará el ejemplo de una base de datos ficticia, que podría ser empleada por una aseguradora médica y una serie de registros concretos para ella *[5]*, *[85]*:

Nº Reg.	Nº Seguridad Social	Nombre	Edad	Sector Puesto de trabajo	Salario Bruto (€/año)	Enfermedad
1	NS-0021-7684	Ana López	40	Medicina	36.000	Bronquitis
2	NS-1107-4730	David Sánchez	29	Informática	24.000	Neumonía
3	NS-0870-0002	Carlos Muñoz	32	Marketing	35.000	Gripe
4	NS-1107-9341	Lucía Pérez	24	Diseño Grafico	24.000	Gastritis
5	NS-0870-0203	Laura González	29	Enfermería	29.000	Úlcera estomacal
6	NS-1107-6151	María Martín	45	Gerencia Negocio	110.000	Bronquitis
7	NS-0870-4005	Jesús Rodríguez	32	Abogacía	34.000	Neumonía
8	NS-0021-9573	Sonia Hernández	32	Docencia	29.000	Cáncer Estómago
9	NS-0021-3464	Juan García	40	Consultoría	34.000	Gripe

Tabla II-1 – Registros originales

En la tabla anterior:

- Identificadores Directos: **Nº Seguridad Social, Nombre[*].**
- Identificadores Indirectos o Cuasi-identificadores: **Edad, Sector Puesto de Trabajo**
- Atributos Sensibles: **Salario Bruto y Enfermedad**

4.3.1.2.2. Métodos basados en la reducción de datos

Estos métodos permiten reducir el número de individuos de una muestra que comparten características, buscando eliminar la posibilidad de que pueda identificarse a una persona concreta.

Como métodos más reseñables en esta categoría destacan (se marcan con fondo verde los campos sobre los que se acometen modificaciones respecto a los registros originales de la *Tabla II-1*):

- **Supresión / Eliminación de variables**: en concreto se eliminan los identificadores directos. Se trata de una técnica que afecta mucho a la utilidad de los datos y sólo debe ser empleada cuando ningún otro método de protección pueda ser aplicado (por ejemplo por tratarse de una variable altamente identificativa de una persona). También se puede suprimir cuando no aporta demasiado valor a efectos de los análisis que se pretenden llevar a cabo.

3	NS-****-****	Carlos *******	32	Marketing	35.000	Gripe

[*] *El nombre puede ser compartido por varias personas, pero a efectos de este análisis se considerará un Identificador Directo.*

Es el método más simple y rápido, pero es el que por el contrario supone una mayor pérdida de información. En este caso, anulando el campo número de Seguridad Social, se impide una identificación inequívoca de la persona.

Este método se puede emplear en caso de que no sean imprescindibles los campos de identificadores directos, por ejemplo cuando se vaya a realizar un análisis estadístico, donde se prioricen la obtención de datos y tendencias agregadas.

- **Supresión / Eliminación de registros**: Se elimina la información completa de una persona en una base de datos. Es el método más extremo y por lo tanto debe ser reservado sólo a los casos excepcionales donde ningún otro método ofrece las debidas garantías de protección, y además la revelación de la identidad de la persona tuviera consecuencias graves. Por ejemplo su uso podría justificarse para el caso de una persona que sufriera de una enfermedad rara (que afecta por ejemplo sólo a 1 persona de cada 1 millón).
- **Generalización o Recodificación global**: Se trata de un método válido para variables numéricas, en el que éstas de generalizan al incluirse en categorías menos específicas. Así por ejemplo, se podrían reagrupar las edades en grupos de 5 (entre 25 y 30 – entre 31 y 35,…), o los números de la seguridad social, dejando por ejemplo los números iniciales.

3	NS-0870-****	Carlos Muñoz	30-35	Marketing	30.000 – 40.000	Gripe

- **Codificación por arriba y por abajo**: Consiste en agregar un valor genérico para enmascarar los valores atípicos, por ser anormalmente altos o bajos, que son los que pueden llevar con más probabilidad a una reidentificación.

6	NS-1107-6151	María Martín	45	Gerencia Negocio	Superior a 36.000	Bronquitis

Con carácter general, los métodos basados en la reducción de datos, tienen las siguientes desventajas:

- Son más vulnerables a los ataques de homogeneidad (aquellos en los que no existe una diversidad suficientemente significativa entre los datos sensibles) o cuando el atacante dispone de un conocimiento de información adicional a la presentada en la base de datos anonimizada de la persona.
- Pueden suponer una disminución de la utilidad de los datos en los casos de reducción más drástica (como la eliminación de registros).

4.3.1.2.3. Métodos basados en la introducción de perturbaciones / modificaciones en los datos

Estos métodos se basan en la modificación de los datos originales para dificultar una posible reidentificación, e incluso cuando ésta se pueda producir, el atacante no podían estar completamente seguro de que los datos coincidan con los reales / originales.

- **Reordenación aleatoria**: Consiste en redistribuir los valores de un determinado atributo entre los diferentes registros. Por ejemplo en la *Tabla II-1* ejemplo, aplicando este método sobre el campo Salario Bruto impediría que pudiera conocerse el sueldo real de una persona concreta, pero al mismo tiempo se seguiría conservando la posibilidad de desarrollar análisis de datos agregados reales.

…

7	NS-0870-4005	Jesús Rodríguez	32	Abogacía	29.000	Neumonía
8	NS-0021-9573	Sonia Hernández	32	Docencia	34.000	Cáncer Estómago

…

- **Sustitución**

  Consistiría en sustituir el valor de un campo de un registro por un valor ficticio diferente. Por ejemplo el registro 6 de la Tabla 2, podría quedar como sigue:

6	NS-1107-7777	Cristina Jiménez	45	Gerencia Negocio	110.000	Bronquitis

  Existiría una tabla de sustitución en el que se guardaría la correspondencia entre el registro original y el sustituido. A efectos estadísticos se mantienen todas las propiedades para análisis, si bien se pierde la posibilidad de una asociación personalizada (por ejemplo para casos de marketing).

- **Empleo de medias estadísticas**: Nuevamente se trata de un método que podría aplicarse para los valores de campos numéricos. En este caso se sustituyen los valores concretos por la media de ese campo en un conjunto de registros. En el ejemplo de la *Tabla II-1*, se sustituirían los datos de salarios que estén entre 20.000 y 30.000 € por 26.500 € (la media estadística de 24.000, 24.000, 29.000 y 29.000).

4.3.1.2.4. Retos y problemas de los métodos estadísticos de anonimización.

Algunos autores estiman que no existen suficientes evidencias que demuestren que la anonimización, al menos en el modo en que está siendo aplicada actualmente, funcione completamente ni en teoría ni en práctica (siempre y cuando al mismo tiempo no se reduzca enormemente la utilidad de los datos, hasta hacerlos prácticamente no aptos para cualquier tipo de análisis) *[69]*. Los métodos para cuantificar la eficiencia de los métodos de anonimización no son absolutamente exactos, dado que hacen unas suposiciones no completamente demostradas de lo que un atacante externo realmente conoce y puede llegar a hacer.

En concreto existen dos factores que reducen la capacidad de protección de la privacidad que proveen los métodos de anonimización:

- La información que los usuarios publican conscientemente en las redes sociales juegan precisamente un papel muy importante a la hora de reducir la eficiencia de la anonimización como método de protección de la privacidad, dado que se dan a conocer a hipotéticos atacantes externos datos que de otra forma le serían muy difíciles de conseguir, como por ejemplo información de localización geoespacial.
- Nuevas posibilidades de inferencia de información derivada de la explotación y análisis de Minería de Datos sobre Big Data. Por ejemplo empleando información pública, como el registro de votantes que empleó Latanya Sweeney, en el caso expuesto en el *Capítulo I* u otras fuentes de información como Wikipedia (en la que por ejemplo se puede conocer el nombre, sexo, edad, y código postal de por ejemplo el Gobernador del Estado de Nueva York, disponiéndose así de una información similar a la que Latanya Sweeney obtuvo del registro de votantes, de manera aún más accesible si cabe).

El acceso a Big Data (estructurados o no) abren una nueva posibilidad de análisis que llevan por ejemplo a que una persona pueda, con un porcentaje significativo que puede superar el 50% *[110]*, ser reidentificada únicamente conociendo dos puntos de localización espacio-temporal, cuando en el plano teórico hasta ahora se había venido defendiendo que deberían ser 4. Así dos puntos típicos como el domicilio y el lugar de trabajo pueden permitir la reidentificación, dado que se trata de información única en un elevado volumen de casos.

Por todo ello se demuestra como la anonimización, como método para proteger la privacidad en un contexto de gestión de Big Data, y siempre con la lógica intención de mantener lo más alta posible la utilidad de los datos, debe ser revisada, adaptada y completada por nuevas técnicas, que incrementen la protección al mismo tiempo que se garantice un nivel suficientemente aceptable de utilidad de los datos, y siempre teniendo en cuenta todas las posibilidades de reidentificación.

Incluso los defensores de la anonimización reconocen que es imprescindible, sobre todo a la vista de las posibilidades que ofrece el análisis de Big Data y registros de datos multidimensionales, complementar los métodos de anonimización con por ejemplo acuerdos legales, en los que las personas con acceso a datos personales (aun estando anonimizados) se comprometan a hacer un uso de ellos que se limite al propósito para el que se les autoriza el conocimiento. Lógicamente los registros de datos multidimensionales son en la actualidad la norma (por ejemplo la mediana del número de amigos de un usuario de Facebook se encuentra en 200 amigos *[89]*, lo que dispara los datos generados en relación con dicho usuario a las 200 dimensiones).

Además las técnicas tradicionales de anonimización no contemplan que la generalización que emplean puede hacer inservible para ciertos análisis lícitos de la información a publicar, especialmente en entornos multidimensionales, como las ya comentadas redes sociales. Por este motivo en estos casos no es factible su empleo.

No se puede establecer a priori y con exactitud una tasa de la posibilidad de reidentificación de una serie de datos anonimizados, ya depende del conocimiento que un atacante pueda tener, que lógicamente es imposible de conocer.

Por todo lo anterior, se presentó la necesidad de evolucionar los métodos estadísticos de anonimización, abordando sus debilidades, bien con otros métodos estadísticos de protección de la privacidad (como la aleatorización o la generación de datos sintéticos) o con otros enfoque como los métodos de anonimización grupales o el control de los resultados de consultas. Estos métodos y enfoques de Preservación de la Privacidad se analizan en el siguiente apartado.

4.3.2. Métodos de Preservación de la Privacidad

En este subapartado se presenta una clasificación de los métodos de preservación de la privacidad que se están empleando hoy en día en el ámbito de la gestión de los Big Data *[24]*.

Estos métodos se clasifican en función de la forma que tiene la información a proteger, esto es, si se encuentra disponible en forma de bases de datos estructuradas o en algún tipo de forma no estructurada (texto, información derivada de redes sociales, información obtenida de dispositivos inteligentes, etc.). La siguiente imagen resume esta clasificación:

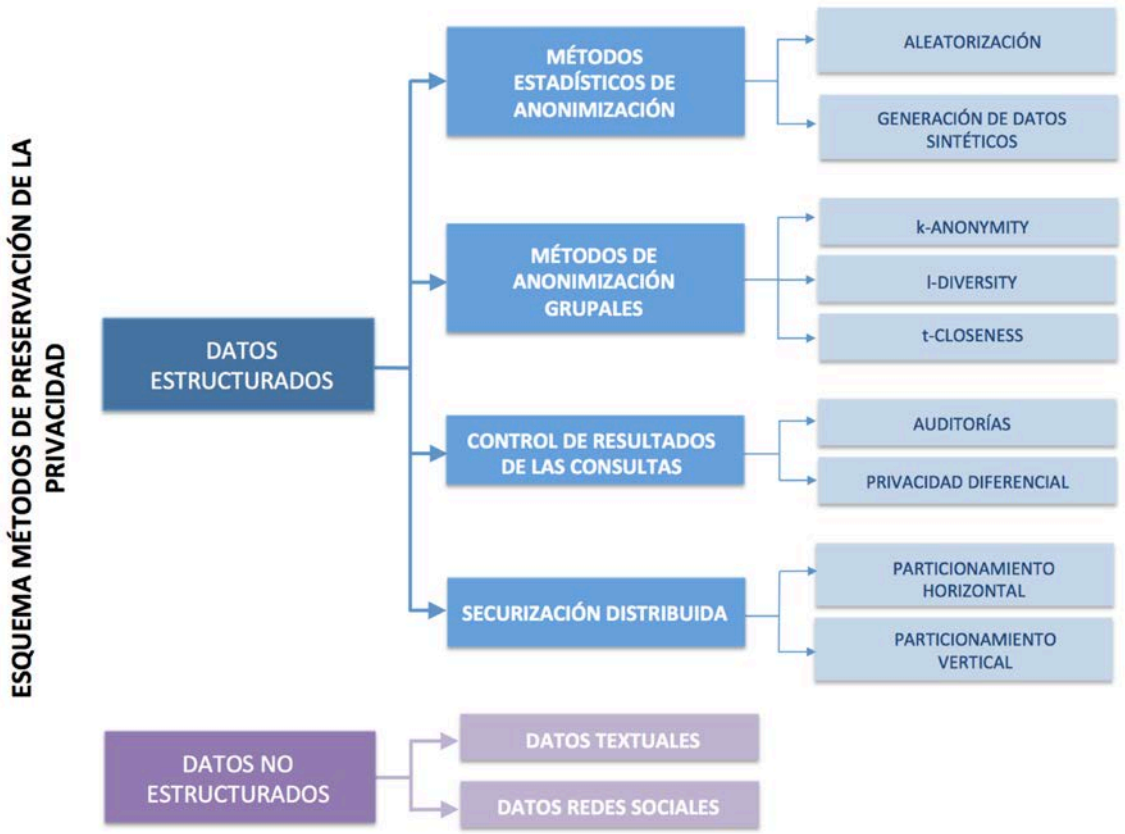


Imagen II-12 – Esquema Métodos de Preservación de la Privacidad

En los siguientes subapartados se da una descripción de cada uno de estos métodos, destacando las fortalezas y debilidades de cada uno de ellos y los casos en que su empleo supone mayores beneficios en términos de protección de la privacidad y al mismo tiempo de utilidad y uso eficiente de los Big Data.

4.3.2.1. Métodos Estadísticos de Anonimización

4.3.2.1.1. Aleatorización

- **Métodos Aditivos**

  En este método se añade ruido, generalmente de media 0 y varianza predefinida σ^2, a la muestra original [4].

  De este modo sobre una serie de registros $X = \{x_1...x_N\}$, y para cada uno de esos registros $xi \in X$, se añade un componente de ruido. Estos componentes de ruido, $Y = \{y_1...y_N\}$, se definen de manera independiente unos de otros, siguiendo una distribución de probabilidad $f_Y(y)$. Con ello los nuevos registros modificados tienen unos valores $x_1 + y_1... x_N + y_N$, que se denotan por $Z = \{z_1...z_N\}$.

  La variación que añade el ruido debe ser lo suficientemente alta como para que los valores originales no puedan ser fácilmente averiguados a partir de los registros Z. De este modo los registros originales no se pueden recuperar, pero sí su distribución (con el comportamiento de X), con los beneficios que ello conlleva en términos de mantenimiento de la posibilidad de elaborar análisis y estimaciones de manera agregada.

Se debe tener en cuenta no obstante que los análisis a efectuar (por ejemplo los algoritmos de minería de datos), deben rediseñarse para poder trabajar con distribuciones en lugar de con registros individuales. En concreto deben permitir eliminar (sustraer) primero la componente de ruido para reestimar X. Para ello se emplean habitualmente métodos como el algoritmo de Esperanza-Maximización (Expectation–Maximization -EM-algorithm), que permite aproximar en primer lugar la distribución de Z a partir de las muestras de datos publicadas, y entonces obtener nuevos parámetros para estimar X.

Partiendo de esta base se han modificado varios algoritmos de minería de datos (clasificadores como el Bayesiano Ingenuo (Naive-Bayes)) y se han definido métodos como los inductores de regla de asociación, las aplicaciones de procesamiento analítico en línea (On-Line Analytical Processing –OLAP-) y los sistemas de filtrado colaborativo basados en Descomposición en Valores Singulares (Singular Value Decomposition –SVD-).

Como principales ventajas de la aleatorización aditiva se encuentra que es relativamente simple de aplicar, dado que no requiere del conocimiento de la distribución de otros registros en el conjunto de datos a anonimizar (lo cual no ocurre por ejemplo para los métodos de anonimización grupales que se presentarán más adelante), y además se puede usar desde el momento de la recopilación de los datos (por lo que no es necesario que los datos se almacenen en un servidor seguro, antes de ejecutar su anonimización).

No obstante debe tenerse en cuenta que, con el objetivo de mantener lo más alta posible la utilidad de los datos al mismo tiempo que se protege la privacidad, no es necesario añadir el mismo nivel de ruido a todos los registros, sino que debe considerarse si se encuentran en zonas de alta densidad de datos (será necesario menos ruido para protegerlos) o si son datos atípicos más susceptibles de un ataque (para los que en consecuencia se necesitará añadir un mayor nivel de ruido).

- **Métodos Multiplicativos**

  Estos métodos se basan en técnicas de reducción de la dimensionalidad, que asocian los datos a un espacio de menor dimensión mediante una multiplicación matricial proyectiva.

  En general, estos métodos permiten el uso de determinados algoritmos de minería de datos (que mantienen la distancia entre registros) directamente sobre los datos aleatorizados.

Con carácter general, los métodos de aleatorización (tanto aditivos como multiplicativos) son vulnerables especialmente a dos tipos de ataques:

- Casos donde un atacante conoce una serie de registros linealmente independientes y su versión modificada (ataques de Entrada-Salida Conocida), con lo que puede aplicar ingeniería inversa (técnicas de algebra lineal).
- Estos métodos tampoco son los más adecuados para proporcionar protección frente a ataques de Muestra Conocida, donde el atacante tiene una serie de muestras independientes de la misma distribución de la que se han obtenido los datos originales.

4.3.2.1.2. Métodos basados en la generación de microdatos sintéticos

Estos métodos consisten en la inclusión en los conjuntos de datos a proteger, de microdatos sintéticos, que si bien no son datos reales (y por lo tanto no comportan problemas de protección de la privacidad), sí preservan las características estadísticas de los datos reales (y permiten el desarrollo de técnicas de minería de datos y la obtención de resultados agregados).

Para garantizar estas características, los microdatos agregados se producen mediante algoritmos de simulación de datos.

4.3.2.2. Métodos de Anonimización Grupales

4.3.2.2.1. k-Anonimización (k-Anonymity)

Este modelo consiste en buscar que la información personal publicada pueda ser trazable a un número k de individuos diferentes. **Lógicamente se buscará que k sea lo más alto posible**, minimizándose de esta manera la posibilidad de que se pueda re-identificar correctamente a una persona. Se basa en los atributos de los cuasi-identificadores y las clases de equivalencia, dado que se busca que cada combinación posible de cuasi-identificadores sea compartida por al menos k registros en los datos que se publiquen.

La k-Anonimización fue definida inicialmente en los artículos *[83]* y *[95]* y se basa fundamentalmente en el empleo de la generalización de los cuasi-identificadores e identificadores directos, así como en la supresión de registros y atributos, que a su vez permitan una mayor generalización.

A continuación se compara una tabla original y su versión anonimizada-3.

Código Postal	Nombre	Edad	Enfermedad
28037	Ana López	40	Bronquitis
28021	Sonia Hernández	37	Bronquitis
28056	Juan García	40	Bronquitis
28934	David Sánchez	29	Neumonía
28920	Lucía Pérez	27	Úlcera estomago
28956	María Martín	28	Escoliosis
08034	Carlos Muñoz	32	Gastritis
08078	Laura González	31	Gastritis
08003	Jesús Rodríguez	32	Neumonía

Tabla II-2a - Datos Originales

Código Postal	Edad	Enfermedad
280**	>36	Bronquitis
280**	>36	Bronquitis
280**	>36	Bronquitis
289**	[26-30]	Neumonía
289**	[26-30]	Úlcera estomago
289**	[26-30]	Escoliosis
080**	[31-35]	Gastritis
080**	[31-35]	Gastritis
080**	[31-35]	Neumonía

Tabla II-2b – Versión Anonimizada-3 de la Tabla II-2a

En la tabla anonimizada se dispone de 3 clases de equivalencia, dentro de las cuáles no es posible diferenciar a ninguna persona del resto de personas en esa clase empleando los identificadores directos y los cuasi-identificadores.

Problemas y ataques que puede sufrir

La k-Anonimización es más adecuada para proteger frente a la revelación de la identidad, que frente a la revelación de los atributos sensibles, dado que no puede proteger frente a ésta última en todos los casos. Es especialmente vulnerable frente a dos tipos de ataque:

- **Ataque de Homogeneidad.**

  Se basa en que a través de la k-Anonimización se pueden crear grupos de los que se puede inferir información debido a la ausencia de diversidad en los atributos sensibles y empleando de forma combinada información que se encuentra disponible en el dominio público.

  En el ejemplo anterior, un usuario externo tiene acceso a la tabla anonimizada. Este usuario sabe por información pública disponible que Sonia Hernández vive en el Código Postal 28021, que tiene 37 años, y además sabe que su información está en la tabla publicada. Por lo tanto, únicamente con esta información puede saber que Sonia tiene Bronquitis, con lo que este dato sensible habría quedado revelado.

- **Ataque por Conocimiento de Background.**

  Otro posible ataque se basa en la información de la que, sin ser pública, puede disponer previamente un determinado actor externo que tiene acceso a los microdatos (ya sea con intenciones ilícitas o no).

  Nuevamente en el ejemplo anterior, un usuario externo sabe que su amiga Laura González, que tiene 31 años, vive en el código postal 08078 y que su registro se encuentra en la tabla publicada de forma anónima.

  Además este usuario sabe que Laura no tiene ningún tipo de problema respiratorio, por lo que puede concluir que sufre Gastritis.

4.3.2.2.2. l-Diversidad (l-Diversity)

El método de anonimización denominado l-Diversidad, tiene como fundamento inicial el intentar superar las limitaciones de la k-Anonimización en relación con la posible ausencia de una representación suficientemente variada de valores para el atributo sensible en cada una se las cases de equivalencia que se establezcan (y de este modo reduciendo la vulnerabilidad frente a los ataques de homogeneidad).

Se dice que una clase de equivalencia cumple con la l-Diversidad, si en ella existen un número l o más valores adecuadamente representados para el atributo sensible. Si cada clase de equivalencia de un conjunto de datos cumple con la l-Diversidad, este conjunto se datos se considera así mismo como l-Diverso.

Este método fue definido inicialmente en el artículo *[61]*.

	Código Postal	Edad	Salario Bruto Anual	Enfermedad
1	28037	25	23.000 €	Hernia Disco
2	28021	26	24.000 €	Escoliosis
3	28056	28	25.000 €	Lumbalgia
4	28934	59	26.000 €	Escoliosis
5	28920	47	31.000 €	Gripe
6	28956	48	28.000 €	Bronquitis
7	28034	32	27.000 €	Bronquitis

	Código Postal	Edad	Salario Bruto Anual	Enfermedad
8	28078	31	29.000 €	Neumonía
9	28003	32	30.000 €	Lumbalgia

Tabla II-3a – Datos Originales

	Código Postal	Edad	Salario Bruto Anual	Enfermedad
1	280**	2*	23.000 €	Hernia Disco
2	280**	2*	24.000 €	Escoliosis
3	280**	2*	25.000 €	Lumbalgia
4	289**	>40	26.000 €	Escoliosis
5	289**	>40	31.000 €	Gripe
6	289**	>40	28.000 €	Bronquitis
7	280**	3*	27.000 €	Bronquitis
8	280**	3*	29.000 €	Neumonía
9	280**	3*	30.000 €	Lumbalgia

Tabla II-3b – Versión 3-Diversa de la Tabla II-3a

Existen 3 tipos de l-Diversidad:

- **l-Diversidad perceptible**: Es la forma más general de l-Diversidad y en ella para cada clase de equivalencia que se establezca existirán al menos l valores distintos para el atributo sensible cuya revelación se trata de proteger. De este modo se consigue una protección frente a los ataques de homogeneidad, pero no frente a ataques de inferencia probabilística.

  Estos últimos se dan cuando por ejemplo en una clase de equivalencia algunos valores aparecen con mucha mayor frecuencia que otros, con lo que un atacante externo puede concluir que existe una gran probabilidad de que el atributo sensible para una determinada persona sea precisamente el más probable. Esta carencia ha dado lugar al desarrollo de otros tipos de l-Diversidad.

- **l-Diversidad Entrópica**: La entropía de una clase de equivalencia E se define como:

$$Entropía\ (E) = -\sum_{s \in S} p(E,s)\ log\ p(E,s)$$

  donde S es el dominio del atributo sensible y $p(E,\ s)$ es el porcentaje de registros en E que tienen el valor sensible S.

La tabla tiene l-Diversidad entrópica si en cada clase de equivalencia E, la $entropía\ (E) \geq log\ l$.

El principal problema con este tipo de mecanismo de anonimización es que restringe mucho los datos, dado que la entropía de la tabla completa puede ser muy baja si unos pocos valores son demasiado frecuentes.

- **Diversidad Recursiva (c,l)**: En este caso se restringen los valores más y menos frecuentes para limitar la posibilidad de ataques de inferencia probabilística.

Problemas y ataques que puede sufrir

Tampoco permite una protección absoluta frente a la revelación de atributos.

Además, este método tampoco ofrece una protección de la privacidad absoluta en todos los casos, siendo especialmente vulnerable frente a dos tipos de ataque:

- **Ataque de Asimetría Estadística (Skewness)**

  Aunque la protección contra la revelación de la identidad es proporcionada por la l-Diversidad, este método no previene completamente frente a la revelación de atributos cuando la distribución presenta Asimetría Estadística.

  Por ejemplo, teniendo en cuenta una tabla de datos en la que se presenta si una serie de personas han aprobado o suspendido un examen, se podrían establecer 3 clases de equivalencia:

  - En el primer caso, se trata de una clase de equivalencia con 50 registros, con igual número de aprobados que de suspensos.
  - En el segundo caso, existen 49 registros suspensos y un solo aprobado.
  - En el tercer caso, habría 49 registros aprobados y uno solo suspenso.

  En los tres casos se satisface la 2-Diversidad, pero se comprueba como existen niveles de sensibilidad diferentes (el pertenecer a la segunda clase de equivalencia supone una posibilidad muy alta (98%) de haber suspendido el examen, mientras que el estar incluido en la tercera, conlleva que la probabilidad de haber suspendido es muy baja (2%)).

- **Ataque de Similitud (Similarity)**

  Estos ataques ocurren cuando los valores sensibles son semánticamente similares.

  Así por ejemplo en la *tabla II-3b*, en caso de conocer que el registro de una persona concreta está incluido en la primera clase de equivalencia (por conocer su código postal y su edad), entonces es posible saber que sufre de un problema relacionado con la espalda (dado que todos los atributos sensible Enfermedad son problemas de este tipo) y además que su salario se encuentra en un rango reducido entre 23.000€ y 25.000€.

  Esta relación entre atributos sensibles en una clase de equivalencia es el motivo principal por el que se buscan nuevos métodos de anonimización grupales como la t-Proximidad.

4.3.2.2.3. t-Proximidad (t-Closeness)

Este método fue definido inicialmente en el artículo *[59]*, y tal y como se ha comentado surge para resolver algunas de las limitaciones que presentan los otros métodos de anonimización grupales.

Una clase de equivalencia cumple la t-Proximidad si la distancia entre la distribución de un atributo sensible en esa clase y la distribución del atributo en la tabla completa de datos, es

menor que un valor umbral t. Una tabla cumple la t-Proximidad si todas sus clases de equivalencia cumplen la t-Proximidad.

En este punto, es fundamental definir cómo se va a medir la distancia que se va a emplear para determinar la diferencia entre la distribución del atributo sensible en una clase de equivalencia y en la tabla completa. Existen métricas como la de *Kullback-Leibler* o la de la *Distancia de Variación*, pero las mismas no tienen en cuenta la distancia semántica, que como se ha visto en el apartado anterior, es importante a la hora de proteger frente a los Ataques de Similitud.

Por este motivo se decide emplear la Earth Mover's Distance (EMD), que se basa en calcular la mínima cantidad de trabajo necesaria para transformar una distribución en otra, moviendo entre ellas la masa de la distribución.

La EMD proporciona un método para determinar la distancia entre dos distribuciones, pero no permite establecer la distancia entre dos elementos de las distribuciones en cuestión. Para hacer esto es necesario tener en cuenta el tipo de datos de los que se trata:

- **Atributos numéricos** (en los ejemplos anteriores las edades o los salarios).

  En este caso el concepto fundamental para medir la distancia, es la ordenación de los valores. Considerando el dominio de los valores $\{v_1, v_2 \ldots v_m\}$, donde v_i es el i-ésimo valor más pequeño, la distancia entre dos valores se calcula teniendo en cuenta el número de valores entre ellos en el total, por ejemplo mediante la siguiente fórmula:

  $$distancia\text{-}ordenada(vi, vj) = \frac{|i-j|}{m-1}$$

  Para medir distancias entre distribuciones, usando la EMD, se considera $r_i = p_i - q_i, (i = 1,2,\ldots,m)$, y entonces la distancia entre las distribuciones P y Q se puede calcular como:

  $$D[P,Q] = \frac{1}{m-1}(|r_1| + |r_1+r_2| + \ldots + |r_1+r_2+\ldots+r_{m-1}|) =$$

  $$= \frac{1}{1-m} \sum_{i=1}^{i=m} \left| \sum_{j=1}^{j=i} rj \right|$$

- **Atributos en categorías** (en los ejemplos anteriores, las enfermedades). En este caso habría a su vez dos opciones para el cálculo de la distancia:
  - **Distancia Igual**, que se empleará cuando no tenga sentido, por la naturaleza del propio atributo hablar de distancia entre los valores implicados (por ejemplo es estado civil: soltero, casado, divorciado o viudo). En este caso se considerará la distancia como siempre 1 (distancia igual). De este modo:

    $$D[P,Q] = \frac{1}{2} \sum_{i=1}^{m} |pi - qi| = \sum_{pi \geq qi} (pi - qi) = - \sum_{pi < qi} (pi - qi)$$

  - **Distancia Jerárquica**, que se empleará cuando los valores puedan ser ordenados según algún tipo de árbol de jerarquía que de algún modo las ordene. Así por ejemplo si se consideran valores de un campo Enfermedad de una Base de Datos, un valor Escoliosis estará más cercano a un valor Hernia Discal (al ser ambas dolencias relacionadas con la espalda), que a un valor Gastritis (al ser una enfermedad del aparato digestivo).

    En este caso se considera H como la altura del dominio jerárquico (el número de niveles con que contaría el árbol de jerarquía) y la distancia entre dos valores vendrá

definida por $\frac{nivel\ (v1,v2)}{H}$, donde el $nivel\ (v_1,\ v_2)$ es la altura del ancestro común más bajo de v_1 y v_2 en el árbol.

En este caso:

$$D[P, Q] = \sum_N Cost\ (N)$$

Se expone el empleo de estos conceptos, nuevamente empleando el mismo ejemplo que se uso para mostrar la aplicación de la l-Diversidad:

	Código Postal	Edad	Salario Bruto Anual	Enfermedad
1	28037	25	23.000 €	Hernia Disco
2	28021	26	24.000 €	Escoliosis
3	28056	28	25.000 €	Lumbalgia
4	28934	59	26.000 €	Escoliosis
5	28920	47	31.000 €	Gripe
6	28956	48	28.000 €	Bronquitis
7	28034	32	27.000 €	Bronquitis
8	28078	31	29.000 €	Neumonía
9	28003	32	30.000 €	Lumbalgia

Tabla II-4a – Datos Originales

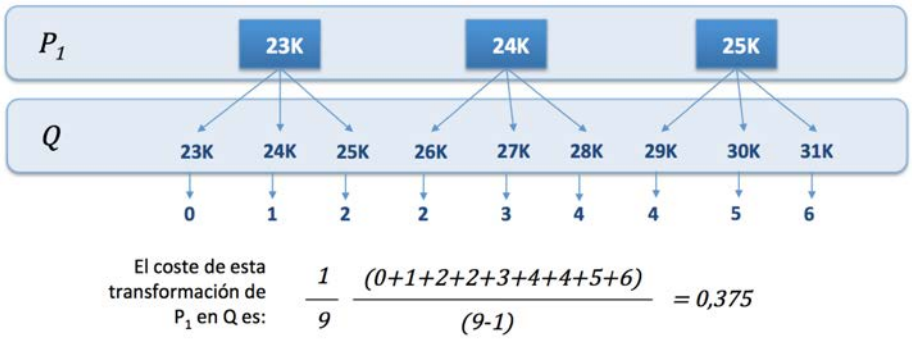
	Código Postal	Edad	Salario Bruto Anual	Enfermedad
1	280**	2*	23.000 €	Hernia Disco
2	280**	2*	24.000 €	Escoliosis
3	280**	2*	25.000 €	Lumbalgia
4	289**	>40	26.000 €	Escoliosis
5	289**	>40	31.000 €	Gripe
6	289**	>40	28.000 €	Bronquitis
7	280**	3*	27.000 €	Bronquitis
8	280**	3*	29.000 €	Neumonía
9	280**	3*	30.000 €	Lumbalgia

Tabla II-4b – Versión 3-Diversa de la Tabla 4a

En primer lugar se aplicará la t-Proximidad sobre el atributo sensible de tipo Numérico *Salario Bruto Anual*, cuyos posibles valores son: Q = {23K, 24K, 25K, 26K, 27K, 28K, 29K, 30K, 31K}

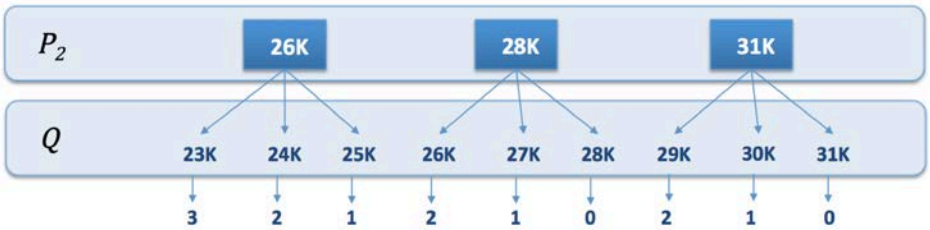
En la primera clase de equivalencia, los valores son: P_1 = {23K, 24K, 25K} y en la segunda clase de equivalencia son: P_2 = {26K, 28K, 31K}

La distancia D[P_1, Q] se calcularía como:



El coste de esta transformación de P_1 en Q es: $\frac{1}{9}\frac{(0+1+2+2+3+4+4+5+6)}{(9-1)} = 0{,}375$

y la Distancia D[P_2, Q]:



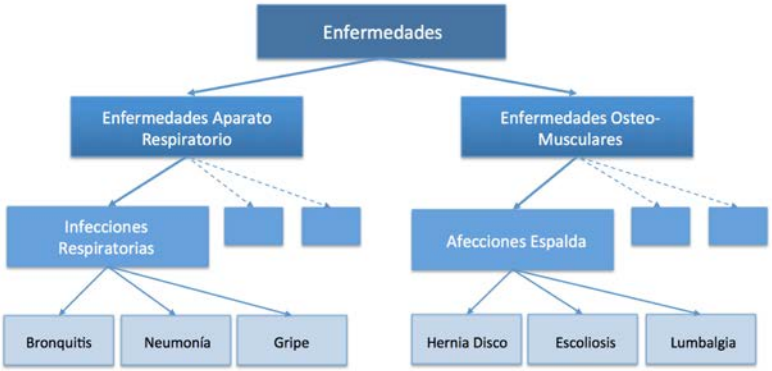
El coste de esta transformación de P_2 en Q es: $\frac{1}{9}\frac{(3+2+1+2+1+0+2+1+0)}{(9-1)} = 0{,}167$

Como D[P_1, Q] > D[P_2, Q], se puede concluir que P_2 revela menos datos privados. Este resultado podía deducirse dado que existirá menos gente en el rango P_1 de salario que en el P_2 al ser este segundo considerablemente más amplio, y será 0,167 el umbral que se emplee.

En segundo lugar se aplicará la t-Proximidad sobre el atributo jerárquico Enfermedad, cuyos posibles valores son: Q = {Hernia Disco, Escoliosis, Lumbalgia, Escoliosis, Gripe, Bronquitis, Bronquitis, Neumonía, Lumbalgia}

En la primera clase de equivalencia, los valores son: P_1 = {Hernia Disco, Escoliosis, Lumbalgia} y en la segunda clase de equivalencia son: P_2 = {Escoliosis, Gripe, Bronquitis}.

En este caso y tal y como se comentó con anterioridad es imprescindible contar con el árbol jerárquico del atributo enfermedad:



Tal y como se apuntó con anterioridad la distancia entre 2 dos valores vendrá definida por $\frac{nivel\ (v1,v2)}{H}$, donde el $nivel\ (v_1,\ v_2)$ es la altura del ancestro común más bajo de v_1 y v_2 en el árbol, por lo que por ejemplo la distancia entre *Hernia de Disco* y *Escoliosis* es de $\frac{1}{3}$ y la de Hernia de Disco y Neumonía es de $\frac{3}{3}$.

Calculando la distancia D[P_1, Q] obtenemos que es igual a 0,5 y la D[P_2, Q] es igual a 0,278 (aplicando las formulas de la EMD para el caso de atributos jerárquicos). Será por lo tanto 0,278 el umbral que se emplee en este caso.

Una vez calculados los umbrales, se puede componer la tabla anonimizada. Se trata de una tabla que cumple la **0,167-Proximidad** con respecto al atributo *Salario Bruto Anual* y la **0,278-Proximidad** con respecto al atributo *Enfermedad*.

En ella se mantiene la clase de equivalencia 2, ya que tanto para el atributo *Salario Bruto Anual*, como para *Enfermedad*, es la que ha establecido los umbrales. Ha sido necesario modificar la generalización del atributo *Edad* para mantener las propiedades de las clases de equivalencia.

	Código Postal	Edad	Salario Bruto Anual	Enfermedad
1	280**	<40	23.000 €	Hernia Disco
8	280**	<40	29.000 €	Neumonía
3	280**	<40	25.000 €	Lumbalgia
4	289**	>40	26.000 €	Escoliosis
5	289**	>40	31.000 €	Gripe
6	289**	>40	28.000 €	Bronquitis
7	280**	<40	27.000 €	Bronquitis
2	280**	<40	24.000 €	Escoliosis
9	280**	<40	30.000 €	Lumbalgia

Tabla II-5 – Versión 0,167-Proxima con respecto a 'Salario Bruto Anual' y 0,278-Próxima con respecto a 'Enfermedad' de la Tabla II-4a

4.3.2.3. Métodos de Control de los Resultados de las Consultas

4.3.2.3.1. Auditorías

El método de establecimiento de Auditorías de Control de las Solicitudes recibidas, permite determinar qué solicitudes deben aceptarse y por lo tanto para ellas debe proporcionarse al solicitante el resultado esperado, y cuáles deben rechazarse al constituir un riesgo para la privacidad de los datos implicados.

En concreto se busca cuantificar la cantidad de solicitudes que pueden ser respondidas sin que ello suponga la revelación de información privada. Para ello son muy relevantes los conceptos de Coste de Privacidad y de Presupuesto de Privacidad (Privacy Cost y Privacy Budget) que permitirán que el responsable de la Privacidad de un determinado conjunto de datos controle el riesgo que supone cada petición, no sólo de manera individualizada, sino teniendo en cuenta el riesgo que puede haberse asumido ya con otras solicitudes previas.

4.3.2.3.2. Privacidad Diferencial

Básicamente este método consiste en insertar ruido aleatorio en las respuestas a las solicitudes (ruido o inexactitudes), para garantizar matemáticamente que la información personal de un individuo queda '*enmascarada'*. En función del riesgo asociado a cada solicitud se decide el nivel de ruido a introducir.

La Privacidad Diferencial fue definida inicialmente en los artículos *[35]* y *[36]*.

Una función aleatoria f proporciona ε-Privacidad Diferencial si, para todos los registros de Datos D_1, D_2, tales que uno puede ser obtenido a partir del otro a través de la modificación de un único registro, y todos los $S \subseteq$ Rango(k), se cumple que:

$$Pr(k(D_1) \in S) \leq exp(\varepsilon) \times Pr(k(D_2) \in S)$$

Dada una solicitud f, el objetivo es encontrar una función aleatoria k_f que cumpla la ε-Privacidad Diferencial y que se aproxime a f lo máximo posible. Para conseguir la privacidad diferencial es necesario introducir incertidumbre a la salida de la solicitud. Tal y como ya se ha destacado, esta incertidumbre se puede conseguir añadiendo ruido en las respuestas:

$k_f(D) = f(D) + N_f(D)$, donde $Nf(D)$ es un ruido aleatorio, cuya distribución depende de los valores de f y D. Una elección que suele emplearse es que el ruido aleatorio tenga una distribución de Laplace con unos parámetros que dependerían de la variabilidad de f entre conjuntos de datos que difieren únicamente en un registro *[91]*.

En este caso los investigadores no tienen acceso a listados de información (ni en su forma original ni en la anonimizada), sino que deben obtener toda la información que necesitan para el desarrollo de sus investigaciones / servicios de unas preguntas que se responderán con información de los listados (como ya se ha dicho tras introducirles una serie de 'inexactitudes').

Para evaluar el nivel de inexactitud a introducir no sólo se tiene en cuenta el riesgo que supone independientemente una solicitud sino también todas aquellas que previamente se hayan hecho (por ese usuario en concreto o por todos en general).

Una de las ventajas de este método es que los datos no tienen que ser modificados en ningún modo.

4.3.2.4. Securización Distribuida

Los métodos planteados hasta este punto, consideran que toda la información sensible está en manos de una única entidad de confianza. Sin embargo tal y como se ha expuesto en

apartados anteriores, la gestión de los Big Data se basa en un modelo de gestión distribuida y en paralelo, en el que intervienen múltiples entidades.

Estas múltiples partes (sistemas esclavos, según la Arquitectura de Hadoop, que ya se ha expuesto, que disponen de un componente Data Node en el que analizan una parte de los datos sensibles totales disponibles), tienen acceso únicamente a los datos que necesitan para el desarrollo de las tareas que les son encomendadas por el Sistema Maestro, no pudiendo por el contrario acceder a la totalidad de los datos disponibles.

- **Particionamiento Horizontal**

  En este esquema de preservación de la privacidad, cada entidad (sistema esclavo) del sistema distribuido tiene acceso a un conjunto de datos con la misma estructura de campos, si bien los registros incluidos en este conjunto son disjuntos con aquellos de los que dispone otro sistema esclavo.

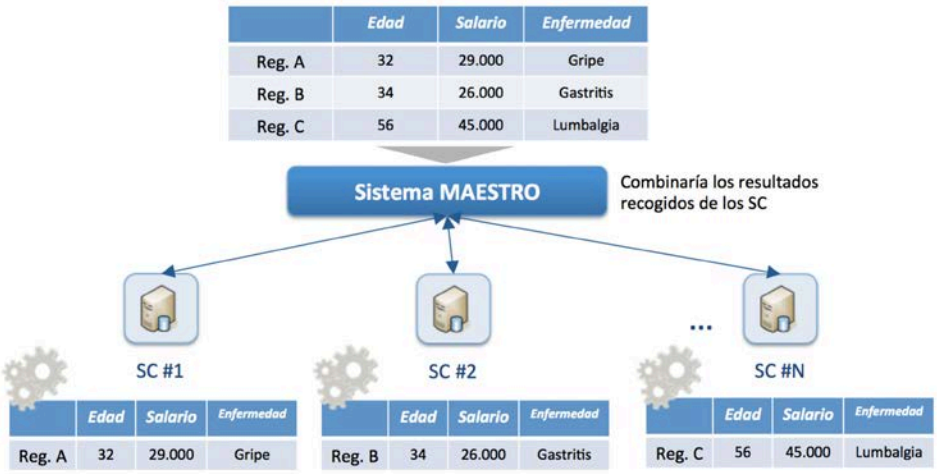


Imagen II-13 – Esquema Aplicación Particionamiento Horizontal

- **Particionamiento Vertical**

  En este caso, cada sistema esclavo en lugar de disponer de registros de datos con la misma estructura, tiene acceso a conjuntos de atributos (campos) disjuntos.

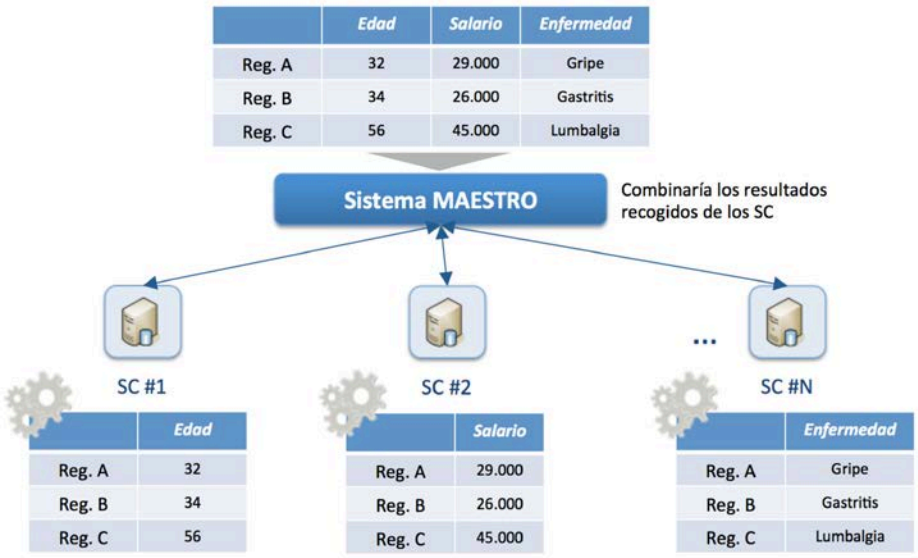


Imagen II-14 – Esquema Aplicación Particionamiento Vertical

4.3.2.5. Preservación de la Privacidad en Datos No Estructurados – El caso de información en formato Texto

Evidentemente no todos los datos que se manejan y mucho menos aún en un contexto con las especificidades ya comentadas para los Big Data, se encuentran estructurados en forma de base de datos.

En muchos casos los analistas deben poder obtener información a partir de información no estructurada, y para ella es imprescindible asegurar también la protección de la privacidad.

Entre esta información no estructurada se encuentra la información textual. Entre ésta, una de los casos más empleados son las búsquedas en Internet que hacen los usuarios.

Como uno de los enfoques mas extendidos que se están siguiendo para la anonimización de este tipo de datos se encuentra la aplicación de algunas de las técnicas ya presentadas en los apartados anteriores para los datos estructurados, como por ejemplo la k-Anonimización *[96]* o la Privacidad Diferencial *[36]*. Para ello se tratan los textos como registros de una base de datos de longitud variable.

En este caso, la anonimización que puede realizarse supone una pérdida de utilidad mayor que la observada para el caso de bases de datos estructuradas.

Además en *[54]* se ha mostrado que mediante el adecuado entrenamiento de clasificadores y regresores relativamente sencillos, es posible reconstruir los cuasi-identificadores presentes en un texto, como la edad, el código postal o el género, que entonces pueden ser empleados para reidentificar por ejemplo al autor de una determinada búsqueda en Internet, a pesar de que el bloque de datos pueda haber sido previamente anonimizado.

Se trata por tanto este tipo de anonimización de uno de los retos fundamentales abiertos actualmente en el ámbito de la protección de la privacidad.

4.3.2.6. Preservación de la Privacidad en Datos No Estructurados – El caso de las Redes Sociales.

En este caso, y dado el alcance de este proyecto, simplemente se van a apuntar algunas nociones generales en relación con los métodos de protección de la privacidad en Redes Sociales.

Las redes sociales se caracterizan por una **estructura de nodos**, que representan a los diferentes usuarios o entidades que se interrelacionan, y **de conexiones**, que muestran precisamente las características de dichas interrelaciones que se producen.

Normalmente los **atributos asociados a los nodos** (como los intereses de una persona, su edad, estado civil, etc.), son los aspectos que van a demandar una mayor protección desde el punto de vista de la privacidad. No obstante, aquellos otros **relativos a las conexiones**, que a priori podrían considerarse menos sensibles, en determinadas ocasiones también requerirán una protección especial (como por ejemplo en una red social de contactos a partir de la cual se pudiera inferir la orientación sexual de una persona). Incluso aspectos como el número de contactos pueden en determinadas circunstancias ser reveladores de algunos parámetros sensibles (nuevamente el caso de una red de contactos).

El elevado número de conexiones que se producen normalmente, hace enormemente complejo abordar tareas como la anonimización de las características de un usuario concreto, sin modificar al mismo tiempo información relativa a otros.

Se trata de uno de los campos de estudio más relevantes en la actualidad, dada la gran cantidad de datos que en ellas se generan y las interacciones existentes y el volumen de su

uso y los retos específicos que en cuanto a la protección de la privacidad plantean [48], [70], [111].

5. Las Métricas de Control de la Privacidad e Impacto sobre el Rendimiento

5.1. Introducción

Existen distintas métricas y parámetros a analizar, que permiten comparar las características de diferentes soluciones de protección de la privacidad y poder determinar con mayor facilidad cual se adecua mejora a cada casuística concreta.

Evidentemente en este caso es muy importante identificar los distintos roles involucrados en la protección de la privacidad, dado que la percepción de la importancia relativa de las métricas para cada uno de ellos evidentemente será diferente. Así se diferenciará entre:

- Titular / Propietario de los datos (dentro de ella habría categorías por ejemplo personajes públicos o famosos pueden estar más preocupados en que se asegure la protección de su PII).
- Usuario de los Datos.
- Responsable de la protección de los Datos.

El titular busca maximizar la protección de la privacidad (o dicho de otra forma minimizar el riesgo), el usuario de los datos, busca maximizar la utilidad que obtiene de los análisis que efectúa, y el responsable de la protección de los datos busca en primer lugar minimizar el riesgo, pero al mismo tiempo debe prestar atención (que no tiene por ejemplo el propietario de los datos) a minimizar el coste de las medidas implementadas e incrementar el rendimiento global de la plataforma.

Dado el enfoque de este proyecto, se podrá el foco de análisis en el punto de vista del responsable de la protección de datos, dado que tiene el enfoque más global y que al ser el nexo entre titular y usuario debe buscar un equilibrio entre los intereses de ambos buscando soluciones que favorezcan a ambos.

5.2. Parámetros a analizar

- **Control del Riesgo / Cumplimiento normativo**: Aquí es fundamental la definición del umbral de riesgo que puede ser asumido. Se deben identificar riesgos de privacidad dentro de escenarios de ataques / pérdidas de información realistas (no sobredimensionar añadiendo muchos costes, pérdidas de eficiencia y utilidad innecesariamente).
- **Utilidad de los datos**: Por ejemplo los métodos de control de acceso no afectan a este parámetro de un modo continuo, ya que normalmente se permite el acceso o no.
- **Fallos en la protección**. Porcentaje de información sensible no protegida por el algoritmo evaluado.
- **Impacto de las medidas de protección de la privacidad** en la plataforma de Big Data.
- **Coste de implementación** de la medida de protección de la privacidad.
- **Eficiencia**. La posibilidad de ejecutar un algoritmo o técnica de preservación de la privacidad en unas condiciones de adecuado aprovechamiento de los recursos implicados en el algoritmo.

- **Escalabilidad**. Posibilidad de continuar asegurando la privacidad en conjuntos mayores de datos.
- **Tiempo / Dificultad de implementación**.

A través de soluciones gráficas como la siguiente se facilitaría esta tarea de comparativa:

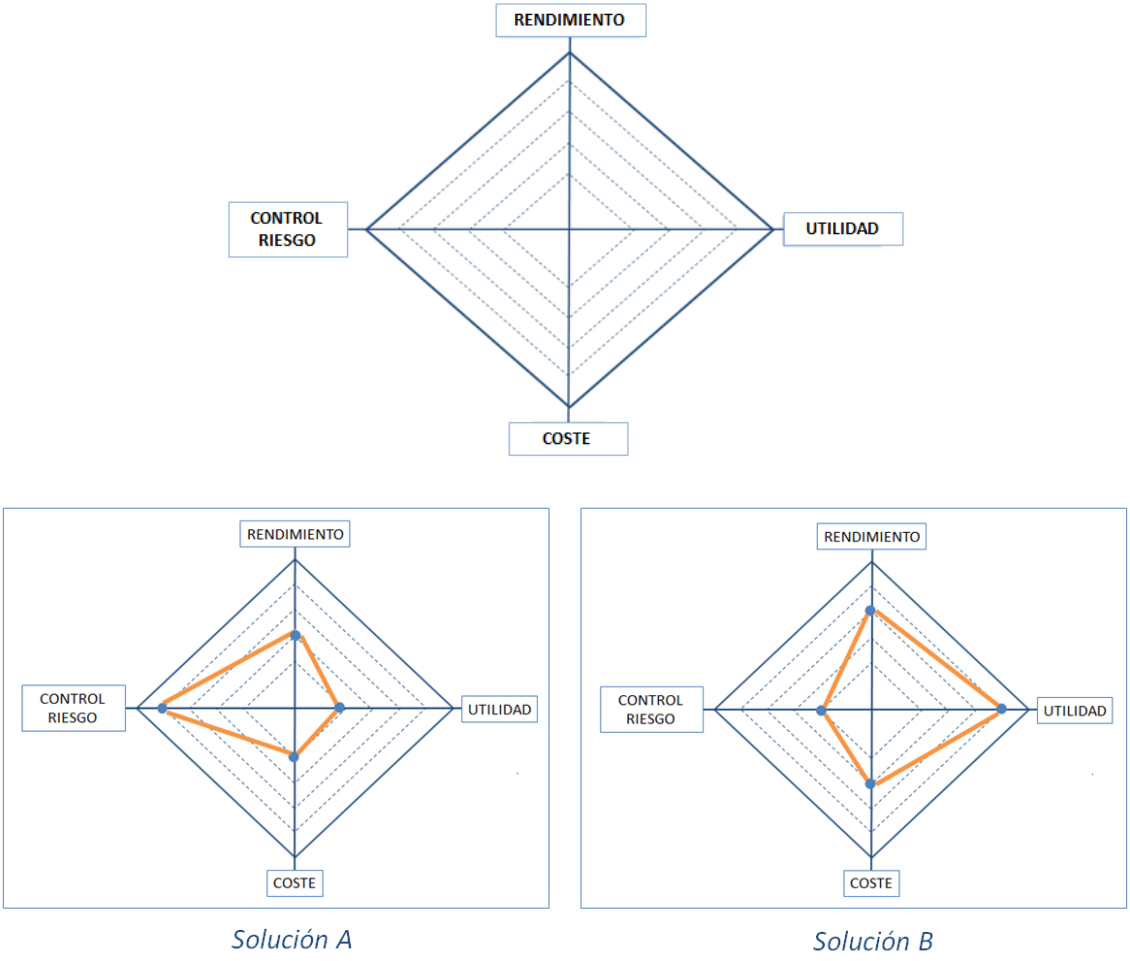


Imagen II-15 – Gráficas de comparativa de parámetros soluciones protección privacidad

Por ejemplo en el caso anterior se comprobaría que la Solución A supondría asumir un riesgo mucho menor que mediante la Solución B, pero a cambio supone una reducción del Rendimiento y de la Utilidad y Beneficios que podrían obtenerse de los análisis de los datos implicados.

En función del contexto en el que se emplearán los datos, se optará por un método de anonimización u otro. Así por ejemplo se puede optar por *[19]*:

- **Maximizar la utilidad, teniendo en cuenta las limitaciones existentes de riesgo y coste.**

  Este caso puede ser útil cuando sea primordial limitar determinados riesgos.

- **Minimizar riesgos de privacidad, teniendo en cuenta las limitaciones de utilidad y coste.**

En algunos casos (como por ejemplo en el ámbito de la Sanidad), no es asumible una reducción en el control del riesgo sobre los datos, dada la criticidad de los análisis que se efectúan sobre los datos y las consecuencias que una identificación errónea o un retraso pueden conllevar.

- **Minimizar el coste, teniendo en cuenta las limitaciones existentes de utilidad y riesgo.**

  En determinados contextos, el coste puede ser el parámetro primordial que hay que buscar reducir (por ejemplo el caso de uso de protocolos criptográficos).

CAPÍTULO III

ESTADO DEL ARTE DE LA PROTECCIÓN DE LA PRIVACIDAD EN LA GESTIÓN DE BIG DATA – DIMENSIÓN NORMATIVA

1. Introducción y Objeto del Capítulo.

En este capítulo, complementando el anterior, se presenta un análisis de la situación actual en relación con la **Normativa y las Regulaciones de Protección de la Privacidad**, centrándose en cómo esta Normativa se ve afectada por la aparición de las soluciones de gestión de Big Data, y qué enfoques se plantean para afrontar los diversos retos existentes.

- En primer lugar se presentan los **Principios Fundamentales que en relación con la Protección de los Datos de Carácter Personal**, han constituido la **referencia principal alrededor de la cual se han establecido las diversas regulaciones** en esta materia desde los años 70 del siglo XX.

  En concreto, se analiza **cómo los Big Data impactan sobre estos Principios** y en qué medida pueden incluso hacer que algunos de ellos deban ser reconsiderados para garantizar que su utilidad se mantenga.

- A título ilustrativo de la **situación actual de la normativa** de Protección de la Privacidad en el mundo, se presentan un análisis para diversos países y Organizaciones, especialmente centrado en el caso de la Unión Europea (particularizando algunos aspectos para España y Reino Unido), aunque también se destacan particularidades para el caso de los Estados Unidos.

- Por último, se plantean cuáles son los **principales retos a los que debe responder la normativa** que en esta materia se elabore para hacer que se convierta realmente en una herramienta que incremente la eficiencia en la protección de la privacidad, conservando al mismo tiempo la utilidad de las técnicas y herramientas que proporcionan los Big Data, así como algunas de las **soluciones** que en este sentido se están planteando.

1.1. Esquema de estructura del Capítulo

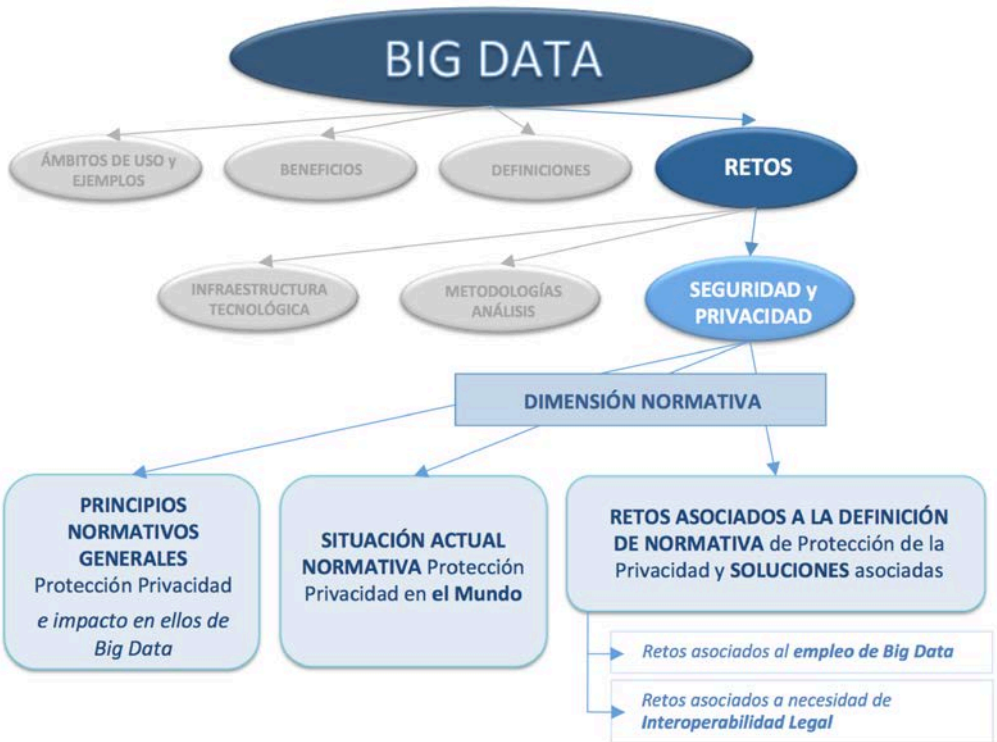


Imagen III-1 – Esquema Estructura Capítulo III

2. La Relevancia de la Normativa en la Protección de la Privacidad

En los últimos años se está asistiendo a un aumento sin precedentes en la elaboración de normativa en relación con la protección de datos de carácter personal, derivado de una mayor conciencia de los usuarios de los peligros que las nuevas tecnologías suponen para su privacidad. Esta normativa tiene un gran impacto sobre la manera en la que las distintas organizaciones y empresas recopilan y gestionan la información relativa a las personas con las que interactúan, bien como clientes o en alguna otra forma de relación.

Las diversas normativas en vigor reflejan la cultura de cada país o área en la que son de aplicación, lo que hace que exista una cierta heterogeneidad y que sea necesaria la búsqueda de una mayor interoperabilidad para poder responder con agilidad y eficiencia a los retos que supone la evolución en las tecnologías y en los métodos de gestión.

En este contexto los Big Data suponen uno de los mayores retos a los que debe responder la normativa, dado que es fundamental que las regulaciones sirvan de apoyo sobre el que consolidar un más rápido y eficiente desarrollo de las soluciones tecnológicas de gestión y protección de la privacidad (como las planteadas en el *Capítulo II*).

En concreto, estas soluciones tecnológicas, se deben modular por la Dimensión Normativa, que por ejemplo establece el Grado de Protección que debe tener cada tipo de dato en diferentes circunstancias.

Así, aspectos como qué información debe tener un mayor grado de protección, o qué tipo de datos se consideran información personal o no (concepto alrededor del cual, y tal y como ya se planteó en el *Capítulo I*, no existe unanimidad a la vista de las posibilidades de reidentificación que ofrecen los Big Data), vendrán determinados normalmente por la normativa o jurisprudencia de aplicación.

En relación con la consideración de un dato como información personal, y como un caso ilustrativo, cabe mencionar por ejemplo, que el Tribunal Supremo de España en su sentencia *S 3-10-2014, rec. 6153/2011*, se ha pronunciado afirmativamente acerca de que las direcciones IP constituyen Datos Personales, dado que estima que contienen información concerniente a personas físicas "identificadas o identificables" *[97]*.

La normativa también establece qué ámbitos son aquellos en los que se requiere una protección especial, debido a que las prácticas de clasificación y segmentación de ciudadanos y usuarios de Servicios y Sistemas de Información y Telecomunicaciones, pueden ser discriminatorias con ellos (o al menos no proteger sus intereses en la medida necesaria). En este caso cabría citar los sistemas de determinación de exención en el pago de impuestos, la analítica predictiva en seguros de salud o la publicidad online basada en comportamiento de usuarios.

A la vista de las múltiples posibilidades que ofrece la gestión de los Big Data y su impacto sobre la privacidad de los usuarios de servicios de Tecnologías de la Información y Comunicaciones (TIC), éste es un ámbito que ha concitado la atención de los responsables de la normativa que regula los servicios digitales y el empleo de las TIC.

En muchas ocasiones la evolución de la normativa y las posibilidades que ofrece la tecnología no avanzan en paralelo:

- Por un lado el desarrollo normativo suele adolecer de un carácter meramente reactivo, regulando aspectos que en el empleo de la tecnología son ya son habituales.

- En un ámbito como las TIC, el alineamiento entre regulaciones es primordial, por lo que se necesitan esfuerzos conjuntos para incrementar la interoperabilidad entre regulaciones de diferentes países / organizaciones supranacionales.
- En ocasiones la regulación puede llegar a ser tan restrictiva, que puede impedir un aprovechamiento de las ventajas que ofrece el análisis de Big Data. Nuevamente se trata de que la normativa permita encontrar el punto de equilibrio entre utilidad y protección.

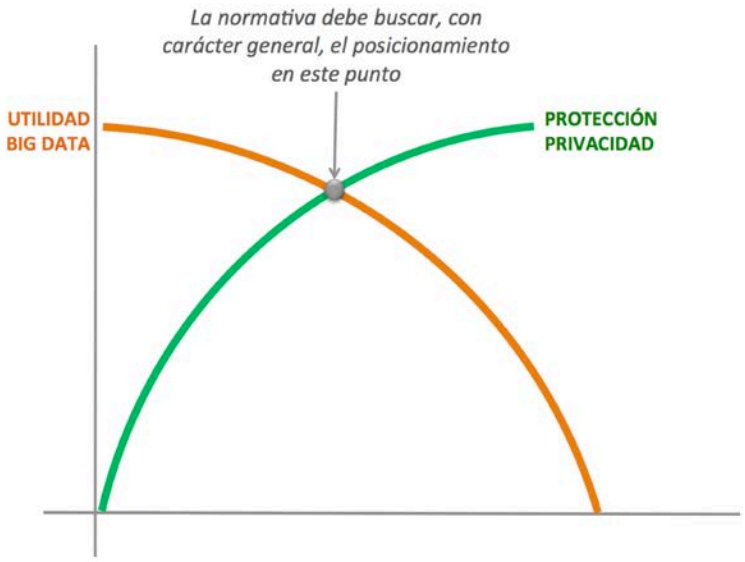


Imagen III-2 – Punto de Equilibrio Utilidad / Protección Privacidad

Lógicamente el anterior es un caso general, y en algunas ocasiones, dada la criticidad de la información que se maneja (por ejemplo datos médicos o datos relativos a menores de edad) será más crítica la Protección de la Privacidad, y la Normativa deberá posicionarse más a la derecha en la gráfica. Por el contrario en otras ocasiones no estará justificada una protección tan estricta (al emplearse datos menos sensibles, o cuando pueda demostrarse que el beneficio que se obtiene de un análisis exhaustivo está por encima de la protección de la privacidad (por ejemplo en casos de lucha frente al terrorismo), y en este caso la normativa buscará el posicionamiento en zonas más a la izquierda de la gráfica.

En resumen, debe evitarse que la normativa se convierta en:

- Un instrumento que ralentice e impida aprovechar todas los beneficios que la tecnología ofrece (extremo derecho de la gráfica).
- Un mero trámite que en realidad no permita ningún tipo de protección (extremo izquierdo de la gráfica).

3. Principios Normativos Generales de Protección de la Privacidad

3.1. Los Principios Fundamentales de la Protección de la Privacidad

Existen, como ya se ha comentado en el *Capítulo I*, unos **Principios Fundamentales de Protección de la Privacidad (Fair Information Privacy Practices –FIPP-)**, que constituyen la

base de prácticamente toda la Legislación moderna en relación con la protección de la Privacidad.

Estos principios fueron definidos en diversos documentos a lo largo de los años 70 del siglo XX[*], que sirvieron de referencia para el documento *Guías para la Protección de la Privacidad y los Flujos Transfronterizos de Datos Personales*, elaborado por la Organización para la Cooperación y el Desarrollo Económicos (OCDE) en 1980 (Guidelines on the Protection of Privacy and Transborder Flows of Personal Data by the Committee of Ministers of the Organization for Economic Cooperation and Development in 1980) *[73]*.

Este documento ha constituido la formulación más extendida de las FIPP, y como tal se ha empleado como base para buena parte de las disposiciones normativas que desde entonces se han promulgado.

En concreto los principios que se identifican son: *[13]* y *[73]*.

1. **Limitación en la recopilación de datos / Minimización de datos.**

   Se deben establecer límites a la recopilación de datos personales y dichos datos sólo podrán obtenerse por medios legales y justos y, cuando sea necesario, con el conocimiento y/o consentimiento del titular de los datos.

2. **Garantía de la Calidad e Integridad de los datos.**

   Los datos personales deben ser relevantes para el propósito para el que han sido recopilados y para ello, deben ser exactos, completos y estar actualizados.

3. **Determinación explícita del Propósito para el que se recopilan los datos.**

   Los propósitos para los cuales se recopilan una serie de datos deben especificarse como muy tarde en el momento en el que se recopilan, y su consiguiente uso debe limitarse a dichos propósitos, o en todo caso a otros que no sean incompatibles con los propósitos iniciales, y siempre que para ellos se especifique claramente en cada ocasión que se ha dado un cambio de propósito.

4. **Limitación de uso.**

   Los datos personales no deben ser revelados, puestos a disposición de terceros o empleados para propósitos distintos a aquéllos para los que fueron recopilados (de acuerdo con el Principio anterior), salvo que exista el consentimiento del titular, o que haya un requerimiento legal.

5. **Seguridad.**

   Los datos personales deben protegerse mediante medidas de seguridad razonables contra riesgos como su pérdida, el acceso no autorizado a ellos, su destrucción, uso indebido, modificación o revelación.

6. **Transparencia.**

   Debe existir una política de transparencia en relación con los desarrollos, prácticas y políticas relativas a datos de carácter personal. Debe ser posible establecer inmediatamente la existencia y naturaleza de los datos personales, el propósito principal que se busca mediante su uso, así como la identidad y otros datos identificativos de la organización que los custodiará.

[*] *En concreto los más importantes han sido el **Código de Buenas Prácticas del Departamento de Salud, Educación y Bienestar** de EE.UU. elaborado en 1973 y los **Principios de la Comisión de Estudios de Protección de la Privacidad** (establecida en EE.UU. en 1977 al amparo de la Privacy Act de 1974).*

7. **Participación Individual.**

   Un individuo tendrá derecho a:

   - Obtener de una organización confirmación de que dicha organización dispone o no de datos sobre él/ella.
   - Recibir los datos en relación con él / ella que pudieran obrar en posesión de alguna organización, en un plazo razonable, con un coste (si lo hubiera) que no sea excesivo, de un modo razonable y en un formato que sea inmediatamente inteligible para el individuo
   - Recibir explicaciones en caso de que alguna de las peticiones hechas bajo los epígrafes anteriores sea denegada, y a poder impugnar dicha denegación.
   - Impugnar la posesión de datos de los que es titular, y en caso de que la impugnación prospere, a que los datos sean borrados, rectificados o corregidos.

   De este modo se confiere poder a los usuarios, al ser ellos, en lugar de por ejemplo una agencia gubernamental, quienes defiendan el derecho a la protección de su privacidad.

8. **Responsabilidad y Auditoría.**

   Una organización que custodia datos personales debe ser responsable de todas las medidas necesarias para el cumplimiento de los principios anteriores.

Estos principios han tenido una importancia fundamental en cuanto a la protección de la privacidad, dado que buscan un equilibrio entre dos aspectos críticos, pero al mismo tiempo enfrentados, como son la Privacidad y el flujo y empleo libre de información *[14]*.

3.2. El Modelo de Notificación y Consentimiento (Notice and Consent)

Tal y como se destaca en *[14]*, los principios anteriores, en esencia, requieren que el procesado de información personal sea conforme a la Ley, lo que significa que o bien es un procesado que está permitido explícitamente por la Ley de aplicación en cada caso, o que la persona titular de los datos que se van a tratar ha dado su consentimiento, tras haber sido informado de del contexto y propósito de las operaciones que se van a efectuar.

No se trata por tanto sólo de recabar el consentimiento del usuario, sino que debe asegurarse que el mismo está debidamente fundamentado, es decir que el usuario al otorgar el consentimiento es plenamente consciente de qué tipo de análisis se van a hacer con sus datos.

Este modelo de Notificación y Consentimiento, que inicialmente se había definido como una más de las herramientas existentes para asegurar un tratamiento legítimo de los datos personales, se ha convertido en el elemento más importante y extendido en la normativa. Esta prevalencia de la Notificación y Consentimiento se ha producido especialmente tras la generalización del empleo de sistemas y tecnologías Web.

En todo caso, la forma que reviste este principio en la mayor parte de los casos prácticos está lejos de ser el mecanismo que iba a potenciar a las personas al permitirles tomar decisiones razonadas sobre el procesamiento de sus datos. En concreto se traduce en notificaciones de privacidad, clausulas y prácticas de uso extensas y difíciles de comprender para personas no expertas en materias legales, por lo que en muchos casos no es posible para ellas conocer en toda su extensión las repercusiones de conceder su consentimiento. Además en la mayor parte de los casos, la no aceptación de esas clausulas y prácticas supone la imposibilidad para usar el servicio deseado, y además en muchos casos no suelen existir alternativas, por lo que el usuario se ve en cierto modo ‘obligado’ a dar su consentimiento.

4. Impacto de los Big Data en los Principios Fundamentales de Protección de la Privacidad y la necesidad de su reenfoque

En un contexto como el derivado de la gestión de Big Data, los sistemas y soluciones de protección de la privacidad se hacen muchísimo más complejos, y los principios que tradicionalmente se han venido empleando (como los FIPP o la Notificación y Consentimiento) se enfrentan a una serie de retos a los que deben hacer frente de cara a garantizar su utilidad.

Tal y como se destaca en *[52]*, se necesita hacer uso de Datos de Carácter Personal (PII), pero, en cumplimiento de la Normativa en vigor en determinados países / regiones (como la *Directiva 95/46/EC* de Protección de Datos de la Unión Europea), es necesario mantener informado a los usuarios de qué análisis se pretende llevar a cabo con sus datos.

En ocasiones esto puede llegar a ser irrealizable, dado el número de usuarios implicados y que en determinados casos será imposible saber a priori qué tipos de análisis se van a hacer con sus datos, al no contarse con un objetivo a priori por ejemplo a la hora de buscar tendencias y patrones de actividad. De este modo, la participación individual y el recabar información de los titulares de los datos puede llegar a ser inviable, especialmente en entornos de gestión de Big Data donde se efectúan múltiples análisis, para los cuales las personas involucradas no pueden estar continuamente otorgando su consentimiento. Además puede ser contrario a actividades de análisis de Big Data que implican por ejemplo a la Seguridad Nacional.

Profundizando más en esta dificultad, una solicitud de transparencia, de acuerdo con la Legislación Europea no sólo cubre los datos personales de un individuo concreto, sino que afecta también a la documentación detallada de los procesos a través de los cuales se procesan dichos datos personales. Así, no sólo es necesario responder con los datos personales de una persona, sino también proporcionar detalles de los algoritmos y procedimientos involucrados en la analítica de Big Data que se hayan empleado en cada caso. Dado que muchos de estos procedimientos contienen algoritmos complejos de minería de datos, que incluso pueden considerarse como secretos empresariales de la organización que efectúa los análisis de los Big Data, esta obligación legal se convierte en un reto enorme para el procesado de Big Data.

En línea con las observaciones anteriores, en *[62]* se reconoce que la complejidad del procesamiento de datos y las posibilidades de las analítica moderna habilitada por los Big Data, limitan la conciencia de los titulares de los datos, su capacidad para evaluar las distintas consecuencias de sus elecciones y su libre e informado consentimiento (para que sus datos puedan ser analizados). Es más, el uso intensivo de los datos personales hace a menudo imposible dar una descripción de los usos potenciales de los datos en el momento de su recogida (no pudiendo cumplirse de esta manera la FIPP 3 de determinación del propósito).

En el escenario de la gestión de Big Data, los marcos de referencia tradicionales definidos en los años 90 del siglo XX han dejado de ser completamente válidos, dado que el nuevo contexto económico y tecnológico (la concentración del mercado en grandes empresas, la creación de clientes cautivos a nivel tecnológico y social, etc.) debilitan dos de sus pilares fundamentales: los principios FIPP de la especificación del propósito y la limitación de uso, y el modelo de notificación y consentimiento.

La normativa se ha ido adaptando a una sociedad que está cada vez más orientada al uso intensivo de datos. La alta demanda de información personal, la complejidad de las nuevas herramientas de análisis y el creciente número de fuentes de datos personales, han generado un contexto en el cual las grandes compañías, agencias gubernamentales, intermediarios, etc.

tienen control sobre la información digital que no está ya compensada por la autodeterminación del usuario.

Sin embargo todas las iniciativas en curso para reformar las regulaciones de protección de datos, tanto en la Unión Europea como en EE.UU., siguen estando focalizadas en los FIPP como pilares principales de las leyes de protección de datos, fundamentalmente la especificación del propósito y la limitación de uso, así como en el modelo de notificación y consentimiento.

Estas tendencias tecnológicas limitan de manera drástica el número de empresas y organizaciones que pueden proporcionar la clase de servicios que se demandan, con lo que en consecuencia las mismas tendrán cientos de millones de usuarios. Esta dimensión de los actores principales produce así mismo efectos de cliente cautivo, tanto a nivel tecnológico como social (si tus conocidos se relacionan a través de Facebook, lo más lógico es que tú también lo acabes empleando en lugar de otras redes sociales, o si tus contactos se comunican mayoritariamente mediante Whatsapp, es más complicado que tú pases a usar otros sistemas de mensajería como Line o Telegram). Esto incrementa la concentración de datos y representa limitaciones directas e indirectas para la capacidad de decisión y elección de los usuarios en relación con el uso de sus datos.

En este escenario descrito, caracterizado por un procesado complejo de datos y una concentración del control sobre la información, la decisión que subyace debe ser si mantener un modelo basado en la notificación y consentimiento representa un riesgo, dado que aunque en la práctica las empresas y organizaciones responsables de la gestión de los datos sí ofrecen notificación y buscan recabar el consentimiento de los usuarios, en la práctica éstas se hacen sin una verdadera capacidad de decisión y libertad de elección por parte de los usuarios (cliente cautivo).

4.1. Impacto de los Big Data sobre los FIPP

Las consideraciones anteriores hacen necesaria la revisión de los FIPP, para adaptarlos en la medida de lo posible a las nuevas condiciones que impone la gestión de los Big Data.

En la siguiente tabla se muestra el impacto que los Big Data suponen sobre los FIPP presentados en el *apartado 3.1.* de este capítulo [*] *[14]*.

FIPP	Impacto de los Big Data
1. Limitación en la recopilación de datos / Minimización de datos.	**ALTO** Se trata de un aspecto cuyo mantenimiento en los términos actuales no es compatible con la naturaleza de los Big Data.
2. Garantía de la Calidad e Integridad de los datos.	**BAJO** La redacción actual de este principio seguiría en principio siendo válida en el contexto de los Big Data.
3. Determinación explícita del Propósito para el que	**ALTO** No es posible a priori determinar todos los propósitos para

[*] *Se evalúa el impacto sobre el principio en sí y si se mantiene o no su validez; no se evalúa el impacto que los Big Data supondrían sobre su implementación práctica, que en todos los casos sería mayor, dada la naturaleza intrínseca de mayor complejidad que este nuevo método de gestión supone.*

se recopilan los datos.	los que una serie de datos personales se recopilan en un contexto de gestión de Big Data.
4. Limitación de uso.	**ALTO** Dada la relación con el principio anterior, también se trata de un principio que debería ser revisado para asegurar su aplicación en entornos Big Data.
5. Seguridad.	**BAJO** La redacción actual de este principio seguiría en principio siendo válida en el contexto de los Big Data.
6. Transparencia.	**BAJO** La redacción actual de este principio seguiría en principio siendo válida en el contexto de los Big Data.
7. Participación Individual.	**ALTO** En muchas ocasiones es imposible que los usuarios puedan saber los propósitos para los que serán usados sus datos, así como analizar detalladamente los algoritmos que se usarán en su procesado.
8. Responsabilidad y Auditoría.	**MEDIO** La redacción original se centra en el cumplimiento de los principios y de la normativa en vigor, si bien debería orientarse más hacia la administración responsable de datos y al uso de mecanismos (como las evaluaciones de impacto de privacidad) para asegurar el cumplimiento y demostrar el mismo a las autoridades responsables.

Tabla III-I – FIPP e Impacto de los Big Data sobre ellos

4.2. Impacto de los Big Data en el Modelo de Notificación y Consentimiento

En lo que respecta a la obtención del consentimiento, el análisis es similar, al planteado para el caso de los FIPP. Es imposible obtener el consentimiento de los usuarios de una manera eficiente (es decir sin afectar a la manera en que se desarrollan los análisis), ya que nuevamente a priori no se sabe para qué se van a usar los datos o en qué análisis se van a ver involucrados.

Dado que muchos tipos de analítica de Big Data se basan en algoritmos de minería de datos alta complejidad, el consentimiento informado y fundamentado implicaría que cada individuo debería recibir una explicación de dichos algoritmos para que pueda entender qué riesgos existen realmente para su privacidad. Esto puede considerarse un aspecto tremendamente complejo en el contexto de los Big Data, dada su naturaleza intrínseca (elevadísimo volumen de datos (muchas veces de distintos titulares) implicados, velocidad de actualización, necesidad de resultados rápidos, en muchas ocasiones en tiempo real, etc.).

Además, debe tenerse en cuenta la diferencia entre el consentimiento en la práctica y el consentimiento significativo; normalmente cuanto más simple sea el procedimiento para dar el consentimiento, menos usuarios podrán entender en toda su extensión lo que realmente están consintiendo, y cuanto más significativo se haga el procedimiento para dar el consentimiento

(suministrando suficiente información sobre lo que se hará con los datos), más dificultades supondrá para el posterior empleo de los datos de un modo ágil y práctico.

Puede verse que a priori, el **consentimiento es simplemente un enfoque poco apropiado para legitimar el procesado de datos en contextos online**. De hecho muchos expertos y legisladores parecen simplemente sugerir que deben buscarse medios para hacer el consentimiento online más informado, más consciente y más obligatorio para los proveedores.

4.2.1. La posibilidad de Obtención de Información de Fuentes Públicas

Los analistas de Big Data son capaces de extraer información de fuentes de datos personales públicas, por lo que aunque un usuario se haya negado a dar su consentimiento para el procesamiento de un registro concreto de sus datos (por ejemplo los que introduce al hacer una compra concreta online), de la información pública que puede haberse publicado en otros lugares, se puede obtener esa información.

Por ejemplo una persona compra online una tabla de surf, pero no da su consentimiento a que la empresa en cuestión haga un uso de sus datos personales; sin embargo por ejemplo de la información que publique en Facebook, los analistas de datos podrán saber que es una persona a la que le gusta el surf y que tiene una tabla nueva, y que por lo tanto puede ser susceptible de comprar otros productos (como por ejemplo viajes a zonas en las que se practique este deporte).

4.2.2. El problema con la Revocación del Consentimiento

Como otra dificultad adicional que supone el Modelo de Notificación y Consentimiento, se encuentra el hecho de que el consentimiento puede ser revocado en cualquier momento por el usuario (por ejemplo al haber tenido conocimiento de que una empresa que maneja sus datos se ha visto afectada por un caso de revelación de información sensible de sus usuarios). No obstante, aun habiéndose revocado el consentimiento de uso de esos datos, los mismos ya pueden haber sido ampliamente utilizados en diversos análisis, haberse empleado para la inferir nueva información (incluso relativa al propio titular) y divulgado en diferentes lugares.

Así, por ejemplo en un esquema como el siguiente:

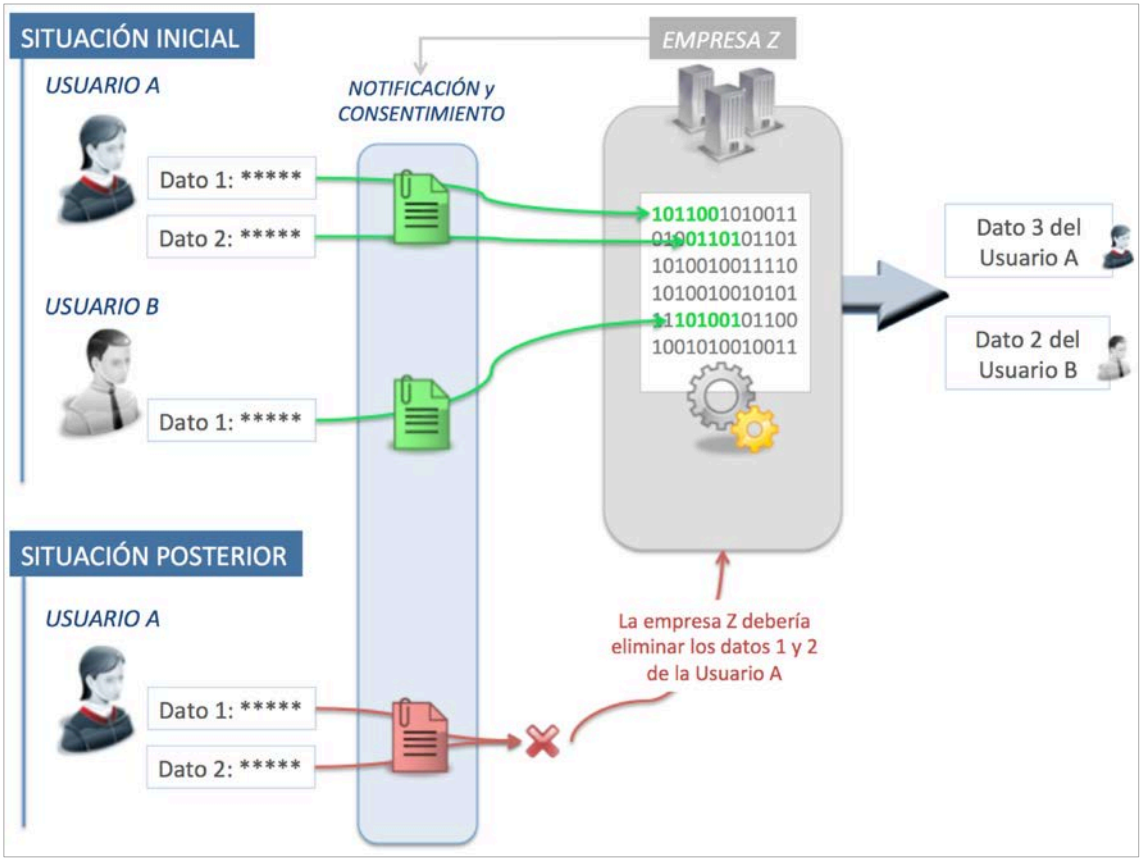


Imagen III-3 - Esquema aplicación Revocación Consentimiento

La usuario A puede retirar su consentimiento a que se usen sus datos 1 y 2, pero no sabe que con ellos se ha generado un nuevo dato 3, relativo a ella.

Del caso expuesto en la imagen anterior se derivan varios retos que deben ser planteados, en caso de que el consentimiento inicial fuera revocado:

- ¿Qué pasaría con el Dato 3 de la Usuario A (averiguado por la empresa gracias a sus procesos de análisis y combinación con otros datos)?
- ¿Qué pasaría con los datos de otros usuarios obtenidos gracias al empleo de los datos 1 y 2 de la Usuario A?

En principio, y en un escenario ideal, la respuesta a esas preguntas debería ser que deberían ser también eliminados, y por lo tanto la normativa debería estipular estos casos.

No obstante se trata de un aspecto que en muchos casos, y dada la dispersión y el grado de distribución y compartición de los datos que se produce, puede llegar a no ser factible. El derecho al olvido, que se prevé incluir como un derecho en la nueva normativa europea sobre Protección de Datos (tal y como se presenta en apartados siguientes) está relacionado con esa casuística y con él se ha demostrado la dificultad que puede darse para su implantación efectiva completa.

4.3. Conclusión

En resumen, muchas veces lo que establece la normativa es inviable y de todas formas en determinadas circunstancias se puede soslayar en el contexto de la gestión de los Big Data.

Por este motivo, es necesario un reenfoque de la Normativa a la vista de las nuevas posibilidades que ofrecen los Big Data.

Como ejemplo, se puede considerar una de las fuentes más comunes de Big Data, como es el tráfico de red. En este caso, para responder a una solicitud de transparencia, sería necesario analizar enormes listas de direcciones IP y sellos de tiempo, junto con otros tipos de datos como URLs a las que se haya accedido, cookies de sesión empleadas, etc. siempre buscando información que puede ser (o puede no ser) correlada para la persona que ha hecho la petición de transparencia. Obviamente esto no es factible de un modo sencillo, por lo que deberá aplicarse algún tipo de simplificación.

El haberse centrado en el principio de Notificación y Consentimiento puede haber impedido el desarrollo de otras alternativas de protección de la privacidad más eficientes. El problema es que las regulaciones y leyes que se aprueban en el mundo normalmente se centran en dicho principio. Evidentemente se debe tener en cuenta la importancia de la Notificación y Consentimiento, pero no emplearlo para todos los casos, sino sólo para aquellos en los que sea más factible y útil su empleo.

Esto hace necesario reconsiderar el papel de la capacidad de decisión de los usuarios y diferenciar las situaciones en las que los usuarios no tienen la posibilidad de entender con la suficiente profundidad los algoritmos y métodos empleados para el procesado de sus datos y los propósitos que se persiguen, o bien no se encuentran en una posición para decidir.

La extensión en el empleo de los Big Data supone un reto fundamental que desafía la validez del principio de Notificación y Consentimiento, dado que en este contexto buena parte del valor que tiene la información de carácter personal no es evidente en el momento de su recogida, cuando la notificación y consentimiento es normalmente requerida.

En la era de los Big Data uno de los retos fundamentales en lo que respecta a protección de la privacidad desde el punto de vista normativo, es que buena parte del valor que la información de carácter personal tiene no es evidente en el momento de su recogida, cuando la notificación y consentimiento es normalmente requerida.

Lo que en realidad se debe buscar es responsabilidad más que un mero cumplimiento normativo, y centrarse no tanto en la recopilación de los datos, sino en su uso. Volviendo al modelo de Notificación y Consentimiento, en ocasiones las organizaciones pueden escudarse en el consentimiento recabado del usuario en cuestión (que como ya se ha apuntado en muchas ocasiones no podrá conocer en toda su extensión las implicaciones de ese consentimiento, y en otras no podrá tener alternativas al consentimiento, si quiere usar dicho servicio que necesita), para eludir al menos parcialmente su responsabilidad.

En todo caso, a pesar de los límites de la notificación y consentimiento, muchos implicados sienten que este mecanismo puede continuar jugando un rol en el futuro, si bien modificado respecto a lo que es hoy en día. Por ejemplo, la notificación y consentimiento puede ser una herramienta clave para fortalecer la transparencia, aunque esto puede sugerir que la divulgación de datos personales a un regulador o a un repositorio central accesible, puede ser más eficientes que la notificación individual.

Apoyarse en los mecanismos de aplicación multinacionales (como la designación de agencias de aplicación, un órgano internacional de aplicación o medios de arbitraje), puede ayudar a construir responsabilidad y confianza transfronteriza, al mismo tiempo que se reducen los costes de aplicación y se evita la duplicación de actuaciones.

Existen aspectos distintivos a nivel nacional y cultural que afectan al modo en que la privacidad es entendida y que enfatizan el rol de las leyes nacionales de protección de datos, si bien debe

existir una interoperabilidad legal derivada de la amplia creencia compartida de que los individuos, las sociedades y los usuarios de datos se pueden beneficiar de una mayor coherencia e interoperabilidad entre los sistemas nacionales.

En la actualidad en múltiples servicios, de manera previa al registro o al uso, se debe aceptar las políticas de uso en las que se acepta el tratamiento de los datos personales del usuario por parte del propietario del servicio. Lógicamente se establece que este tratamiento se hará de acuerdo con la normativa legal vigente, si bien es probable que esta consideración no sea suficiente.

5. Situación Actual en relación con la Normativa de Protección de la Privacidad

5.1. Análisis de la Normativa en distintos Países y Organizaciones

Se expone en este apartado la situación en relación con la protección de la Privacidad en distintos Países y Organizaciones, centrándose con mayor detalle en el caso de la Unión Europea, y cómo se están enfocando los proyectos de revisión de la normativa para adaptarlos a los retos que suponen los Big Data, tal y como se ha indicado en los apartados anteriores.

Es un factor fundamental que la Legislación permita abordar de la forma más eficiente posible la protección de la privacidad en los sistemas y aplicaciones de gestión de Big Data, pero es importante que al mismo tiempo estas leyes no sean tan restrictivas que supongan limitaciones considerables a la innovación. Es decir, el reto en este sentido es cómo hacer que la Normativa pueda ser un elemento que realmente favorezca la protección de la privacidad de una manera ágil y dinámica, sin suponer trabas que reduzcan la eficiencia en el empleo y explotación de los datos.

5.1.1. La Unión Europea

- La referencia normativa fundamental en materia de Protección de Datos en la Unión Europea es la ***Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos***.

  Esta Directiva, al igual que la mayoría de la relativa a protección de la privacidad, toma como referencia los FIPP, adaptándolos y ampliándolos, resultando los siguientes principios fundamentales: *[13]*

  - El principio de la **limitación en la recopilación de datos**. Los datos deben ser procesados con un propósito específico y por consiguiente usados o comunicados sólo en caso de que esto no sea incompatible con el propósito por el que se recopilaron. Cuando los datos sean transferidos para propósitos de marketing directo, el titular de los datos debe tener la posibilidad de no autorizar (hacer 'opt-out') a que sus datos sean usados para dichos propósitos.
  - El principio de **calidad de los datos y de proporcionalidad**. Los datos deben ser exactos y cuando sea necesario, actualizados. Los datos deben ser adecuados, relevantes y no excesivos en relación con los propósitos para los cuales fueron transferidos y procesados.
  - El principio de **transparencia**. Las personas deben recibir información en relación con el propósito por el cual se procesan sus datos, así como de la identidad de la

organización que custodia sus datos, y cualquier otra información necesaria para asegurar la ecuanimidad de la relación.

- El principio de **seguridad**. Las medidas técnicas y de organización deben ser tomadas por parte del responsable de custodiar los datos y deben ser apropiadas para los riesgos que presentan cada procesado concreto. Cualquier persona que actúe bajo la autoridad de este responsable, no debe poder procesar datos, salvo en los casos en que dicho responsable se lo encomiende.
- Los derechos de **acceso, rectificación y oposición**. Los titulares de los datos deben tener derecho a obtener una copia de todos los datos relativos a ellos que han sido procesados por una organización, así como a rectificar aquellos datos que no sean exactos. En determinados casos, el titular debe poder oponerse al procesado de sus datos.
- **Restricciones a las transferencias futuras**. Las transferencias futuras de datos personales por parte del receptor original de los datos deben permitirse sólo cuando el segundo receptor esté sujeto también a las reglas que permitan un nivel adecuado de protección.
- **Tratamiento de datos sensibles**. Cuando se vean involucradas categorías sensibles de datos (como origen racial o étnico, opiniones políticas, creencias religiosas, convicciones éticas o filosóficas, etc., o que afecten a la salud o vida sexual), deben tomarse medidas de seguridad adicionales, como por ejemplo que el titular de los datos dé su consentimiento explícito al procesado.
- **Decisión individual automatizada**. Cuando el propósito de la transferencia sea la ejecución de una decisión automatizada, la persona afectada debe tener el derecho de conocer el procedimiento involucrado en esta decisión. Además en este caso, deben tomarse medidas adicionales para salvaguardar los intereses legítimos de la persona.
- Además se consagra el **Principio de la Supervisión Independiente** y el **Principio de Rectificación Individua**l.

- Desde 1995 cuando fue aprobada la *Directiva 95/46/EC*, se han producido una serie de cambios en la manera en que son compartidos y analizados los datos de carácter personal, especialmente debido a la extensión en el uso de tecnologías Web y en especial de nuevas y revolucionarias formas de procesado como Big Data, Cloud Computing, etc.

  Por este motivo, en 2010, la Comisión Europea publicó una Comunicación en la que se concluía, entre otros aspectos, que si bien sus principios centrales seguían siendo válidos, la Directiva no podía seguir respondiendo a los retos de los rápidos desarrollos tecnológicos y de la globalización *[20]*. De este modo y para dar respuesta a estos nuevos retos que se presentan, se ha propuesto a nivel de la Unión Europea una readaptación y evolución de la Directiva a través de una nueva ***Regulación General de Protección de Datos***. Esta Regulación se encuentra actualmente en proceso de aprobación.

  Como algunas de las novedades que esta Regulación prevé incluir cabe mencionar, tal y como se destacan en diversos análisis como *[30]*, *[55]* y *[81]*:

  - La Comisión Europea ha optado por una Regulación en lugar de por una Directiva, **para que no sea necesaria una transposición posterior** en leyes locales (como ocurrió para el caso de la *Directiva 95/46/EC*). De esta forma se asegura que la Regulación se aplicará directamente y de igual forma en todos los Estados miembros, y como consecuencia se prevé acabar con la fragmentación actual, eliminar cargas

administrativas y simplificar los procedimientos de gestión, especialmente para organizaciones que actúan en varios países de la UE.

- Se establecen tres nuevos objetivos que se persiguen, que se añaden a los objetivos originales perseguidos con la *Directiva 95/46/EC*:
  - **Incrementar la efectividad del derecho fundamental a la protección de datos y dar a las personas el control de sus datos**, particularmente en el contexto de los nuevos desarrollos tecnológicos y de la cada vez más implantada globalización.
  - **Fortalecer la dimensión interna del mercado de la protección de datos**, mediante la reducción de la fragmentación, fortaleciendo la consistencia y simplificando el entorno regulatorio, eliminando así costes innecesarios y reduciendo las cargas administrativas.
  - Establecer un **marco de referencia completo para la protección de datos**, que cubra todas las áreas.
- Se marcan **nuevas obligaciones de las organizaciones que custodian y procesan datos**, que deben realizar **Evaluaciones de Impacto en la Protección de Datos** (Data Protection Impact Assessments (DPIAs), e implementar **mecanismos de Protección de Datos por Diseño y por Defecto** (Data Protection by Design and by Default), que incluyen medidas como las técnicas de anonimización, control de acceso y auditorías.

  Las DPIAs son instrumentos útiles y obligarán a muchos controladores a pensar realmente de antemano en qué datos necesitan recopilar realmente y qué hacer con ellos, en lugar de aplicar aproximaciones de recabar todo lo que se pueda, y pensar más adelante en que usos se les podrá dar.

  Se requiere que una organización que gestiona datos adopte políticas y que sea capaz de **demostrar que el procesado de los datos se hace con arreglo a la Regulación**. Además estas organizaciones estarán obligadas a mantener documentación sobre todas sus operaciones de procesado de datos.
- Se reconocen nuevos derechos para los usuarios, como el **Derecho al Olvido** y el **Derecho a la Portabilidad de Datos** (posibilidad de transferir los datos de un titular y otras informaciones que haya facilitado, de un sistema de tratamiento electrónico a otro).

  En lo que respecta al derecho al olvido, se busca responder a diversos casos que han acontecido en los que usuarios han solicitado a diversas organizaciones el borrado de los datos que en relación con ellos disponían. Destaca en este sentido por ejemplo el caso de Max Schrems, un estudiante de Derecho austriaco que denunció a Facebook por conservar todos los datos que había publicado, y toda su actividad en esta red social, incluidos datos que había eliminado *[39] y [90]*.

  También destaca el caso de la denuncia contra Google España, cuando una persona solicitó que no aparecieran en las búsquedas los links con información sobre él *[103]*. En el plazo de 2 meses desde la resolución sobre Google España, Google recibió 91.000 peticiones de borrado que involucraban a 328.000 URLs, lo que da muestra de que los titulares de datos han ejercido activamente su derecho a solicitar el borrado *[55]*.
- A partir de la aprobación de la Regulación la ubicación física de la organización que procesa los datos no será ya relevante para determinar la aplicabilidad de la Ley

Europea de Protección de Datos. De hecho el borrador de Regulación (en contraposición a la actual *Directiva 95/46/EC*) se refiere no sólo a que el custodio de los datos deba estar radicado en la UE, sino que también lo esté o no la organización encargada del procesado efectivo de los datos.

Una organización que custodia información personal que no esté establecida en la UE, puede ser requerida también a cumplir con la Legislación Europea, cuando:

- Procese datos personales de personas que residen en la UE, y
- dichas actividades de procesado afecten a la prestación de bienes o servicios a titulares de datos de la UE, o se monitorice su comportamiento.

Se puede considerar (tal y como se pone de manifiesto en *[55]* y *[81]*), que el enfoque que se ha tomado en la nueva revisión tiene algunos puntos débiles:

- Se pone un **énfasis excesivo en el Modelo de Consentimiento de los Usuarios**, que como ya se ha expuesto en el *apartado 4.2.* tiene algunas limitaciones en un contexto de gestión de Big Data, y por lo tanto se convierte en un instrumento principalmente teórico sin una aplicación práctica eficiente.
  - Al emplear servicios de Internet, es ampliamente reconocido que la mayoría de los usuarios simplemente marcan las casillas de consentimiento sin asumir o entender realmente las declaraciones de privacidad.
  - En algunos casos se da por otorgado el consentimiento simplemente si el usuario por ejemplo continúa navegando por la Web en cuestión.
  - En muchas ocasiones para usar un servicio es condición obligatoria la aceptación de las condiciones y términos legales, y además no existen alternativas realistas para dicho servicio.
- El buscar que el control de la recopilación y el uso de los datos pase de las empresas y organizaciones a las personas, supone problemas derivados de que en muchos casos las personas no podrán saber exactamente lo que se haría con sus datos y más en un contexto de Big Data.
- En relación con las nuevas obligaciones para las organizaciones que custodian y gestionan datos personales, como la necesidad de elaborar DPIA o certificar el empleo de políticas de protección y el cumplimiento normativo, se corre el **riesgo de que sean considerados simplemente como hitos a cumplir, generándose documentos que en la práctica no supongan ninguna mejora real sobre la protección de los datos**.

  Este problema se deriva de que en la Legislación de Protección de Datos, el espacio entre la Ley en el papel y la Ley en la práctica es especialmente notorio.
- En lugar de permitir la diferenciación, la Ley de Protección de Datos de la UE aplica una **aproximación de todo o nada**: los datos son personales (siéndoles por tanto de aplicación toda la normativa) o no lo son, pero no es posible que se den posiciones intermedias, que permitirían una más eficiente protección de la privacidad, especialmente en lo que respecta a la búsqueda del equilibrio con la Utilidad en la gestión de los datos en cuestión.
- Las leyes de protección de la privacidad aplican únicamente a datos personales, es decir a datos relativos a una persona identificada o identificable. Pero **no está suficientemente claro si los principios centrales de la Regulación** (transparencia, consentimiento, minimización de datos, acceso, así como los nuevos derechos al

olvido y a la portabilidad de datos) **son de aplicación al nuevo conocimiento descubierto y derivado de datos personales**, gracias a los métodos de análisis de Big Data. Con ello se establece el dilema de si la Legislación en materia de Big Data debe cubrir sólo datos personales o también otros datos en principio no personales derivados y obtenidos del análisis de Big Data.

- Tal y como se destaca en *[62]* la solución propuesta por los legisladores de la Unión Europea está en línea con el enfoque tradicional centrado en los elementos del procesado de datos (naturaleza de los datos, categorías de los titulares de los datos, tiempo de procesado, etc.), en lugar de en las nuevas tecnologías empleadas y en la dificultad de los usuarios de ser consciente de todas las implicaciones que supone el procesado de sus datos.

  El foco debe estar en las herramientas empleadas para gestionar la información, y la dimensión de los datos recopilados sólo representa una condición complementaria. Así por ejemplo el criterio adoptado en la Propuesta de Regulación de la UE, según el cual se establecería un umbral de 5.000 usuarios, puede parecer subjetivo, dado que en función de la naturaleza de los datos, sería adecuado o no.

A pesar de estas consideraciones, algunas autoridades como el Information Commissioner's Office del Reino Unido (ICO – autoridad con responsabilidad en materia de protección de la privacidad) consideran *[98]* que la normativa existente actualmente sigue siendo válida y que simplemente hace falta adaptarla (lo cual es posible dada su inherente flexibilidad), a un contexto como el de los Big Data.

- Además de las normas principales anteriores, existen otras disposiciones de impacto en materia de Protección de la Privacidad, entre las que destaca la Directiva *2002/58/CE del Parlamento Europeo y del Consejo, de 12 de julio de 2002, relativa al tratamiento de los datos personales y a la protección de la intimidad en el sector de las comunicaciones electrónicas.*

  Esta Directiva dio respuesta al creciente intercambio de información a través de servicios públicos de comunicaciones electrónicas como Internet y la telefonía móvil y fija, así como de sus redes de apoyo. Estos servicios y redes exigen normas y salvaguardas específicas para garantizar el derecho de los usuarios a la intimidad y la confidencialidad.

  En la Directiva se establecen normas para garantizar la seguridad en el tratamiento de los datos personales, la notificación de las violaciones de los datos personales y la confidencialidad de las comunicaciones. Así mismo, prohibe las comunicaciones no solicitadas en las que el usuario no ha dado su consentimiento. Se fijan además obligaciones para los proveedores de servicios de comunicaciones electrónicas, que deben garantizar el acceso a los datos de carácter personal únicamente de las personas autorizadas, así como proteger los datos personales frente a formas de tratamiento no autorizadas y garantizar la aplicación de políticas de seguridad relativas al tratamiento de datos de carácter personal *[100]*.

5.1.1.1. España

En España la referencia fundamental la constituye la ***Ley Orgánica 15/1999, de 13 de diciembre de Protección de Datos de Carácter Personal (LOPD)***, transposición a nivel nacional de la *Directiva 95/46/CE*. De este modo adapta al caso español los preceptos y principios de esta Directiva.

Destaca además la creación de la Agencia Española de Protección de Datos (AEPD) como la autoridad estatal de control independiente encargada de velar por el cumplimiento de la

normativa sobre protección de datos. Garantiza y tutela el derecho fundamental a la protección de datos de carácter personal de los ciudadanos [2].

Además existen otras Leyes, como la *Ley 34/2002, de 11 de julio, de Servicios de la Sociedad de la Información y del Comercio Electrónico*, en la que también se regulan consideraciones de impacto sobre el tratamiento de datos personales, como por ejemplo el envío de marketing electrónico a usuarios de servicios digitales.

5.1.1.2. Reino Unido

En el Reino Unido la referencia fundamental la constituye el ***Data Protection Act (DPA) de 1998***, transposición a nivel nacional de la *Directiva 95/46/CE*, que permite adecuar al Reino Unido los preceptos y principios de la citada Directiva.

Además se ha establecido la figura de la Oficina del Comisionado de Información (Information Commissioner's Office – ICO-), como responsable de la aplicación de la Directiva 95/46/CE y de la DPA, y por ende de la protección de datos de carácter personal en el Reino Unido.

El ICO [99], ha establecido una serie de consideraciones clave respecto al procesado de datos personales en relación con la analítica de Big Data. En concreto se trata de:

- La necesidad del procesador de datos de adherirse al primer principio fundamental de la DPA de Legalidad y Legitimidad en el procesado.
- La necesidad del procesador de datos de informar claramente a los titulares de los datos del propósito para el que dichos datos se procesan.
- La necesidad de que el procesado de los datos esté alineado con el concepto de Minimización de Datos establecido en la DPA.

Tal y como se ha mencionado, la determinación del propósito del procesado de la información personal es muy difícil de determinar a priori en entornos de gestión de Big Data. Juega por lo tanto un papel fundamental la posibilidad de redefinir o modificar el propósito inicial para el uso de los datos, si bien esto debe hacerse desde un doble punto de vista:

- En primer lugar, se debe especificar el propósito para el que los datos se recopilaron, el cual debe cumplir con la legalidad vigente.
- En segundo lugar, si los datos son procesados para otro propósito diferente, éste no debe ser incompatible con el inicial.

El ICO considera que la limitación de propósito es algo más parecido a una limitación de no-incompatibilidad, es decir a establecer ámbitos funcionales o tipos de operaciones para los cuales un dato no puede nunca ser empleado, actuando de este modo como un medio para establecer salvaguardas en la gestión de Big Data (tal y como se reconoce en *EU Article 29 Data Protection Working Party, "Opinion 03/2013 on purpose limitation"* [21]).

Así por ejemplo, datos que inicialmente se procesaron para propósitos estadísticos u otros propósitos de investigación, no deberían estar disponibles para apoyar medidas o decisiones que se tomen en relación con los individuos titulares de los datos (a menos que esté explícitamente autorizado por dichos individuos).

Si la información que la gente publica en sus redes sociales se va a usar para evaluar los riesgos de su salud o su aptitud para recibir un crédito, o para ofrecerles publicidad de unos determinados productos, entonces, a menos de que sean convenientemente informados de esto y deban dar su consentimiento, es improbable que sea una situación justa o compatible.

5.1.2. Estados Unidos

En el caso de Estado Unidos, el enfoque que se emplea en relación con la protección de datos de carácter personal es algo diferente al de por ejemplo la Unión Europea, ya que inicialmente existen disposiciones diferentes según los distintos ámbitos funcionales, en lugar de una normativa global que abarque la protección de datos personales en cualquier ámbito (como ocurre con las normativas europeas). Así como ejemplos cabe citar:

- En el ámbito sanitario se encuentra en vigor el ***Health Insurance Portability and Accountability Act (HIPAA)***. Esta norma establece una serie de información que nunca puede publicarse en relación con registros de información médica para proteger la privacidad.
- El ***Sarbanes-Oxley Act,*** para el sector financiero,
- El ***Family Educational Rights and Privacy Act (FERPA)***, para el sector educativo.

Además el *Privacy Act de 1974*, establece el código de buenas prácticas que rige la recopilación, mantenimiento, uso y publicación de información personal que se encuentra en poder del Gobierno Federal.

Tal y como se destaca en *[38]*, al igual que ocurre en Europa, la recopilación en curso de información personal en los Estados Unidos sin las suficientes salvaguardas de privacidad ha conducido a notables incrementos en robos de identidad, violaciones de seguridad y fraudes financieros.

De manera adicional, el uso de información personal para la toma de decisiones automatizadas y la segregación de personas basada en factores secretos, imprecisos y a menudo no permisibles, presenta claros riesgos para la justicia e imparcialidad. De este modo se comprueba como en ocasiones se emplea la analítica predictiva habilitada por los Big Data para usos discriminatorios (cálculo de primas de seguros de salud, inclusión en una lista de personas a las que no se permite viajar en avión, etc.).

Un caso que atrae especialmente el foco de interés en EE.UU. en relación con la protección de la privacidad, es el de las entidades que acceden a Big Data asumiendo pocas responsabilidades, como por ejemplo los agregadores de contenidos que venden perfiles de usuarios que no están claramente protegidos en los marcos de referencia legales actuales. Destaca en este sentido por ejemplo la empresa Spokeo *[92]*, un servicio de búsqueda de personas que vende perfiles detallados de usuarios, incluyendo direcciones de email, direcciones físicas, números de teléfono, estado civil, ocupación, etc. Aunque Spokeo obtiene beneficios de la venta de perfiles de usuarios, no ofrece garantías de la exactitud de los perfiles. La profesora Anita Ramasastry de la School of Law de la Universidad de Washington, ha declarado que Spokeo y otros agregadores de información deben estar sujetos a algún tipo de regulación *[38]*. Como mínimo, los usuarios debería poder acceder a sus datos, para corregirlos si es necesario y comprender quién puede estar comprando sus datos para propósitos comerciales. La transparencia no es suficiente y son también necesarios mecanismos de supervisión.

Ante estas consideraciones, se ha identificado la necesidad de **actualizar las leyes actuales de protección de la privacidad** para minimizar la capacidad de recogida de información, securizar la información recopilada y prevenir abusos sobre los datos disponibles a través del uso de la analítica predictiva que se ha potenciado gracias a los Big Data.

Así, los siguientes principios deben ser tenidos en cuenta en las revisiones de la Normativa a acometer:

- **Transparencia**. Las organizaciones que recopilan datos personales deben ser transparentes en relación con dicha información que recopilan, cómo lo hacen, quién tiene acceso a ella, y cómo se supone que será usada. Más aún, los algoritmos empleados en la gestión de Big Data, también deberían ponerse a disposición del personal afectado.
- **Supervisión**. Deben implementarse mecanismos de supervisión independientes para asegurar la integridad de los datos y de los algoritmos que analizan los datos. Estos mecanismos ayudarían a asegurar la exactitud y la ecuanimidad de la toma de decisiones que se pueda hacer sobre la base de los datos personales.
- **Responsabilidad**. Las organizaciones que usen de manera indebida los datos o los algoritmos para hacer evaluaciones por perfiles con criterios que puedan ser discriminatorios, deben asumir responsabilidades por ello. Las personas deben poder recurrir a soluciones que resuelvan las decisiones injustas o no legítimas en relación con sus datos, siendo capaces de acceder fácilmente a ellos y corrigiendo la información errónea que se pueda haber recopilado sobre ellos.
- **Técnicas de Privacidad Robustas**. Deben potenciarse las técnicas que ayuden a aprovechar las ventajas de los Big Data, mientras al mismo tiempo se minimizan los riesgos sobre la privacidad. Estas técnicas deben ser, para asegurar su más eficiente implantación, robustas, escalables, verificables y adaptadas a las necesidades concretas que existan.
- **Evaluación Significativa**. Las organizaciones que emplean Big Data deben evaluar su utilidad regularmente y abstenerse de recopilar y almacenar datos que no son necesarios para los propósitos concretos que persiguen en un momento dado. Se ha argumentado que la recopilación masiva de datos no aporta beneficios significativos para la consecución de los objetivos de las organizaciones, y suponen un grave riesgo para la privacidad (en este sentido se destaca en *[38]* el programa de captación masiva de metadatos de la Agencia de Seguridad de los EE.UU. (NSA), que no ha tenido un papel tan significativo como se esperaba en la investigación contra el terrorismo).
- **Control**. Las personas deben poder ejercer el control sobre los datos que ellas crean o que están asociados con ellas, y decidir si esos datos deben poder ser recopilados, y en caso afirmativo como deben ser empleados (gestionados, almacenados, compartidos, etc.)

Además en la evolución de la normativa que se acometa se deben evaluar los avances que se están dando en la normativa europea (en la Regulación pendiente de aprobación, como el derecho al olvido o la privacidad por defecto y por diseño), no sólo dadas las ventajas que para una más eficiente protección de la privacidad supondrían, sino para un incremento en la imprescindible interoperabilidad. En todo caso será necesario evaluar la aplicabilidad práctica de estas medidas con carácter previo.

En el momento actual las Leyes sobre la Privacidad en EE.UU. y los principios autoregulatorios varían mucho, según el ámbito, pero en general requieren para la recopilación y uso de información personal, que con carácter previo se recabe el consentimiento de los usuarios. Además las reglas para el Opt-In (el consentimiento explícito para el empleo de sus datos personales que otorga un usuario) son de aplicación a casos especiales que afectan a información considerada sensible por la Legislación estadounidense, como información sobre la salud, el uso de créditos, o aquella relativa a menores de 13 años.

En EE.UU. no son de aplicación restricciones para la transferencia geográfica de datos, salvo en lo relativo a cierta información gubernamental, a diferencia de en buena parte de los

Estados Europeos donde las autoridades de protección de datos requieren que los datos transferidos a EE.UU. bajo el amparo del Acuerdo Safe Habor entre EE.UU. y la Unión Europea (Los principios internacionales Safe Harbor en materia de privacidad hacen referencia a un proceso de cooperación por el que las organizaciones de Estados Unidos cumplen con la *Directiva 95/46/CE* de la Unión Europea, relativa a la protección de datos personales), no pueden ser transferidos a su vez fuera de EE.UU. sin que concurra alguna circunstancia legal que lo haga necesario.

En el momento actual se encuentra en proceso de aprobación la ***Personal Data Notification and Protection Act***, propuesta por la Administración Obama como una norma para crear una norma uniforme a nivel nacional en el ámbito de la notificación de violaciones de la privacidad.

Actualmente, aparte de la ya comentada dispersión por ámbitos de especialidad (sanidad, educación, etc.), 47 Estados, tres territorios dependientes, y las ciudades de Washington D.C. y de Nueva York tienen sus propias leyes en relación con la notificación de violaciones de la privacidad.

Con esta Ley se requeriría que las organizaciones notificaran a las personas cuya información personal sensible (sensitive personally identifiable information –SPII-) haya sido, o sea razonable pensar que haya sido, adquirida o accedida sin autorización, salvo en el caso de que no exista un riesgo razonable de daño o engaño para esa persona.

Se incluye el concepto de la evaluación de Privacidad después de un incidente, según el cual en un plazo de 30 días desde el descubrimiento de una violación de seguridad, una organización deberá notificar a la Federal Trade Commission (FTC), como autoridad responsable de la protección de la privacidad, los resultados de una evaluación de riesgo o en su caso una solicitud de exención de la necesidad de abordar esa evaluación.

Además de asegurar que las personas reciben notificaciones puntuales en relación con las violaciones de seguridad que afectan a sus SPII, la propuesta de norma también obliga a las organizaciones a informar de las violaciones de seguridad a las autoridades gubernamentales, cuando se cumplan una serie de criterios.

5.2. Conclusión

Puede comprobarse como la normativa en la Unión Europea y en Estados Unidos (como casos paradigmáticos a nivel mundial), comparten actualmente una situación común.

Por un lado se basan en unos principios fundamentales básicos muy similares, lo cual responde a que en ambos casos se reconoce la validez y ventajas de las FIPP (siempre con las modificaciones que sean necesarias). De este modo, aspectos como la Transparencia para los usuarios de qué se hace con sus datos, la supervisión por parte de agencias públicas de las organizaciones que captan y analizan datos personales y la mayor responsabilidad que deben asumir estas últimas, son compartidos a ambos lados del Atlántico.

En ambos casos además se reconoce así mismo la necesidad de evolucionar las normas existentes para adaptarlas a los nuevos modos de procesamiento de datos personales y al enorme incremento en la generación de éstos, especialmente derivados de la generalización de las tecnologías de Big Data.

No obstante las características inherentes a los Big Data, hacen que las nuevas normativas que se elaboren no deban basarse en las FIPP y en modelos como el de Notificación y Consentimiento, al menos con su redacción actual, siendo necesario que las mismas evolucionen también.

6. Retos y propuesta de Soluciones

Como puede concluirse de los apartados anteriores, la evolución de la normativa que regula la protección de datos de carácter personal se enfrenta a una serie de retos para asegurar su adecuada adaptación a los nuevos entornos y modelos de gestión, como los Big Data, sin interferir de manera excesiva en el potencial de innovación que supone el análisis de los datos.

Entre ellos cabe destacar la necesidad de redefinición de los FIPP, la sustitución o al menos replanteamiento del Modelo de Notificación y Consentimiento, la potenciación del papel de los organismos públicos con responsabilidad en materia de protección de los datos personales, o la implementación efectiva de mecanismos como la Evaluaciones del Impacto de las diferentes actuaciones que se acometen en relación con Privacidad, o la Privacidad por diseño.

Aparte de estos retos, otro aspecto fundamental en el que será necesario seguir profundizando es el fortalecimiento de la Interoperabilidad Legal con respecto a la protección de la Privacidad.

6.1. Reto de la Interoperabilidad Legal en relación con la Protección de la Privacidad

La coexistencia y la mayor o menor homogeneidad en cuanto a la protección de datos de carácter personal en distintos países del mundo es un aspecto fundamental, especialmente en un contexto como el actual, donde la globalización y el empleo masivo de Tecnologías de la Información y Comunicaciones posibilita un intercambio de datos intensivo entre organizaciones y empresas radicadas en las regiones más diversas del planeta.

Por este motivo se deben buscar mecanismos que permitan un mayor alineamiento entre legislaciones y de este modo una protección más homogénea de la privacidad, evitando o al menos reduciendo todo lo que sea viable la posibilidad de que ciertas organizaciones puedan acogerse a 'sombras legales' en las que las exigencias de protección requerida sean menores.

Como un caso a destacar en este sentido se encuentran los ya mencionados en apartados anteriores principios internacionales Safe Harbor en materia de privacidad, que representan un paso importante para potenciar la cooperación, al comprometerse las organizaciones y empresas de Estados Unidos a cumplir con los preceptos de la *Directiva 95/46/CE* de la Unión Europea.

La Unión Europea desarrolló un documento de trabajo sobre Transferencias de datos personales a terceros países, en aplicación de los artículos 25 y 26 de la Directiva sobre protección de datos de la UE. Aprobado por el Grupo de Trabajo el 24 de julio de 1998 *[22],* en el que se desarrollan los criterios por los que un país externo a la Unión Europea puede ser considerado como con un adecuado nivel de protección de la información personal.

La principal consecuencia de que un país sea declarado adecuado es que se podrán transferir datos desde los Estados miembros de la Unión Europea sin necesidad de ningún tipo de trámite o autorización especial. A día de hoy en la Unión Europea se han considerado como países con un adecuado nivel de protección los siguientes: Andorra, Argentina, Canadá (Sector privado), Suiza, Islas Feroe, Guernsey, Israel, Isla de Man, Jersey, Estados Unidos (Entidades que cumplen el Acuerdo Safe Harbor), Nueva Zelanda y Uruguay *[3]*.

Además también será necesario asegurar la compatibilidad con otras normativas relacionadas con la Conservación de Datos (plazos, modos de conservación, garantías de seguridad a aplicar, etc.) *[87]*.

6.2. Soluciones Posibles

La Declaración de Madrid de noviembre de 2009 *[78]*, en la que Grupos de la Sociedad Civil y expertos en Privacidad, ratificaron la importancia de las Leyes Internacionales de Protección de la Privacidad, identificaron nuevos retos y solicitaron acciones concretas para salvaguardar la Privacidad, constituye de este modo un compromiso para la protección de la privacidad.

En ella se reafirma el apoyo a continuar con la implementación de las FIPP, así como de nuevas técnicas protectoras de la Privacidad, de las Evaluaciones de Impacto sobre la Privacidad, y a potenciar el rol de las autoridades independientes para la protección de datos.

6.2.1. La Política Do-not-Track del W3C

Do-not-Track es una tecnología y una propuesta de política de uso de datos que permite a los usuarios poder decidir no ser monitorizados y que sus datos no sean empleados por terceros cuyos servicios no utilizan o cuyas webs no visitan (materializando de este modo el denominado opt-out), incluyendo sistemas de análisis de Big Data, redes de publicidad y redes sociales *[37]*. En el momento actual, pocos de esos terceros ofrecen una opción fiable para hacer opt-out por defecto, y las herramientas para bloquearlos no son ni amigables desde un punto de vista de usuario ni completas.

En este sentido Do-not-Track ofrece a los usuarios una opción simple y persistente para hacer opt-out respecto a un tercero que pueda tener interés en la recopilación y análisis de sus datos, reseñando la preferencia de opt-out de un usuario con una cabecera HTTP, lo cual la hace compatible con los diseños Web existentes.

Mientras algunas organizaciones se han comprometido a respetar las prácticas de Do-not-Track, otras no lo han hecho. En febrero de 2012 los principales grupos de publicidad online de EE.UU. se comprometieron ante el Gobierno de ese país a apoyar la política de Do-not-Track antes del fin de ese año, si bien dicha promesa permanece incumplida *[37]*.

De hecho, los esfuerzos para estandarizar el Do-not-Track en el seno del World Wide Web Consortium (W3C) se encuentran en un punto muerto, a pesar de que los reguladores en Europa y los Estados Unidos suelen abogar por completar su desarrollo, y se han producido ejemplos que denotan la falta de apoyo de importantes empresas a esta política (como el caso de Microsoft que ha deshabilitado la opción de Do-not-Track por defecto en la versión 10 de Internet Explorer). Se comprueba de este modo como los controles de privacidad y la creciente transparencia no han respondido a las preocupaciones derivadas de la clasificación y segmentación que habilitan los análisis de los Big Data. *[31]*, *[56]*.

En muchos casos no importa que se haya dado consentimiento (dado que en muchas ocasiones es imposible conocer todo lo que se puede llegar a hacer con los datos), sino que hay que proteger que, se hayan conseguido como se hayan conseguido los datos, se garantice un mínimo de protección sobre la privacidad (balanceada lógicamente con la utilidad).

Es más, la tendencia actualmente tiende a defender que la protección no debe centrarse tanto en impedir conocer la identidad de una persona, sino en evitar que esa identificación pueda conllevar consecuencias negativas para él/ella, por ejemplo aplicándole una tasa mayor en su seguro. De esta manera, implementado prácticas como el Do-not-Track, lo que la normativa permitiría reducir este tipo de impacto discriminatorio de los análisis que se hagan con los Big Data.

6.2.2. El Opt-Out

Estrechamente ligado con la Política Do-not-Track, otro de los pilares fundamentales que los expertos en protección de datos están recomendando es la adopción de esquemas de opt-out y la definición asociada de una evaluación rigurosa de la protección de datos, que se encontraría disponible públicamente *[62]*.

Con respecto a esta última, la misma aproximación que se emplea en el ámbito de la seguridad de productos (por ejemplo la autorización para la comercialización de medicamentos), debe extenderse al procesamiento de datos. Ante un sistema complejo de procesado de datos o de recopilaciones de datos dentro de efectos de cliente cautivo, la evaluación de riesgos y beneficios no debe ser hecha por los usuarios, sino por compañías bajo la supervisión y control de las autoridades de protección de datos. Para asegurar el éxito de esta medida, debe garantizarse la autonomía política y financiera de las autoridades de protección de datos, respecto de gobiernos y fundamentalmente empresas.

Los usuarios únicamente decidirían ceder su información o ejercer su derecho al opt-out, en función del resultado de esas evaluaciones.

En *[62]* se reconoce la necesidad de proceder a una revisión del principio de Notificación y Consentimiento, que actualmente está centrado en el opt-in (el usuario da consentimiento a que sus datos se empleen en análisis). Por el contrario se propone un enfoque diferente basado en el opt-out (los usuarios optan explícitamente por la opción de que un determinado proveedor de servicios no pueda en ningún caso hacer uso de sus datos) y en un control preventivo más profundo por parte de las autoridades de protección de datos, que deberían ser adoptadas cuando el titular de los datos no pueda estar completamente seguro de las herramientas de análisis y sus efectos potenciales. Puede comprobarse como este esquema de opt-out está estrechamente relacionado con la ya mencionada práctica del Do-not-Track.

Para el caso en que los usuarios no puedan ser capaces de entender en profundidad el sistema de procesamiento, su propósito e implicaciones, el rol de las autoridades independientes debe ser incrementado. Las autoridades de protección de datos, en lugar de los usuarios, tienen el conocimiento tecnológico para evaluar los riesgos asociados al procesamiento de datos y pueden adoptar soluciones legales para afrontarlos. Además, se encuentran en una mejor posición para equilibrar los diferentes intereses en juego en relación con los proyectos de recopilación y minería de datos.

En los otros casos, el modelo de notificación y consentimiento, tal y como se diseñó inicialmente, puede aún ser efectivo, aunque es necesario que se refuerce incrementando su transparencia, la responsabilidad de los proveedores de servicios y las arquitecturas orientadas a la protección de datos.

6.2.3. Otras aproximaciones

Además es necesario seguir analizando la posibilidad de aplicar nuevos modelos, que den una respuesta más eficiente a las diversas consideraciones que emergen del actual entorno digital. Entre ellos cabe mencionar:

- Privacy-by-design *[15]*.
- La Privacidad Contextual *[72]*.
- El uso de los datos *[14]*.
- Soluciones combinadas.

CAPÍTULO IV

El Alineamiento Tecnología – Normativa en Relación con la Protección de la Privacidad en Big Data

Enfoques Innovadores para garantizar la coexistencia Privacidad – Utilidad en la gestión de Big Data

1. Introducción y Objeto del Capítulo.

En este capítulo, y una vez planteada la situación actual en relación con la protección de la privacidad en la gestión de los Big Data desde el enfoque tecnológico y el normativo, se plantea una propuesta de solución para asegurar un alineamiento entre dichos enfoques en aras de una más eficiente protección.

La solución que se presenta se enfoca desde un punto de vista funcional explorando, con una perspectiva holísitica, la forma más adecuada en la que podrían afrontarse algunos de los retos más importantes derivados de los análisis de los capítulos anteriores, al tiempo que se mantiene en la mayor medida posible las ventajas y posibilidades que ofrecen los Big Data

1.1. Esquema de estructura del Capítulo

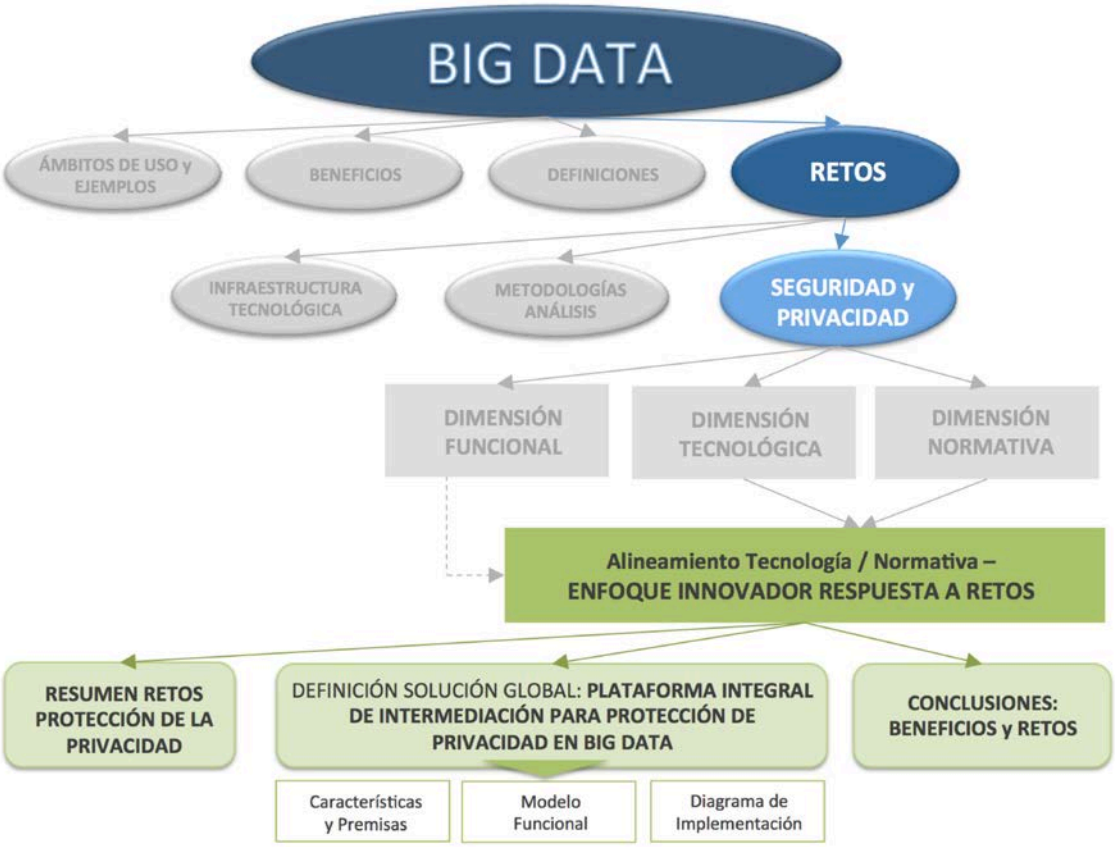


Imagen IV-1 – Esquema Estructura Capítulo IV

2. Resumen de retos abiertos en relación con la protección de la Privacidad

Como conclusión de los capítulos anteriores, se puede comprobar como ninguno de los métodos existentes, ni tecnológico, ni normativo, pueden proteger de manera independiente y sólida la Privacidad en el extremo necesario, y al mismo tiempo garantizar la imprescindible utilidad de los análisis que se hagan sobre los datos, que al fin y al cabo constituyen la verdadera razón de ser y beneficio que pueden aportar los Big Data.

Actualmente la protección de la privacidad se centra en dos amenazas:

- La que representa el acceso a los datos de **atacantes externos que buscan el robo de información o la alteración del proceso correcto de ejecución de trabajos**. Los mecanismos de control de acceso existentes y su evolución para hacerlos más eficientes en contextos de Big Data y computación distribuida (múltiples puntos de autenticación, ejecución de trabajos en sistemas distintos a aquél en el que se produjo la autenticación, uso concurrente de un mismo nodo por distintos usuarios, dificultad para determinar con exactitud el lugar en el que se ejecuta un trabajo, etc.), suponen la línea de trabajo para hacer frente a estas amenazas.

  Se trata de un enfoque de protección de la Privacidad que se articula alrededor del refuerzo de las distintas dimensiones de la Seguridad de la Información (Confidencialidad, Integridad, Disponibilidad).

- La que suponen **usuarios autorizados**, que pudiendo acceder a los datos, hagan un uso de ellos que vaya más allá del propósito inicialmente previsto y pueda suponer una violación de la privacidad de las personas cuyos datos se encuetran involucrados.

  En este caso la respuesta a estas amenazas 'no tradicionales' (ya que no buscan el robo o la alteración de la información), será más compleja y multidimensional, ya que no puede abordarse únicamente desde puntos de vista de refuerzo de la Política de Seguridad de la Información como el caso anterior, sino que debe implicar otros aspectos que pueden ir desde la formación y concienciación de usuarios, el replanteamiento del modelo de negocio de empresas proveedoras de servicios, el cumplimiento normativo, además de lógicamente elementos tecnológicos como el cifrado o la anonimización de datos.

Este capítulo, siguiendo la línea general del proyecto, se va a centrar en el segundo de los aspectos anteriores, partiendo de los diversos retos que en relación con él existen:

- Desde el punto de vista **Tecnológico**:

  Por si misma la **Criptografía, así como cualquier otra herramienta o método de seguridad de la información de la capa lógica, no puede resolver completamente los problemas de protección de la privacidad en toda su extensión** y se deben completar con otros mecanismos como el hardware seguro a prueba de cualquier ataque, o los ecosistemas complejos de confianza, dando lugar a una verdadera aproximación holística.

  El cifrado tiene un elevado impacto en la eficiencia del proceso de análisis de los Big Data, al necesitarse más recursos y tiempo para procesar datos cifrados que no cifrados, además de que puede hacer inaccesibles los datos a algunos analistas externos que no tengan acceso a las correspondientes claves de descifrado. Todo ello puede reducir considerablemente las ventajas de los Big Data, por lo que serán necesarios nuevos enfoques en relación con el uso de la criptografía en este ámbito, como por ejemplo la evolución de los modelos actuales de Criptografía Homomórfica, mejorando el rendimiento y el consumo de recursos de la computación de datos cifrados.

  Otro reto fundamental consiste en **adaptar las tecnologías habilitadoras de la privacidad** presentadas en el *Capítulo II* (métodos estadísticos de anonimización, métodos de anonimización grupales, control del resultado de las consultas, etc.), a contextos de datos no estructurados, fundamentalmente al **Internet of Things y a las redes sociales**. Estas últimas, dadas sus características inherentes (posibilidad elevada de inferencia de datos de otras entidades de información, dificultad para anonimizar estructuras de grafos sin una pérdida considerable de información, necesidad de anonimizar no sólo información propia de una persona, sino también relativa a su relación con otras personas dado que se ella sería posible extraer información privada, etc.),

suponen la necesidad de replantear las distintas técnicas de anonimización y protección de la privacidad empleadas actualmente con los datos estructurados, para asegurar que se mantiene su validez y utilidad.

Por otra parte, se debe tener en cuenta que en la actualidad muchas de las herramientas que se están empleando para que los usuarios controlen su privacidad suponen una reducción en las capacidades que se le ofrecen y una reducción de la potencialidad que los Big Data aportan. Así por ejemplo las tecnologías que bloquean el rastreo de navegación, si bien incrementan la privacidad, también deshabilitan aspectos útiles como los log-in preestablecidos o las características web personalizadas, las cuales mejoran la experiencia de usuario a la hora de navegar *[94]*.

De este modo mecanismos como la navegación en modo seguro, el bloqueo de cookies o el uso de herramientas de búsqueda seguras (con navegadores como DuckDuckGo *[33]*, que no ofrece resultados de búsqueda diferentes en función de características personales del usuario, como su ubicación geográfica, o Ixquick *[51]* que no almacena la dirección IP del usuario o las búsquedas realizadas y no vende o cede dicha información a otras empresas, o el plug-in contra seguimiento de la empresa Disconnect *[29]*, que anonimiza las consultas que un usuario haga en cualquier motor de búsqueda), suponen mecanismos que, si bien incrementan la protección de la privacidad, lo hacen a costa de perder funcionalidades y características que pueden ser beneficiosas para los usuarios.

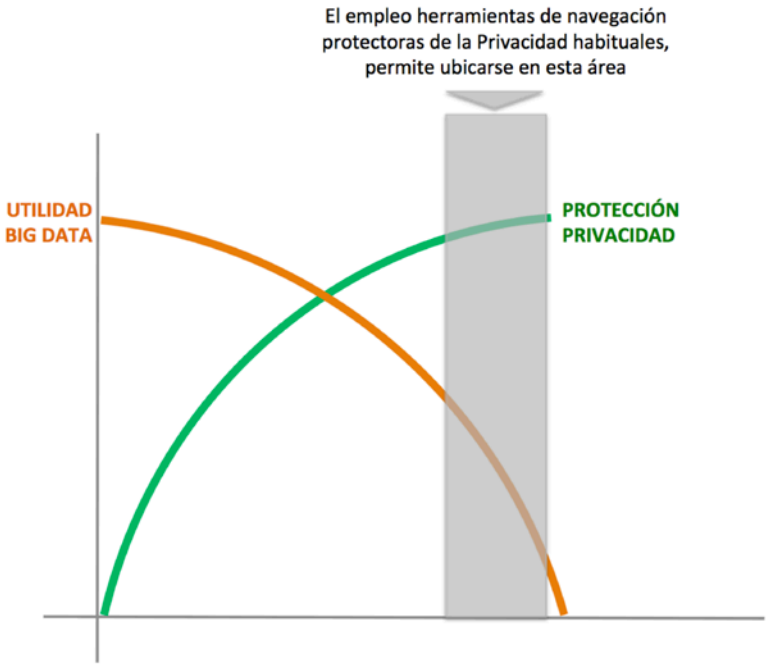


Imagen IV-2 – Zona ubicación soluciones actuales de navegación protectora de la Privacidad

Es por tanto necesario la búsqueda de herramientas protectoras de la privacidad que impacten lo menos posible en la experiencia de usuario, permitiendo mantener la utilidad de los Big Data y las características que habilitan, al mismo tiempo que se incrementa la protección de la Privacidad. Ser objeto de seguimiento de nuestra actividad en Internet y recibir publicidad personalizada de servicios y productos se ha considerado como una contraprestación por poder usar un Internet gratuito. Además en ocasiones este seguimiento ofrece ventajas a los usuarios. No obstante es imprescindible que el mismo

no se extralimite, respete unos parámetros de privacidad mínimos y se haga siempre dentro del conocimiento del usuario, o en caso de no ser posible, dentro de los limites generales que dicho usuario haya establecido en relación con la protección de sus datos personales.

- Desde el punto de vista **Normativo**:

  En este caso los principales retos se centran en la búsqueda de maneras eficientes (y en ocasiones factibles) de **aplicar las FIPP y el principio de Notificación y Consentimiento**, que siguen siendo la base fundamental de las regulaciones en vigor y de las evoluciones que se encuentran en proceso de aprobación, tal y como se ha expuesto en el *Capítulo III*. La imposibilidad para los usuarios de comprender todas las actuaciones susceptibles de realizarse con sus datos (y sus implicaciones sobre su privacidad y sobre si mismos), hace irrealizable una eficiente aplicación de las FIPP y de la notificación y consentimiento en contextos Big Data.

  Relacionado con el principio de Notificación y Consentimiento, otro aspecto que se destaca en la normativa en elaboración en este campo (fundamentalmente en la nueva Regulación Europea) es la búsqueda de un incremento de la responsabilidad de las empresas que recopilan datos. Éstas no deben poder escudarse en que un usuario ha otorgado su consentimiento para dejar de acometer las tareas de seguimiento y control de qué se hace con los datos que recopilaron y que puedan haber cedido a terceras organizaciones o empresas.

  Para la certificación del cumplimiento de esta responsabilidad, la **labor de las Autoridades Públicas de Gestión y Protección de la Privacidad es fundamental**.

En definitiva, cualquier solución que se plantee en relación con la protección de la privacidad debe buscar la posibilidad de **incrementar la transparencia, la responsabilidad, la supervisión y el control y el empleo de técnicas habilitadoras de la privacidad más eficientes y robustas**.

3. Solución global propuesta: La Plataforma Integral de Intermediación para la Protección de la Privacidad en Entornos Big Data.

3.1. Características y Premisas Generales para la Plataforma Integral de Intermediación para la Protección de la Privacidad en Entornos Big Data

3.1.1. La Importancia de un Enfoque Holístico y Generalista

A tenor de los retos anteriores, es posible concluir que una solución que se enfoque únicamente en aspectos tecnológicos sin tener en cuenta una adecuada evolución del ámbito normativo y procedimental de aplicación no será optima (en términos como ya se ha indicado de proteger la privacidad y al mismo tiempo asegurar la máxima utilidad posible de los análisis que se hagan sobre los Big Data concernidos). Del mismo modo tampoco lo sería la situación contraria, en la que se pretendiera dejar la mayor parte del peso de la protección de la Privacidad en la dimensión normativa, sin reforzar y evolucionar adecuadamente los sistemas tecnológicos implicados.

Centrándose en el eje *Usuario Final – Organización Gestora de Datos*, tampoco sería eficiente abordar posibles soluciones que se orientaran de un modo claro hacia uno de los extremos, siendo imprescindible implicar decisivamente a ambos actores en el enfoque que se decida.

De este modo, deben buscarse soluciones intermedias a lo largo de estos dos ejes *Usuario – Final – Organización Gestora de Datos* y *Tecnología – Normativa*, de manera que, gracias a un **enfoque holístico**, se aprovechen las fortalezas de cada uno de estos extremos y se trate de minimizar en la medida de lo posible sus debilidades.

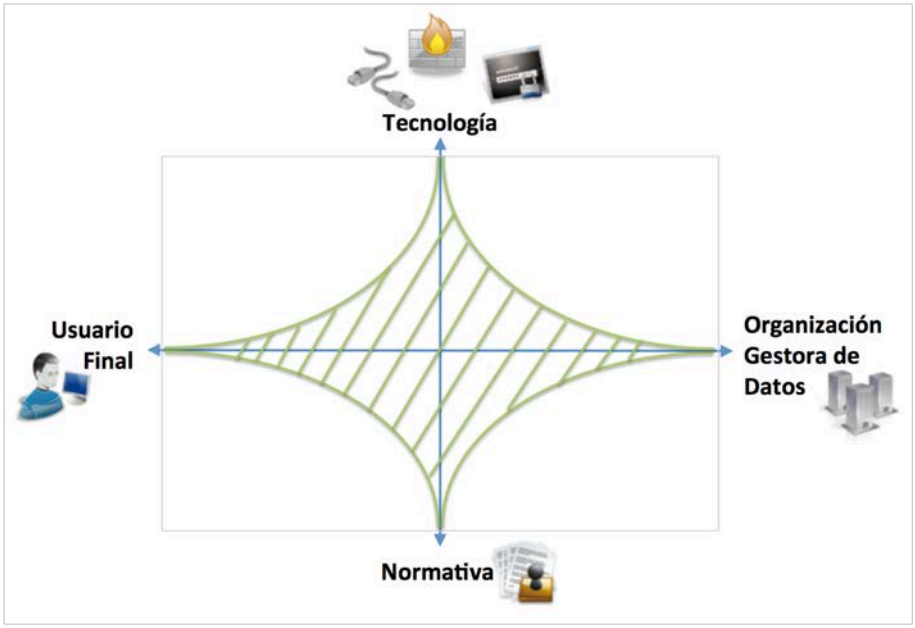


Imagen IV-3 – Esquema Ubicación Soluciones Óptimas Holísticas

Evidentemente, en función de cada caso particular, el peso de la solución a adoptar bascularía dentro del espacio acotado marcado en la imagen anterior. Por ejemplo, en algunas ocasiones en caso de que esté implicada algún tipo de información que requiera una protección especial (como la información médica), será necesario un mayor refuerzo en la dimensión normativa, aun a costa de perder algo de agilidad y flexibilidad en las soluciones a adoptar.

Por otra parte y dada la multiplicidad de usos y aplicaciones que puede tener un mismo conjunto de datos en función de las circunstancias concretas de cada momento, o de quién analice y explote los mismos, es importante poder contar con sistemas de protección los más **generalistas** posibles. Con la modulación adecuada estos sistemas permitirían una adaptación a diferentes casuísticas.

3.1.2. El punto de equilibrio óptimo entre Protección de la Privacidad y Utilidad

Además de las características y premisas anteriores, y como elemento decisivo, que como idea fuerza principal ya se ha destacado a lo largo de todo el proyecto, la **solución planteada debe también enfocarse con el objetivo principal de conseguir un equilibrio entre la Protección de la Privacidad y la Utilidad de los Big Data**:

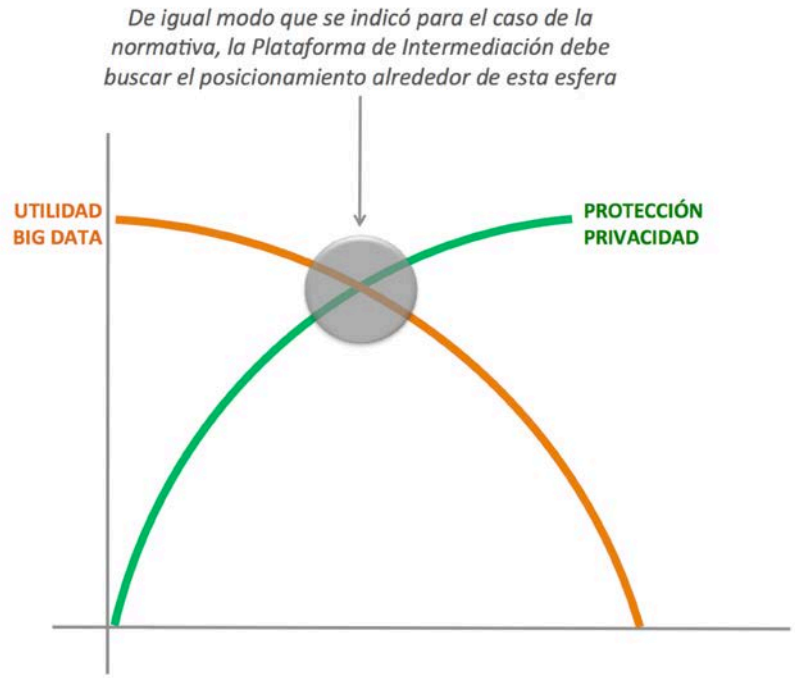


Imagen IV-4 – Punto de Equilibrio Utilidad / Protección Privacidad

3.1.3. Cobertura del Ciclo de Vida Completo de los Datos y enfoque Proactivo

Como una de las buenas prácticas fundamentales que se han identificado para la protección de la privacidad, por ejemplo en enfoques como el Privacy-by-design, reconocidos por ejemplo en la futura Regulación General de Protección de Datos de la Unión Europea, se establece que los elementos de protección deben abarcar el Ciclo de Vida completo de los Datos *[16]*:

- **Generación**, por ejemplo cuando un usuario rellena un formulario en una Web, hace una búsqueda en Internet o publica información en una Red Social.
- **Recopilación *I* Almacenamiento**: Por parte de una empresa u organización proveedora de los servicios que demanda el usuario cuyos datos personales se encuentran involucrados.
- **Análisis e Inferencia de resultados**.
- **Uso de los resultados de los análisis.**
- **Destrucción y borrado seguro de datos.** En este punto será especialmente destacado abordar el reto planteado en el *Capítulo III* de cómo eliminar no sólo los datos sobre los que un titular haya retirado los derechos de acceso y análisis, sino qué hacer con los datos que puedan haber sido inferidos de los primeros.

Además, en una solución de este tipo se deben cumplir el resto de los preceptos del Privacy-by-design, ya que éste permite gracias a su modularidad y adaptabilidad a distintos entornos tecnológicos, contemplar desde la misma definición y diseño de los mismos, las distintas cuestiones relativas a Privacidad que van a resultar de aplicación, empleando enfoques proactivos en lugar de reactivos *[16]*.

3.2. Modelo Funcional de una Plataforma Integral de Intermediación para la Protección de la Privacidad en Entornos Big Data

Partiendo de las características y premisas anteriores, se va a diseñar una Plataforma Integral de Intermediación para la Protección de la Privacidad, describiéndola desde un enfoque funcional.

Para ello en primer lugar se van a presentar los diferentes actores involucrados en el proceso, para a continuación dar una primera aproximación al modelo general de relación entre dichos actores en un contexto de generación y uso de Big Data.

3.2.1. Actores Involucrados

En una solución de este tipo, y tal y como se ha destacado en el *apartado 3.1.*, es imprescindible analizar el rol que jugarían los distintos actores involucrados. La siguiente imagen esquematiza el proceso de relación simplificado entre dichos actores:

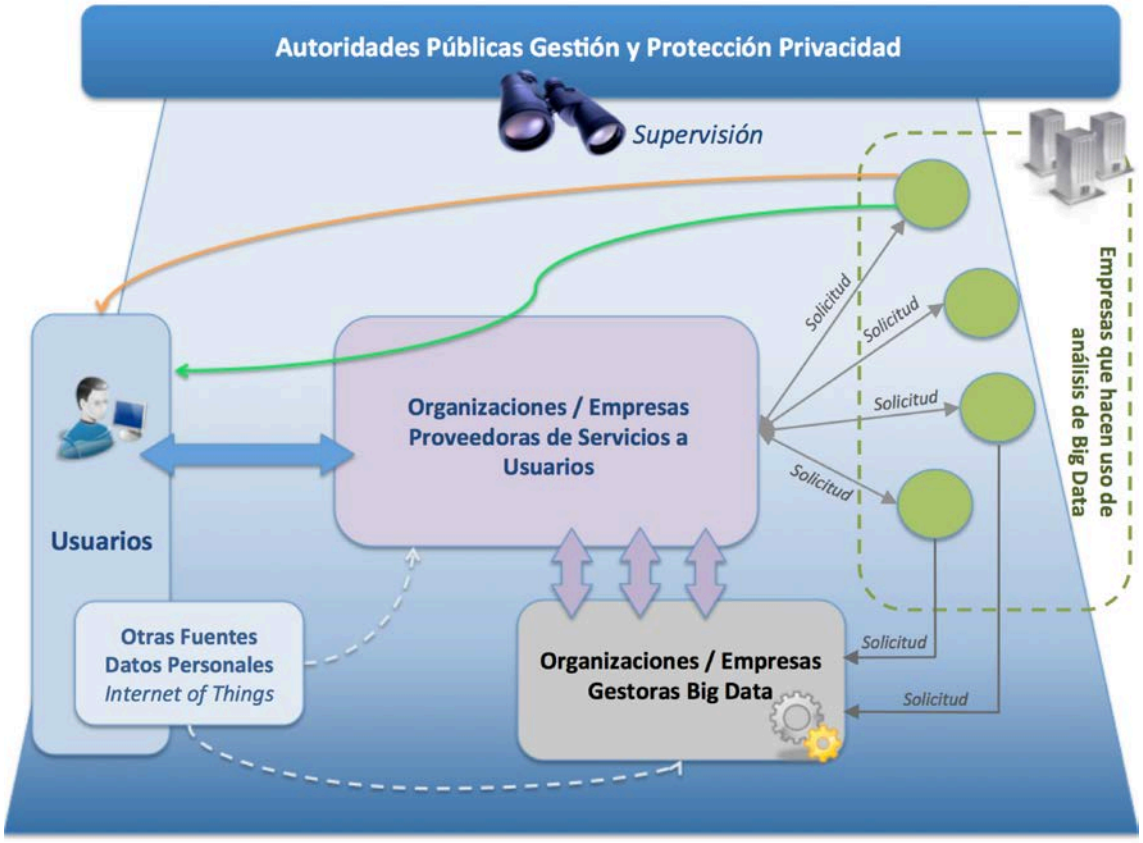


Imagen IV-5 – Esquema de Relación Actores Involucrados Plataforma de Intermediación para la Protección de la Privacidad en Entornos Big Data

- **Usuarios**. Es el actor central del proceso en un contexto centrado en la Protección de la Privacidad. La normativa ya lo reconoce así, y el enfoque aquí presentado busca mecanismos para incrementar sus capacidades para que sus datos personales se vean protegidos, así como para que al mismo tiempo pueden recibir servicios, ofertas e información que derivada de sus datos personales, puedan resultarles de utilidad.

  En este sentido destaca el caso especial que suponen los datos personales no generados conscientemente por los usuarios (por ejemplo datos acerca de su

posicionamiento geográfico, dirección IP, etc.) que gracias a la explotación que se hace del Internet of Things, complementan los análisis que las empresas proveedoras de servicios y las gestoras de Big Data pueden hacer, ofreciendo resultados más precisos, pero constituyendo a la vez un mayor riesgo para la privacidad.

- **Organizaciones / Empresas Proveedoras de Servicios a Usuarios**.
- **Organizaciones / Empresas Gestoras de Big Data**. Son aquellas que procesan, analizan, almacenan y extraen información de los datos que reciben. Pueden coincidir con las empresas proveedoras de servicios a usuarios, que en determinadas ocasiones disponen de capacidad de procesado suficiente como para desarrollar ellas mismas sus propios análisis, como por ejemplo ocurre en el caso de grandes empresas como Google o Facebook.

  Se trata de una situación que, al reducir un actor en el proceso simplifica las operaciones y flujos de protección en el contexto de la Plataforma Integral de Intermediación.

- **Terceras empresas que hacen uso de los resultados de los análisis efectuados**. Estas empresas buscan maximizar sus beneficios tratando de emplear los resultados de los análisis sobre Big Data en su beneficio, ofreciendo productos, servicios o condiciones personalizadas a usuarios o potenciales usuarios de sus productos / servicios.

- **Autoridades Públicas de Gestión y Protección de la Privacidad**. Deben jugar un papel primordial, que en todo caso debe ser potenciado respecto al que desempeñan ahora en la mayor parte de los casos, al ser las encargadas de **supervisar que el proceso se desarrolla a través de procedimientos que protegen la información personal**.

  Organizaciones como la *Agencia Española de Protección de Datos en España*, la *Information Commissioner's Office* en Reino Unido, el *Privacy Commissioner* dependiente de la *Office of the Australian Information Commissioner* en Australia, la *Office of the Privacy Commissioner* de Canadá, el *The Data Protection Ombudsman* en Finlandia, la *Commission Nationale de l'Informatique et des Libertés (CNIL)* en Francia, las *Autoridades de Protección de Datos* de los Länder alemanes, la *Office of the Privacy Commissioner for Personal Data* en Hong Kong, la *Office of the Data Protection Commissioner* en Irlanda, la *Israeli Law, Information and Technology Authority* en Israel, el *Instituto Federal de Acceso a la Información y Protección de Datos* en México, el *College Bescherming Persoonsgegevens* en Holanda, el *Federal Service for Supervision of Communications, Information Technologies and Mass Media* en Rusia, el *Federal Data Protection and Information Commissioner* en Suiza, la *Federal Trade Commission* en EE.UU., o la Unidad Reguladora y de Control de Datos Personales de Uruguay *[30]*, demuestran como en múltiples países existen Autoridades responsables de la Gestión y Protección de la Privacidad.

  Será fundamental la interoperabilidad y colaboración entre estas Autoridades para los diversos países del mundo, en un contexto como el actual de servicios altamente globalizados, con el fin de alcanzar una protección homogénea y evitar la aparición de las ya comentadas '*sombras legales*' en relación con la protección.

  No obstante cabe destacar como un reto futuro para la consecución de la necesaria homogeneidad en la Protección, que países con una Industria TI pujante que están adquiriendo una importancia decisiva en la explotación de Big Data como India, o

China, no disponen de Autoridades de Gestión y Protección de la Privacidad *[11]*, *[17]*, *[46]*.

Se debe potenciar el cumplimiento normativo que obliga a que, se ubiquen donde se ubiquen los datos personales y se analicen donde se analicen, sea siempre de aplicación la Ley del país del que sea natural la persona a la que pertenezcan. Las Autoridades de Gestión de los Big Data deben tener capacidad (normativa y tecnológica) para supervisar y certificar que esto se cumple.

Al mismo tiempo, estas Autoridades también deben ser responsables de las **tareas de concienciación, sensibilización y formación**, fundamentalmente de usuarios en relación con la protección de la privacidad, aspectos críticos tal y como se expondrá más adelante.

3.2.2. Introducción a la Solución

Teniendo en cuenta las características que debe cumplir la Plataforma Integral de Intermediación para la Protección de la Privacidad en entornos Big Data (holística, generalista, búsqueda en cada caso del punto más óptimo de equilibrio entre Privacidad y Utilidad de Big Data, cobertura del ciclo de vida completo), así como todos los actores involucrados en el proceso que se acaban de presentar, esta solución debe **abordarse teniendo en cuenta las necesidades y roles de dichos actores y asegurando que la interacción entre ellos se desarrolla** siempre teniendo en cuenta la necesidad de aseguramiento de la Privacidad.

La Plataforma Integral de Intermediación para la Protección de la Privacidad en entornos Big Data, si bien como su propio nombre indica se centra en la Protección de la Privacidad, busca desde su propia definición y diseño, un **enfoque Win-Win** de modo que este **incremento en la protección de la Privacidad** de las personas **no sólo no reduzca la competitividad y las oportunidades de negocio** de las empresas que ofrecen servicios y/o productos a esas personas, **sino que pueda incluso llegar a incrementarla** en determinadas condiciones. A su vez este incremento en las oportunidades de negocio por parte de las empresas al poder aprovechar la potencialidad de los Big Data redundaría en beneficios para los usuarios a los que se les ofrecerían nuevas formas de afrontar sus retos y necesidades de formas más fáciles y accesibles (y con el plus añadido de que su privacidad se encuentra salvaguardada en un porcentaje muy elevado). Se articularía de este modo una especie de círculo virtuoso de realimentación de beneficios que en todo caso constituye el objetivo y la referencia a alcanzar por parte de la solución que en este capítulo se describe.

De este modo se comprueba como el **reto fundamental** es conseguir que las empresas proveedoras de servicios que captan datos de sus usuarios, implementen mecanismos de protección de la privacidad que en principio pueden ser vistos como una manera de **renunciar voluntariamente al uso libre y sin restricciones de los datos de sus usuarios que son su verdadero motor y en muchos casos fuente de ingresos multimillonarios**. ¿Cómo se les debería hacer ver de que este enfoque puede en última instancia ser beneficioso para ellas?

En principio la respuesta inmediata apunta al **cumplimiento normativo**. La aplicación de unas prácticas no alineadas con la normativa en vigor daría lugar a sanciones (que en virtud de las últimas modificaciones legales pueden llegar a ser muy importantes). Pero en un contexto de normativa no completamente homogénea, con la posibilidad para las empresas para escudarse en el principio de la Notificación y Consentimiento que pueda haber otorgado un usuario, este cumplimiento normativo no debe ser nunca la única respuesta a esgrimir, y el éxito en la aplicación de soluciones protectoras de la privacidad debe sustentarse sobre el enfoque Win-Win antes referido.

Las empresas proveedoras de servicios son conscientes de que los asuntos relativos a la protección de la Privacidad son cada vez más importantes para sus usuarios. Así por ejemplo, y como ya se ha puesto de manifiesto en el *Capítulo I*, según un estudio de la Universidad de Viena *[93]*, uno de los factores más importantes por el que los usuarios dejan de emplear Facebook, es la percepción por su parte de que esta red social no protege en la medida necesaria su privacidad.

La privacidad también se puede destacar como uno de los factores que explican en mayor medida el éxito que ha tenido la aplicación Snapchat, que permite a los usuarios publicar fotos y videos de hasta 10 segundos de duración que se pueden compartir con amigos de tu red o públicamente. Los contenidos compartidos permanecen en la aplicación durante 24 horas, desapareciendo posteriormente, y si alguien hace una captura de pantalla de algo que tú has publicado, el sistema te envía una notificación *[1]*. Si bien es cierto que ha existido cierta polémica con Snapchat, dado que se han presentado casos en los que algunos contenidos no se han borrado pasado el periodo establecido *[80]*, lo cierto es que han sido sus funcionalidades orientadas a la protección de la privacidad lo que ha representado en buena medida su éxito, que ha sido tal que hoy en día la empresa tiene un valor estimado próximo a los 20.000 millones de dólares *[60]*.

Por lo tanto las redes sociales tradicionales, sin renunciar completamente a su modelo de negocio actual (la venta de datos personales a terceras empresas), sí pueden buscar mecanismos que al mismo tiempo hagan ver a sus usuarios que velan por su privacidad y así incrementar el grado de compromiso hacia ellos.

Evidentemente y como ha puesto de manifiesto por ejemplo el robo de información que sufrió Snapchat en diciembre de 2013 *[76]* y *[23]*, estas plataformas no están a salvo de los ataques de usuarios ilícitos. Por este motivo la Plataforma Integral de Intermediación para la Protección de la Privacidad en entornos Big Data se debe complementar con los mecanismos de control de acceso y de protección frente a usuarios ilícitos externos (o internos) que buscan el robo de información o la alteración en el normal desarrollo de análisis y operaciones, que se han presentado en el *Capítulo II* y con la evoluciones que sobre la base de éstas se están produciendo.

A las **terceras empresas implicadas** (aquellas con las que inicialmente el usuario implicado no interacciona ni comparte sus datos directamente), que compran datos de las proveedoras de servicios, bien para explotarlos ellas mismas o ya procesados en función de sus intereses, se les debe presentar la visión de que un enfoque como el presente también puede redundar en beneficios para ellas.

En concreto las interacciones comerciales que se producirían en un entorno protector de la privacidad podrían ser más efectivas, ya que sólo recibiría datos de personas que en una proporción más alta estarán más interesadas en lo que dichas terceras empresas les pueden ofrecer. Así por ejemplo que a un determinado usuario reciba anuncios de portales tipo Meetic, únicamente porque Facebook ha identificado que no tiene una relación sentimental (es decir haciendo un uso de un dato privado de manera oculta para el usuario en cuestión), no es del todo efectivo, pero sí lo sería en el caso en que el mismo usuario hubiera aceptado que el dato relativo a su situación sentimental fuera compartido con terceros (desligándolo previamente de nombre, contactos, etc.). En este caso para la tercera empresa, el acceder a este dato sería mucho más efectivo en términos comerciales.

En todo caso, el **responsable principal** de que los servicios que se ofrecen sean conformes con la protección de la privacidad consensuada en cada caso con los usuarios finales y dentro del marco que supone la normativa de aplicación, deben ser las empresas proveedoras de servicios (siempre que los servicios finalmente se proporcionen a través suyo, como los

anuncios y las sugerencias de terceras empresas que se publican en el muro de Facebook o la cuenta de Gmail de los usuarios).

Todo el enfoque de Privacidad balanceada con Utilidad de los Big Data, supone la necesidad de que los **usuarios** tengan una mucho mayor sensibilización y concienciación en relación con aspectos de su privacidad, en un punto intermedio entre **posiciones alarmistas** que consideran que cualquier información compartida en Internet va a suponer una amenaza para su privacidad que a su vez se va a traducir en un riesgo contra sus intereses de algún tipo, y **posiciones completamente despreocupadas**, derivadas de un desconocimiento de los peligros que ciertas prácticas pueden tener.

Estas Políticas / Campañas de Concienciación, Sensibilización y Formación, que serán un instrumento más en el enfoque de la Plataforma Integral de Intermediación, serán una responsabilidad fundamental de las **Autoridades de Gestión de la Privacidad de los distintos países y regiones**, con el apoyo de Organizaciones y Empresas proveedoras de servicios a usuarios.

3.3. Diagrama de Implementación por bloques de la Plataforma Integral de Intermediación

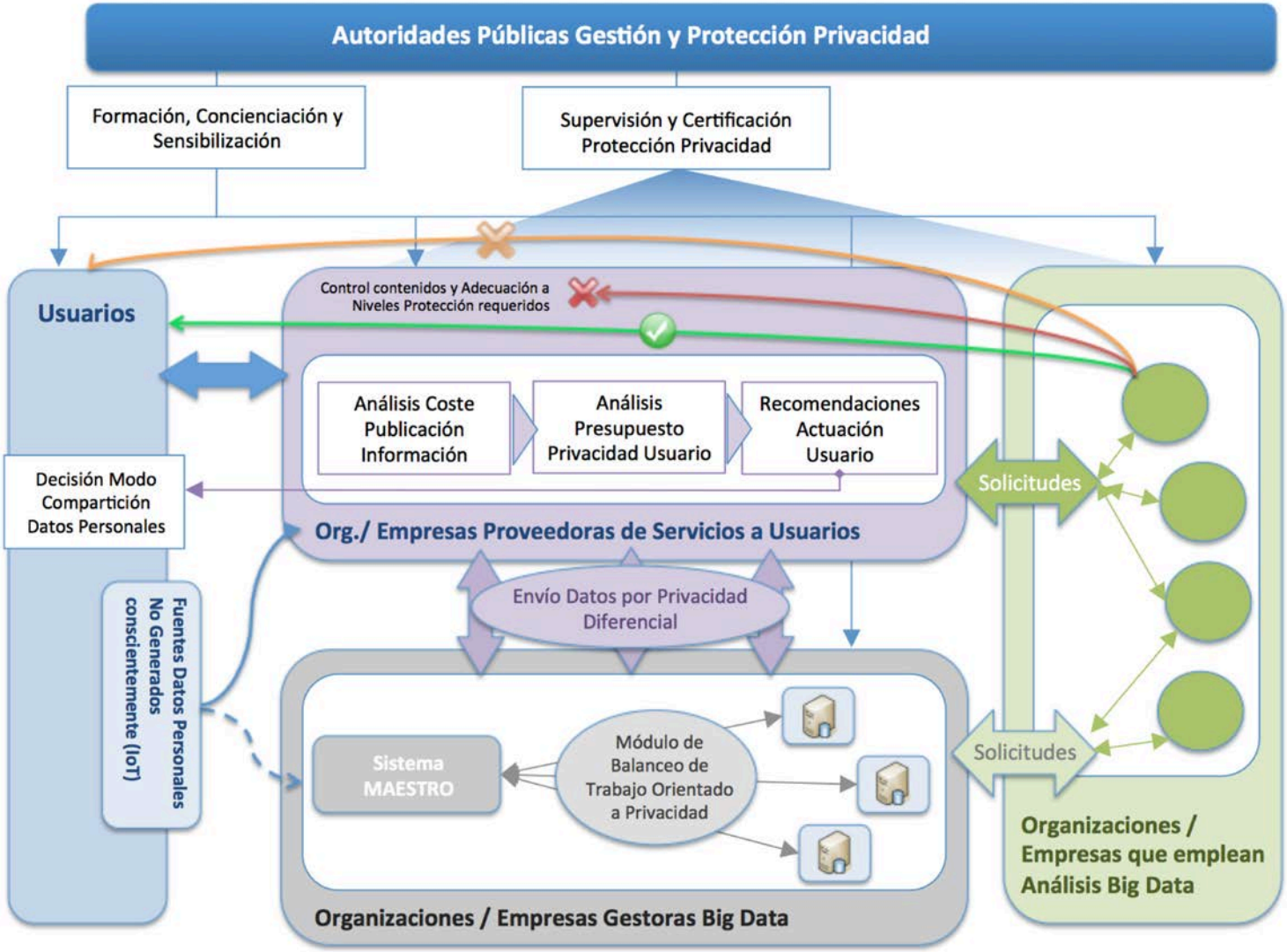


Imagen IV-6 – Diagrama de Implementación por bloques de la solución

3.3.1. Bloque correspondiente a la Interacción Usuario-Empresa Proveedora de Servicios

El siguiente esquema representa el modelo de implementación de las interacciones de los usuarios y las empresas y organizaciones prestadoras de servicios en el modelo de la Plataforma de Intermediación para la Protección de la Privacidad en entornos Big Data.

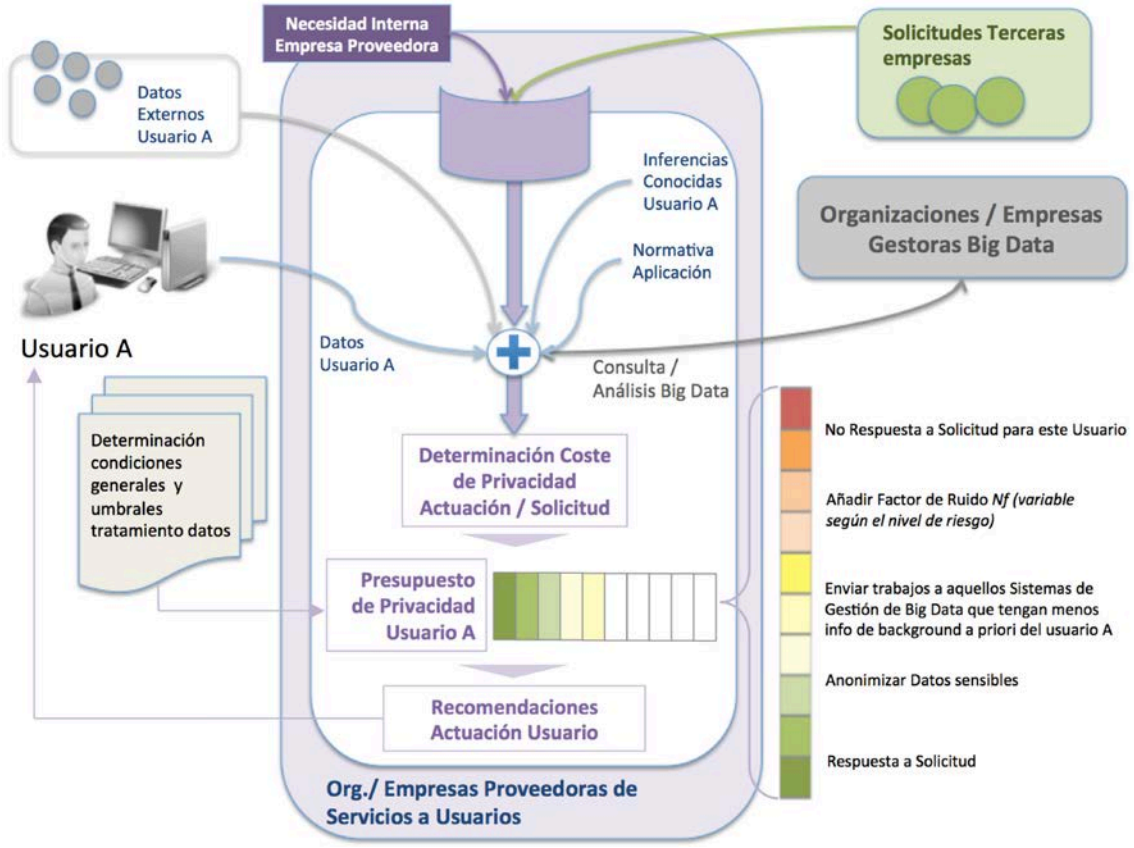


Imagen IV-7 – Esquema Interacción Usuario – Organización/Empresa Proveedora de Servicios en la Plataforma de Intermediación para la Protección de la Privacidad en Big Data

Tal y como se destaca como una directriz en varias disposiciones normativas de Protección de la Privacidad y en especial de la nueva *Regulación General de Protección de Datos de la Unión Europea*, **se debe reforzar la posición de los usuarios, dotándoles de mayor capacidad para poder decidir lo que se puede hacer y lo que no con sus datos**, desde una postura informada. Para ello deben poder conocer en cada caso (o en cada categoría de casos) las ventajas y riesgos que supone cada acción para que puedan ponderar y decidir en consecuencia.

Muchas veces el no informar adecuadamente de las ventajas puede hacer que los usuarios sean celosos en exceso de su privacidad y pueda llegar incluso a perder oportunidades útiles para ellos, al no poder incluirse en análisis de Big Data que en determinados casos pueden suponer unas condiciones especiales en la prestación de un servicio, una oferta especial por un determinado producto, etc.

No obstante, tal y como se ha destacado en el *Capítulo III*, es imposible que los usuarios, con los medios a su disposición, puedan conocer en todos los casos lo que supone para su privacidad el hecho de ceder una información personal referente a ellos, dadas las casi infinitas

posibilidades de inferencia de información que representan el análisis de los Big Data que pueden desarrollar las empresas y organizaciones gestoras de datos y proveedoras de servicios. Por este motivo **una verdadera decisión informada por parte de los usuarios debe componerse de varios pasos, e involucrar a las empresas que captan los datos y a las Autoridades Públicas de Gestión**.

En concreto, y teniendo en cuenta que todo el proceso cuenta como entradas fundamentales con la información publicada o solicitada a los usuarios finales y con la solicitud de información requerida por terceras empresas a las empresas que recopilan datos, se trataría de los siguientes pasos:

1. El usuario determinará unas **condiciones generales y unos umbrales por defecto para la protección de la privacidad**. Estos umbrales generales que se determinarán dentro de cada servicio online que se emplee, deben poder ser modificados de manera sencilla y dinámica por su parte.

   Inicialmente, por ejemplo cuando se da de alta una cuenta en un servicio determinado, se establecerían unos parámetros básicos, que después, con cada caso de uso concreto, se irían actualizando y/o concretando. Se deberían regular aspectos como qué tipo de información es más sensible para el usuario y qué tipo de empresas / organizaciones puede tener acceso a qué tipo de datos.

   Se trata por tanto de una especie de **Do-not-track selectivo y dinámico**, no filtrando completamente para un sitio web completo, sino que se tiene en cuenta los datos concretos en un esquema de análisis caso por caso. De esta forma se permite mejorar el equilibrio Privacidad-Utilidad de los Big Data.

   Por ejemplo, un usuario puede no tener problema con que Google pueda usar la información de sus correos de Gmail para ofrecerle servicios personalizados, pero para aquéllos que tengan que ver con su situación médica (y que intercambia con su médico personal) no deben ser susceptibles de poder ser empleados. Además, esta restricción general que inicialmente puede haber hecho para todos sus datos médicos, posteriormente puede concretarse más y no aplicarla para asuntos relativos a lesiones deportivas (dado que le puede interesar recibir información personalizada de fisioterapeutas de su zona).

   Se trataría de una evolución de las condiciones de uso y privacidad que actualmente permiten configurar empresas como Google o Facebook para sus servicios, pero añadiendo alternativas adicionales. Así por ejemplo en Facebook actualmente es posible configurar quién puede ver tus publicaciones (sólo amigos, amigos y amigos de éstos, público en general), quién puede enviarte solicitudes de amistad, si se quiere que los motores de búsqueda muestren el enlace a tu biografía, etc. En lo relativo a anuncios de terceras empresas, en la configuración de Privacidad de Facebook simplemente se indica que el nombre y la foto de los usuarios no se podrá usar por terceras empresas en anuncios (lo que en principio no significa que no puedan usar la información de tu biografía para componer que anuncios personalizados).

2. Estas condiciones serán comunicadas conforme se vayan generando a la empresa proveedora de servicios, que junto con el resto de la información que obra en su poder (o que puede inferir) y la normativa de aplicación en cada caso, le permitirá **tomar decisiones sobre el Coste que para la Privacidad puede suponer** cada una de las solicitudes que recibe de terceras empresas.

3. En función de este Coste, la empresa Proveedora actualizará el **Prepuesto de Privacidad de un Usuario concreto**, esto es la cantidad de información personal que es posible

gestionar de un determinado usuario sin que ello suponga un riesgo inasumible para su privacidad, a tenor de sus preferencias personales y de la normativa de aplicación.

Según el nivel en el que se encuentre el Presupuesto de Privacidad (y de si el umbral límite para el usuario en cuestión se ha alcanzado o no), se decidirá la manera en que se responderá o no a la solicitud de las terceras empresas.

Así por ejemplo, y tal y como se muestra en la *Imagen IV-7*, se podrá decidir responder a la solicitud sin modificaciones en datos, anonimizar los datos más sensibles, enviar trabajos de análisis a empresas / nodos gestores de Big Data que a priori puedan tener menos información de background del usuario en cuestión, modificar los datos con un determinado factor de ruido, o en los casos más extremos de riesgo no atender a la solicitud.

4. Como se puede comprobar el resultado de este proceso decisor tiene repercusiones sobre el modo en que se efectuarán los análisis sobre los Big Data para poder inferir resultados, y por lo tanto, como se expondrá al analizar la siguiente interacción en el *subapartado 3.3.2.*, servirá para actualizar la manera en que se realizará la distribución de trabajos en los sistemas gestores de Big Data y las políticas que regirán este proceso.
5. Además, también permitirá, cuando sea necesario, informar al Usuario de una serie de **recomendaciones** para que actualice sus condiciones generales de protección de la privacidad.

3.3.2. Bloque correspondiente a la Interacción Empresa Proveedora de Servicios – Empresa Gestora de Big Data

En un diseño de protección de la Privacidad centrado en el usuario como el presente (tanto desde el punto de vista de proteger sus datos personales, como el de ofrecerle en cada momento los mayores beneficios que los Big Data pueden proporcionar), el Proveedor de Servicios es el actor principal a la hora de proveerle de los medios para ejercer la Protección de la Privacidad.

La gestión de la privacidad en las organizaciones / empresas gestoras de Big Data se basará en la definición de una funcionalidad que se integre en el marco de referencia de gestión de Big Data que se emplee (como podría ser Hadoop). Para ello se podría establecer una aplicación independiente dentro del ecosistema establecido o como una parte de una aplicación existente encargada de tareas de gestión de datos, como por ejemplo Ambari o Chukwa en el caso de Hadoop.

Este módulo se encargaría, en función de las directrices que haya establecido el Proveedor de Datos al Sistema Maestro en cada caso y de manera dinámica, de ir adaptando la política de asignación de datos y tareas a los Nodos Esclavos, de manera que la distribución resultante sea lo más óptima posible en términos de protección de la privacidad, al mismo tiempo que se garantiza la máxima utilidad.

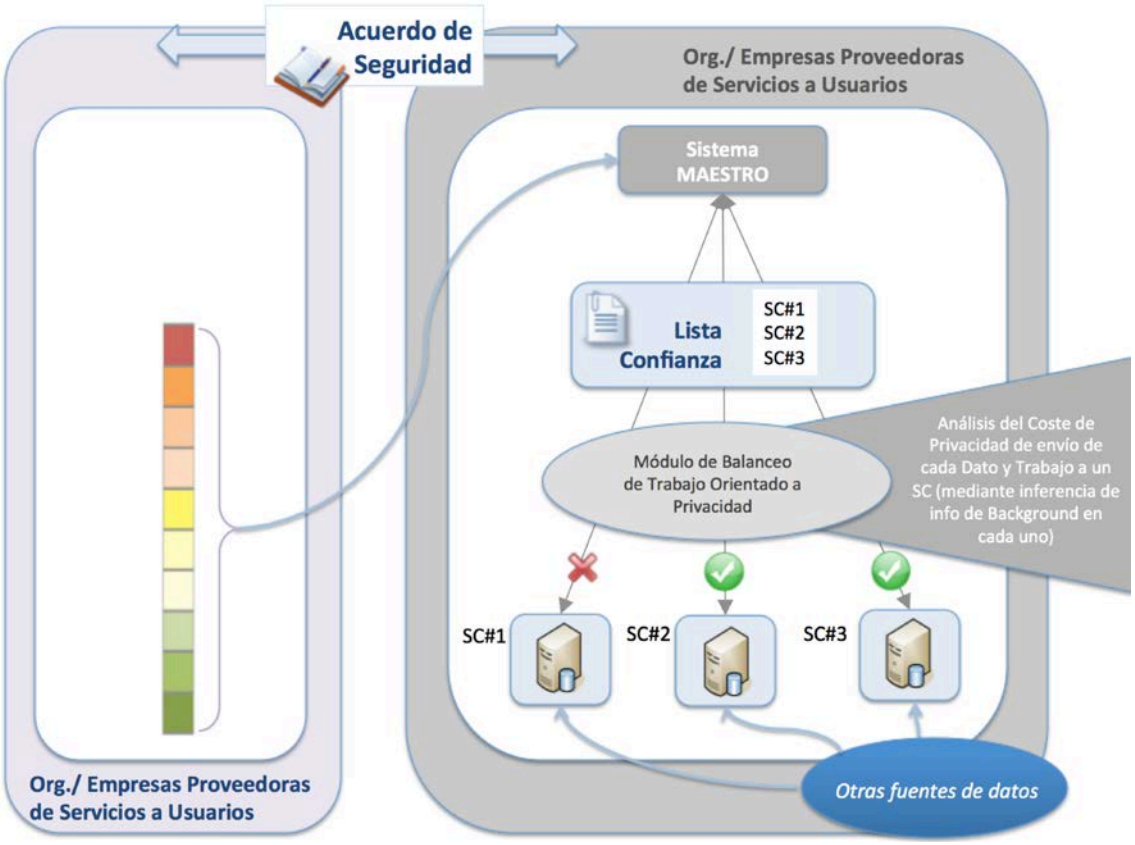


Imagen IV-8 – Esquema Interacción Organización / Empresa Proveedora de Servicios – Organización / Empresa Gestora de Big Data en la Plataforma de Intermediación para la Protección de la Privacidad en entornos Big Data

3.3.3. Bloque correspondiente a la Interacción Empresa Proveedora de Servicios – Empresa que emplea los Análisis de los Big Data para ofrecer Servicios / Productos

Por defecto en el esquema de relación de la Plataforma de Intermediación para la Protección de la Privacidad en Entornos Big Data, la interacción entre la empresa principal proveedora de servicios que capta los datos de sus usuarios y aquellas terceras empresas que requieren estos datos para poder efectuar personalizar su cartera de productos y servicios (y de este modo ser más eficientes), se basará en un **enfoque determinado por la Privacidad Diferencial**.

Tal y como se ha expuesto en el *Capítulo II*, la Privacidad Diferencial hace que la empresa u organización solicitante no reciba conjuntos de datos en bruto, sino simplemente respuestas a preguntas y consultas concretas, que además, en función del riesgo existente (el Coste de la Operación y el Presupuesto de Privacidad restante para un caso concreto), podrá ser modificado mediante la suma de un determinado factor de ruido.

De esta forma se permite acotar mejor el propósito de los análisis (sin reducir completamente el grado de descubrimiento de nuevas posibilidades conforme a los análisis se producen, que es uno de los beneficios de los Big Data, como ya se ha destacado).

Una vez que las terceras empresas proveedoras de servicios han recibido los resultados de los análisis de los Big Data y han podido conformar su oferta de servicios personalizados, éstos

deberán ser siempre controlados por la empresa que originalmente captó los datos, que al fin y al cabo es el responsable principal de la protección de la privacidad, y respondería frente a las correspondientes autoridades de gestión por un hipotético incumplimiento de la normativa de aplicación. Así por ejemplo, si un usuario ha hecho opt-out a que su información de localización se pueda usar, se debe controlar que los servicios que se ofrecen cumplen con esta premisa.

Para posibilitar este control, se debe asegurar que las terceras empresas siempre empleen la plataforma de la empresa recopiladora de datos para ofrecer sus servicios.

3.3.4. Bloque correspondiente a las Autoridades Públicas de Gestión y Protección de la Privacidad

Todo el proceso que supone la Plataforma de Intermediación para la Protección de la Privacidad debe contar como factor consolidador y transversal, que asegure su éxito, con la Legislación y con la actividad de supervisión y control de las Administraciones Públicas a través de los Organismos que para la Protección de la Privacidad se hayan constituido (como los ya mencionados en el *apartado 3.2.1*.).

La colaboración y el trabajo estrecho entre estas Autoridades Públicas y las grandes empresas de Internet que manejan datos (Google, Facebook, Twitter,…) es fundamental, no sólo para el funcionamiento de una solución holística como la aquí presentada, sino para cualquier medida que busque garantizar la privacidad de los usuarios.

La labor de las Autoridades Públicas de Gestión y Protección de la Privacidad se centra en dos tareas fundamentales:

- **Supervisión y Certificación de la Protección de la Privacidad.**

  En este caso y como uno de los mecanismos que podrían implementarse se encontrarían las siguientes:

  - **Auditorías de Control Periódicas del Cumplimiento de la Ley.**
  - Análisis de las **Evaluaciones de Impacto en la Protección de Datos (DPIAs)** que harían las empresas proveedoras de servicios, en función de las decisiones tomadas teniendo en cuenta costes y presupuestos de privacidad, según lo que ya se ha expuesto.
  - Empleo de **usuarios de control (perfiles de supervisión creados ad-hoc)**, con los que se emplearían los servicios de diversas empresas siguiendo patrones de uso tipo, para poder comprobar se primera mano incumplimientos de la normativa o de las buenas prácticas de protección de la Privacidad.

- **Formación, Concienciación y Sensibilización de los diversos actores involucrados.**

4. Conclusiones

El reto fundamental en el ámbito de la protección de la privacidad en contextos de gestión de Big Data consiste en equilibrar al mismo tiempo Privacidad y Utilidad de los datos. Los métodos expuestos en los capítulos anteriores de este proyecto pueden hacer que los usuarios reduzcan de una manera considerable sus datos, dado que pueden entender que su privacidad no se encuentra adecuadamente protegida. Este hecho podría llevar a una reducción de las posibilidades y beneficios que aportan los Big Data.

Por este motivo la solución propuesta con la **Plataforma Integral de Intermediación para la Protección de la Privacidad**, trata de superar estos enfoques, dando al usuario no tanto la

posibilidad de decidir que sus datos no puedan usarse, sino que puedan hacerlo de un modo razonado e informado que le proporcione las mayores ventajas.

Además para aquellos casos en que no sea posible para el usuario conocer con exactitud las implicaciones que una cesión de su información personal puede suponer, deben ser las empresas proveedoras de servicios las que garanticen que la privacidad de dichos usuarios se salvaguarda en función de sus preferencias personales (unas mínimas preferencias personales siempre van a tener que indicarse) y de la normativa de aplicación.

Todo este proceso se supervisaría por parte de las Autoridades responsables de la Protección de la Privacidad en los distintos países. Pero la normativa y el cumplimiento no debe ser la única herramienta, y gracias al empleo de las diferentes soluciones tecnológicas disponibles (anonimización, cifrado, control sobre las consultas, etc.), las empresas que recopilan y explotan datos pueden incrementar las garantías de protección de la privacidad que ofrecen a sus usuarios (mejorando por lo tanto el grado de confianza que éstos tendrán hacia ellas), al mismo tiempo que mantienen e incluso incrementan las posibilidades de obtener resultados útiles y eficientes de los Big Data que ofrecer a terceras empresas o directamente a los usuarios.

CAPÍTULO V

CONCLUSIONES Y LÍNEAS DE INVESTIGACIÓN FUTURAS

1. Conclusión

A lo largo del mes de agosto de 2015, al iniciar la navegación en Google se presentaba la siguiente pantalla, indicando que para continuar la navegación y poder hacer cualquier búsqueda, se debían aceptar previamente los términos expuestos:

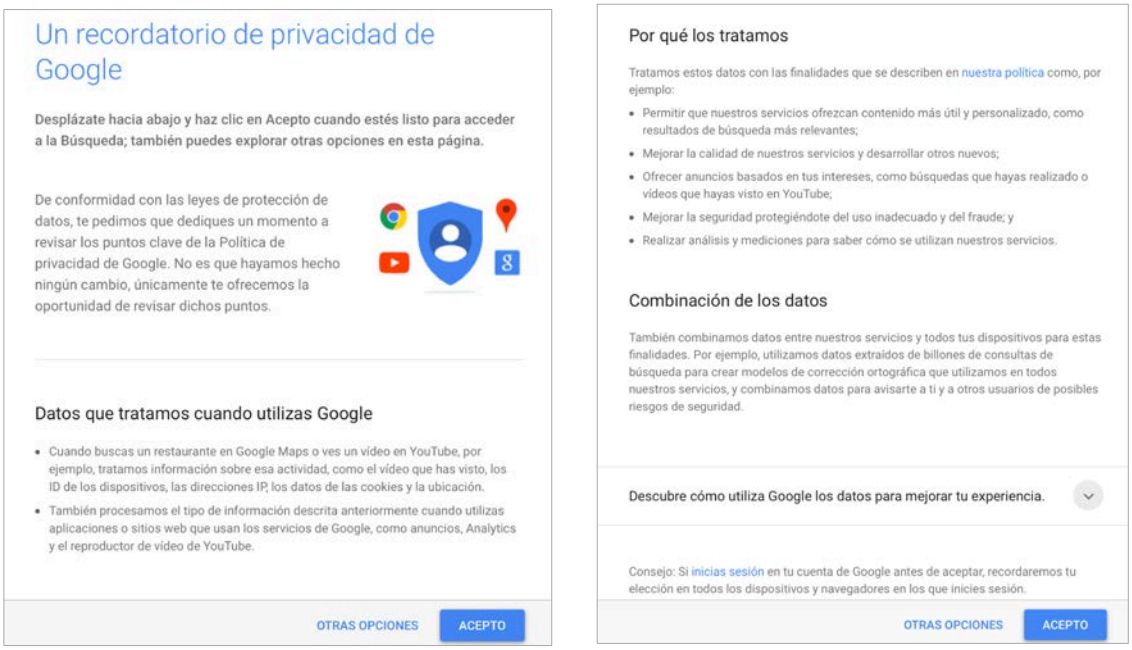


Imagen V-1 – Términos de Privacidad de Google
(fuente: *www.google.es*) y *https://privacy.google.com)*

Si bien a partir de esta información inicial muy escueta y resumida, es posible seguir investigando en mayor profundidad los términos y condiciones que aplica Google en el empleo de los datos personales de sus usuarios, la misma pone de manifiesto algunos de los aspectos que ya han sido puestos de manifiesto en este proyecto, como el hecho de que sea imprescindible aceptar los términos de manera general, completa y para todos los casos si se quiere seguir usando el servicio, sin dar la posibilidad a un análisis y discriminación de casos más granular por parte del usuario, en función de sus preferencias personales, sus datos concretos implicados y el tipo de empresa / servicio con el que se compartirán los datos en cuestión.

Todo ello demuestra la **necesidad de plantear nuevas prácticas y soluciones que permitan a las empresas mantener sus modelos de negocio, al mismo tiempo que se sitúa a los usuarios en una situación más favorable para decidir según los casos concretos, que nivel de protección se debe aplicar a sus datos**.

El hecho de que las empresas proveedoras de servicios presenten información como la anterior, demuestra como cada vez se encuentran **más concienciadas de la importancia de cumplir con la legislación en materia de privacidad**, ante las repercusiones (multas, pérdida de reputación e imagen de marca, etc.).

Sin embargo, estas mismas empresas deben ser también conscientes de que sus usuarios, con carácter general, cada vez van a estar **más informados, y van a ser más conocedores de sus derechos y de las posibilidades que las diferentes organizaciones y empresas con las que comparten sus datos tienen para inferir nueva información** que luego puedan

emplear para optimizar el desarrollo de su actividad. Todo ello debe obligarlas a adoptar nuevas, diferentes y más eficientes medidas en relación con la protección de la privacidad.

Además, siguiendo la idea fuerza que se ha destacado a lo largo de este proyecto, la premisa fundamental que se debe tener en consideración es la **búsqueda de un equilibrio razonable y adaptado a cada situación concreta entre la protección de la privacidad y la garantía de la utilidad de los Big Data**, para que el hecho de proteger los datos personales de los usuarios no suponga al mismo tiempo una reducción considerable de las ventajas y posibilidades que ofrece el análisis de los Big Data.

Los diferentes aspectos que en los capítulos anteriores se han presentado, permiten concluir que cualquier solución eficiente que se desarrolle de cara a la protección de la privacidad en el ámbito de la gestión de los Big Data **debe tener en cuenta no sólo las tecnologías disponibles y sus potenciales evoluciones, sino también los aspectos normativos y procedimentales implicados, así como el papel que deben jugar todos los actores** que participan en el proceso.

De este modo un enfoque como el representado por la Plataforma Integral de Intermediación para la Protección de la Privacidad en entornos Big Data, puede considerarse un punto de partida adecuado, a partir del cual ir añadiendo los avances que las líneas de investigación futura vayan aportando.

2. Líneas de Investigación Futuras

Aparte de los ya apuntados en los diferentes capítulos anteriores, es imprescindible que una solución del tipo de la Plataforma Integral de Intermediación para la Protección de la Privacidad tenga un diseño abierto y adaptativo que permita en cada momento adoptar los diferentes avances que en cada uno de los diferentes ámbitos de actividad (tecnológico y normativo/procedimental) y para cada uno de los distintos bloques que la componen, supongan beneficios y una mejora en el desempeño de su objetivo final.

Se presentan a continuación, y para cada uno de los ámbitos de actividad implicados, algunas de las **líneas de Investigación abiertas, que en el corto/medio plazo se podrían incorporar a soluciones holísticas generalistas de protección de la privacidad**, como la presentada en el *Capítulo IV*.

2.1. Ámbito Tecnológico

- Incremento de la **eficiencia en el procesamiento de Datos Cifrados**, a partir de la evolución de enfoques como la Criptografía Homomórfica.
- Avances en los **Métodos de Anonimización Grupales** (como la k-Anonimización, la l-Diversidad o la t-Proximidad) para su aplicación no sólo en conjuntos de datos estructurados, sino también en **datos no estructurados como textos o información derivada de Redes Sociales y del Internet of Things**.
- Definición, desarrollo e implementación de técnicas y procedimientos para la protección de la privacidad en relación con **información contenida en formatos de imagen y/o video.** En este sentido, puede resultar un caso de especial relevancia la **protección de la privacidad** (manteniendo al mismo tiempo la imprescindible utilidad) **asociada a variables biométricas**, cuyo empleo será cada vez más extendido, por ejemplo para implementar mecanismos de autenticación y protección de la seguridad de la información.

- Búsqueda de nuevos enfoques como la **anonimización** de **datos por parte de los usuarios finales que son sus titulares**, antes de que puedan compartirse y combinarse con otros datos *[104]*, así como plantear la **posibilidad de anonimización de metadatos asociados a una transferencia de datos concreta** (datos de posicionamiento, la dirección IP, tiempo de conexión, etc.), a partir de las posibilidades que los diferentes formatos para estos datos ofrecen (Exchangeable Image File Format (EIFF), International Press Telecommunications Council (IPTC), Adobe's Extensible Metadata Platform (EMP), etc.).
- Mejora de la **traza de datos en los análisis de Big Data y aplicación de técnicas de control de la dispersión de datos**, para poder posibilitar el borrado en casos en los que se requiera la aplicación de por ejemplo el Derecho al Olvido. Se trata de un aspecto en principio muy complicado, por lo que una solución en paralelo debe buscar que se tenga que recurrir las menos veces posibles a este tipo de soluciones drásticas. Para ello el usuario debe ser en cada momento consciente de qué se hace con sus datos, teniendo un cierto grado de control sobre ellos y sentir que efectivamente el diseño de soluciones de análisis de sus datos (y combinación con otros que pueda haber generado otros usuarios o él/ella de manera inconsciente) ha tenido desde el primer momento en un lugar central la protección de su privacidad.

2.2. Ámbito Normativo / Procedimental

- Evolución de las **disposiciones normativas** en materia de protección de datos personales sean **más flexibles y al mismo tiempo eficientes sin centrarse en exceso en conceptos como la Notificación y Consentimiento**, que como se ha presentado, no posibilita todas las ventajas que se requieren (puede suponer una reducción de la utilidad de los Big Data y al mismo tiempo tampoco garantizan una adecuada protección de la privacidad).
- **Refuerzo** de la posición y las competencias de las **Autoridades Públicas de Gestión y Protección de la Privacidad**, para poder desarrollar de una manera más ágil y efectiva su labor de seguimiento y supervisión del proceso de gestión de los datos personales.
- **Incremento** de la **Homogeneidad e Interoperabilidad entre normativas y procedimientos**, especialmente crítico en entornos altamente globalizados, como aquéllos en los que actualmente se desarrolla el intercambio de datos/información entre empresas y organizaciones (y cada vez en mayor medida).
- **Extensión** de la implementación de **enfoques proactivos y de privacy-by-design**.

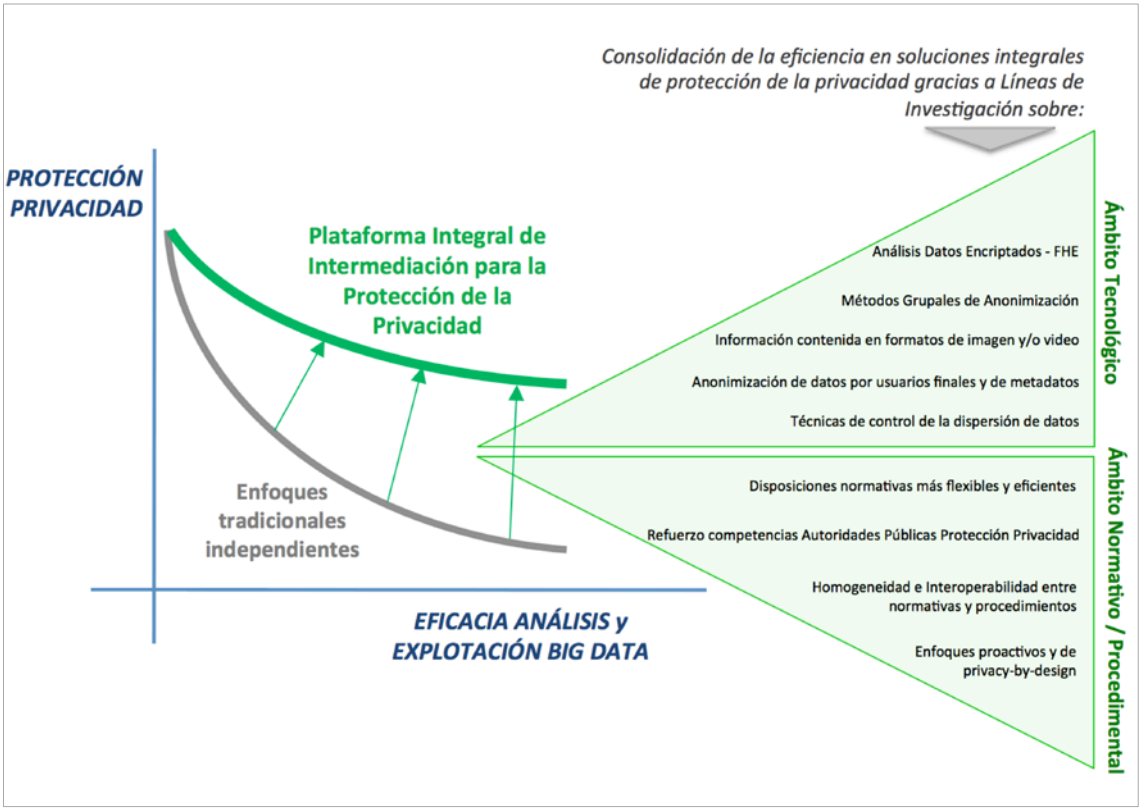


Imagen V-2 – Líneas de Investigación Futuras

ANEXO I - BIBLIOGRAFÍA

[1] K. Acuna, Business Insider http://www.businessinsider.com/james-gunn-explains-why-snapchat-is-successful-2015-4. Abril 2015.

[2] Agencia Española de Protección de Datos. http://www.agpd.es/portalwebAGPD/LaAgencia/informacion_institucional/conoce/index-ides-idphp.php

[3] Agencia Española de Protección de Datos. http://www.agpd.es/portalwebAGPD/internacional/Proteccion_datos_mundo/index-ides-idphp.php

[4] C. C. Aggarwal, IBM T.J. Watson Research Center, Hawthorne NY, and P. S. Yu - University of Illinois at Chicago. ***A General Survey of Privacy-Preserving Data Mining Models and Algorithms***.

[5] K.Anbazhagan, Dr. R.Sugumar, M.Mahendran, R.Natarajan. ***An Efficient Approach for Statistical Anonymization Techniques for Privacy Preserving Data Mining***, - International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 7, September 2012.

[6] Apache Software Foundation. https://hadoop.apache.org

[7] Apache Software Foundation. https://knox.apache.org

[8] Apache Software Foundation. www.spark-project.org

[9] Axelos. ***Proceso de Gestión del Conocimiento*** - ITIL v3 para Gestión de Servicios TI.

[10] E. Bertino and M. Kantarcioglu, Workshop Chair: B. Thuraisingham. ***Big Data – Security with Privacy Response to RFI for National Privacy Research Strategy***. From The Participants of the NSF Big Data Security and Privacy Workshop (September 16-17, 2014) Track Chairs: DRAFT October 16, 2014.

[11] Big Data Congress. http://bigdatacongress.net/brick-by-b-r-i-c-how-big-data-is-strengthening-emerging-markets/

[12] S. Blazhievsky - Nice Systems. Storage Networking Industry Association (SNIA). ***Introduction to Hadoop, MapReduce and HDFS for Big Data Applications***.

[13] F.H. Cate - Center for Applied Cybersecurity Research, Indiana University. ***The Failure of Fair information Practice Principles. Forthcoming in Consumer Protection in the Age of the 'Information Economy'***.

[14] F.H. Cate, Maurer School of Law, Indiana University, and V. Mayer-Schönberger, Oxford Internet Institute, University of Oxford. ***Notice and Consent in a World of Big Data*** - Microsoft Global Privacy Summit Summary Report and Outcomes.

[15] A. Cavoukian, Information & Privacy Commissioner Ontario – Canada, and J. Jonas IBM Fellow Chief Scientist, IBM Entity Analytics. ***Privacy by Design in the Age of Big Data***. Junio 2012. (http://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf).

[16] A. Cavoukian, Information & Privacy Commissioner Ontario – Canada. https://www.privacybydesign.ca/content/uploads/2009/08/7foundationalprinciples-spanish.pdf

[17] China Daily. http://guizhou.chinadaily.com.cn/2015-06/16/content_21016316.htm

[18] Cisco Webpage. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

[19] Cloud Security Alliance. ***Top Ten Big Data Security and Privacy Challenges***. 2012

[20] Comisión Europea. ***European Commission Communication, 'A comprehensive approach on personal data protection in the European Union' COM (2010) 609 final.***

[21] Comisión Europea. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[22] Comisión Europea. http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/1998/wp12_es.pdf

[23] M. Contreras. FayerWayer. https://www.fayerwayer.com/2014/01/publicados-4-6-millones-de-usuarios-y-numeros-de-snapchat/

[24] C. Cumby (editor). Active Project European Union.***State-of-the-art Review of Privacy-preserving Data-mining***. Marzo 2010

[25] D. Das, O. O'Malley, S. Radia and K. Zhang, Hortonworks and IBM. ***Adding Security to Apache Hadoop***

[26] J. Dean and S. Ghemawat, Google Inc. ***MapReduce: Simplified Data Processing on Large Clusters***

[27] Y. Demchenko, C. Ngo, P. Membrey. ***Architecture Framework and Components for the Big Data Ecosystem***. Draft Version 0.2. System and Network Engineering Group. Amsterdam University. Septiembre 2013

[28] M. van Dijk and A. Juels, RSA Laboratories ***On the Impossibility of Cryptography Alone for Privacy-Preserving Cloud Computing***.

[29] Disconnect webpage. https://disconnect.me

[30] DLA Piper Global Data Protection and Privacy Team. ***Data Protection Laws of the World*** – 2015.

[31] Do Not Track Website. http://donottrack.us

[32] Domo webpage. https://www.domo.com/learn/data-never-sleeps-2

[33] Duck Duck Go webpage. https://duckduckgo.com

[34] E. Dumbill. https://beta.oreilly.com/ideas/what-is-big-data. Enero 2012

[35] C. Dwork. ***Differential Privacy***. International Colloquium on Automata Languages and Programming – ICALP 2006.

[36] C. Dwork, F. McSherry K. Nissim and A. Smith. ***Calibrating noise to sensitivity in private data analysis***. Theory of Cryptography Conference-TCC 2006.

[37] C. Dwork. Microsoft Research and D. K. Mulligan, School of Information, Berkley Law. ***It's Not Privacy and It's Not Fair***, Stanford Law Review. Septiembre 2013.

[38] Electronic Privacy Information Center***. Big Data and the Future of Privacy.*** Comments of the Electronic Privacy Information Center to the office of Science and Technology Policy. Abril 2014.

[39] Europe versus Facebook webpage. http://europe-v-facebook.org/FAQ_ENG.pdf

[40] J. Gantz and D. Reinsel, IDC IVIEW. ***Extracting Value from Chaos***. Junio 2011.

http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[41] Gartner, Inc. ***Gartner IT Glossary - Big Data definition,*** http://www.gartner.com/it-glossary/big-data/

[42] C. Gentry, ***A Fully Homomorphic Encryption Scheme***, https://crypto.stanford.edu/craig/craig-thesis.pdf. Septiembre 2009

[43] Google Flu Trends Project. https://www.google.org/flutrends/es/#ES

[44] Google, Self-Driving Car Project. http://www.google.com/selfdrivingcar/how/

[45] A. Gosain and N. Chugh – USICT Guru Gobind Singh Indraprastha University, Delhi India. ***Privacy Preservation in Big Data***. International Journal of Computer Applications (0975-8887) Volume 100 – Nº17. Agosto 2014.

[46] Greater Pacific Capital. http://greaterpacificcapital.com/article/big-data-analytics-at-the-tip-of-an-iceberg/

[47] M. Gualteri http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data. Diciembre 2012.

[48] M. Hay, G. Miklau, D. Jensen, P. Weis, S. Srivastava. Computer Science Department – University of Massachusetts – Amherst. ***Anonymizing Social Networks***. 2007.

[49] V. C. Hu, T. Grance, D.F. Ferraiolo, D. Rick Kuhn, National Institute of Standards and Technology, Gaithersburg, MD, USA. ***An Access Control Scheme for Big Data Processing***.

[50] IBM Big Data and Analytics Hub: http://www.ibmbigdatahub.com/gallery/quick-facts-and-stats-big-data

[51] Ixquick Webpage. https://ixquick.com/esp/

[52] M. Jensen, Independent Centre for Privacy Protection Schleswig-Holstein (ULD) Kiel, Germany. ***Challenges of Privacy Protection in Big Data Analytics***.

[53] M. Jiménez, Cinco Días. http://cincodias.com/cincodias/2014/06/04/tecnologia/1401910197_449355.html. Junio 2014.

[54] R. Jones, R. Kumar, B. Pang, A. Tomkins. ***"I Know What You Did Last Summer": Query Logs and User Privacy***. In Proceedings of CIKM'07, 2007

[55] B.J. Koops - Tilburg Institute for Law, Technology, and Society, Tilburg University, the Netherlands. ***The Trouble with European Data Protection Law***. August 2014.

[56] F. Lardinois. http://techcrunch.com/2015/04/03/microsoft-disables-do-not-track-as-the-default-setting-in-internet-explorer/. Abril 2015.

[57] D. Lazer. MIT Technology Review. http://www.technologyreview.com/view/526416/mistaken-analysis/. Abril 2014.

[58] D. Lazer, R. Kennedy, G. King, A. Vespignani. ***The Parable of Google Flu: Traps in Big Data Analysis***. Marzo 2014

[59] N. Li[1], T. Li[1], S. Venkatasubramanian[2]. [1]: Department of Computer Sciences, Purdue University. [2]: AT&T Labs Research ***t-Closeness: privacy beyond k-anonymity and l-diversity***. International Conference on Data Engineering-ICDE 2007. IEEE.

[60] I. Lunden, Alexia Tsotsis. Tech Crunch. http://techcrunch.com/2014/12/31/snapchat-485m/

[61] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam. Department of Computer Science, Cornell University ***l-Diversity: privacy beyond k-Anonymity***. ACM Transactions on Knowledge Discovery from Data-TKDD. 2007.

[62] A. Mantelero. Polytechnic University of Turin. ***Rethinking E.U. data protection in the Big Data World.*** 6th International Conference on Information Law and Ethics (Thessaloniki, Greece, 30-31 mayo, 2014).

[63] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung Byers***. Big Data: The next frontier for innovation, competition and productivity***. McKinsey Global Institute. Mayo 2011: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[64] B. Marr, LinkedIn. ***The Awesome Ways Big Data Is Used Today To Change Our World***. Noviembre 2013

[65] C. McTaggart – Object Oriented Analysis and Design Spring 2011. Computer Sciences Department - University of Colorado Boulder. ***Hadoop / MapReduce***.

[66] Microsoft News Center. ***The Big Bang: How the Big Data Explosion Is Changing the World. Febrero 2013*** - http://news.microsoft.com/2013/02/11/the-big-bang-how-the-big-data-explosion-is-changing-the-world/

[67] K. Naganuma, M. Yoshino, H. Sato, Y. Sato. Service Innovation Research Department, Yokohama Research Laboratory, Hitachi Ltd. Hitachi ***Privacy-preserving Analysis Technique for Secure, Cloud-based Big Data Analytics***. Review Vol. 63 (2014), Nº 9

[68] V. Narasimha Inukollu, S. Arsi, and S. Rao Ravuri, Department of Computer Engineering, Texas Tech University, USA and Department of Banking and Financial Services, Cognizant Technology Solutions, India. ***Security Issues associated with Big Data in Cloud Computing***, - International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3. Mayo 2014

[69] A. Narayanan, E. W. Felten. Princeton University. ***No silver bullet: De-identification still doesn't work.*** Julio 2014.

[70] A. Narayanan and V. Shmatikov, The University of Texas at Austin. ***De-anonymizing Social Networks****.* 2009.

[71] A. Narayanan and V. Shmatikov. The University of Texas at Austin. ***Robust De-anonymization of Large Sparse Datasets***. Mayo 2008.

[72] H. Nissenbaum, Stanford University. ***Privacy in Context. Technology, Policy, and the Integrity of Social Life***. 2009.

[73] Organización para la Cooperación y el Desarrollo Económicos (OCDE). ***Guidelines on the Protection of Privacy and Transborder Flows of Personal Data by the Committee of Ministers of the Organization for Economic Cooperation and Development.*** OECD 1980.

[74] Oracle webpage. http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf

[75] M. A. Pérez. http://blogthinkbig.com/senales-de-trafico-inteligentes/. Marzo 2015.

[76] N. Perlroth and J. Wortham. Bits – New York Times. http://bits.blogs.nytimes.com/2014/01/02/snapchat-breach-exposes-weak-security/?_r=0

[77] S.G. Reddy, Nuviun. ***Proyecto Artemis***: http://nuviun.com/content/big-data-saves-small-babies-by-detecting-nosocomial-infections-earlier-than-clinicians. Noviembre 2014.

[78] Resolución de Madrid - Estándares Internacionales sobre Protección de Datos Personales y Privacidad. http://privacyconference2011.org/htmls/adoptedResolutions/2009_Madrid/2009_M1.2.pdf

[79] M. Rezaei Jam - Department of Computer Engineering University of Tabriz Tabriz, Iran, L.M. Khanli - Department of Computer Engineering University of Tabriz Tabriz, Iran, M. K. Akbari - Department of Computer Engineering and IT Amirkabir University of Technology Tehran, Iran, M. Sargolzaei Javan Department of Computer Engineering and IT Amirkabir University of Technology, Tehran, IRAN. ***A Survey on Security of Hadoop***, 4th International Conference on Computer and Knowledge Engineering (ICCKE).

[80] P.F. Roberts. IT World. http://www.itworld.com/article/2705499/security/snapchat-is-less-private-than-you-think.html

[81] I.S. Rubinstein, New York University School of Law. International Data Privacy Law Advance Access. ***Big Data: The End of Privacy or a New Beginning?*** Enero 2013.

[82] H. Russo – Geek's Room. http://geeksroom.com/2015/01/que-sucede-en-un-minuto-en-internet-comparando-el-ano-2013-contra-el-2014/91383/. Enero 2015.

[83] P. Samarati – Computer Sciences Laboratory, SRI International, and L. Sweeney – Laboratory for Computer Science, Massachusetts Institute of Technology. ***Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression***. 1998.

[84] B. Santhosh Kumar, Anna University of Technology, Tiruchirappalli, and K.V. Rukmani, CSI College of Engineering, Ketti ***Novel Privacy Notion T-Closeness: Privacy Preserving Data Mining***.

[85] Securosis. ***Understanding and Selecting Data Masking Solutions: Creating Secure and Useful Data***. Agosto 2012.

[86] J. Sen. Department of Computer Science, National Institute of Science & Technology Odisha, India ***Homomorphic Encryption: Theory & Application***.

[87] E. Serbeto. ABC. http://www.abc.es/economia/20140408/abci-sentencia-datos-operadoras-201404081038.html

[88] S. Sicular, Gartner, Inc. ***Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s,*** 27 March 2013. http://www.forbes.com/sites/gartnergroup/2013/03/27/ gartners-big-data-definition-consists-of-three- parts-not-to-be-confused-with-three-vs/

[89] A. Smith. Pew Research Center. http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook

[90] A. Solana. La Vanguardia. http://www.lavanguardia.com/internet/20111025/54236037686/un-estudiante-de-derecho-denuncia-a-facebook-por-conservar-datos-borrados.html

[91] J. Soria-Comas and J. Domingo-Ferrer, Universidad Rovira i Virgili. ***Connecting privacy models: synergies between k-anonymity, t-closeness and differential privacy***. Joint UNECE / Eurostat work session on statistical data confidentiality (Ottawa, Canada, 28-30. Octubre 2013).

[92] Spokeo webpage http://www.spokeo.com

[93] S. Stieger, C. Burger, M. Bohn, and M. Voracek, Cyberpsychology, Behavior, and Social Networking.http://www.adweek.com/socialtimes/study-why-people-quit-facebook/428402 - Who Commits Virtual Identity Suicide? Differences in Privacy Concerns, Internet Addiction, and Personality Between Facebook Users and Quitters, September 2013, Vol. 16, No. 9: 629-634
http://medienportal.univie.ac.at/uniview/forschung/detailansicht/artikel/quitting-facebook-whats-behind-the-new-trend-to-leave-social-networks/

[94] N. Stokes. Techlicious. http://www.techlicious.com/tip/the-best-browser-privacy-tools/

[95] L. Sweeney - School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. ***k-Anonymity: A Model for Protecting Privacy***. 2002

[96] M. Terrovitis, N. Mamoulis, P. Kalnis. ***Privacy-Preserving Anonymization of Set-Valued Data***. In Proc. of VLDB Endowment, 2008.

[97] Tribunal Supremo de España.
http://www.poderjudicial.es/search/doAction?action=contentpdf&databasematch=TS&reference=7195354&links=%226153/2011%22&optimize=20141023&publicinterface=true

[98] UK Information Commisioner's Office. ***Proposed new EU General Data Protection Regulation: Article-by-article analysis paper*** — February 2013.

[99] UK Information Commisioner's Office. ***Big Data and Data Protection – Data Protection Act 1998.*** – Julio 2014.

[100] Unión Europea. http://eur-lex.europa.eu/legal-content/ES/TXT/?uri=uriserv:l24120

[101] Universia España. http://noticias.universia.es/cultura/noticia/2015/05/15/1125062/dia-internet-pasa-red-1-minuto.html. Mayo 2015

[102] D.K. Vallabhadas. Department of Computer Science and Engineering National Institute of Technology Rourkela, India. ***Comparative Study for Distance Metrics for t-closeness***.

[103] G. Vassall-Adams, M. Chambers. Eutopia Law.
http://eutopialaw.com/2014/05/16/case-comment-google-spain-sl-google-inc-v-agencia-espanola-de-proteccion-de-datos-mario-costeja-gonzalez/

[104] S.J. Vaughan-Nichols, IT World, julio 2013. http://www.itworld.com/article/2829511/big-data/big-data--metadata--and-traffic-analysis--what-the-nsa-is-really-doing.html

[105] Veena and Devidas, NMAM Institute of Technology, Nitte, Karnataka, India. ***Data Anonymization Approaches for Data Sets Using Map Reduce on Cloud: A Survey***.

[106] J. Wakefield - BBC News. http://www.bbc.com/news/technology-23253949. Agosto 2013.

[107] D. Wu. Stanford University Security Lunch, ***Somewhat (Practical) Homomorphic Encryption.*** Febrero 2014

[108] S. Yakoubov, V. Gadepally, N. Schear, E. Shen, A. Yerukhimovich. MIT Lincoln Laboratory, Lexington. EE.UU.***A Survey of Cryptographic Approaches to Securing Big-Data Analytics in the Cloud***.

[109] M. Zaharia et al. Amplab. UC Berkeley. ***Spark. Fast, Interactive, Language-Integrated Cluster Computing***.

[110] H. Zang and J. Bolot. ***Anonymization of location data does not work: A large-scale***

measurement study. Int. Proc. 17th International Conference on Mobile Computing and Networking 145-156 (2011)

[111] B. Zhou, J. Pei and W. Shun Luk, School of Computing Science, Simon Fraser University, Canada. ***A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data***. 2008.

ANEXO II – PRESUPUESTO DEL PROYECTO FIN DE CARRERA

El presupuesto para la realización del presente Proyecto Fin de Carrera tiene como principal componente, dado su carácter eminentemente teórico, el coste de las horas de trabajo que se han empleado.

Para proceder a su cuantificación se han tenido en cuenta el coste asignado por la Universidad Carlos III a un mes de trabajo (131,25 horas) de un ingeniero, que asciende a 2.694,39 € / mes de trabajo. Dado que se han invertido 125 horas en la realización de PFC, el coste para este caso ascendería a 2.566,09 €.

Además se han tomado en consideración el coste complementario de otros elementos, como la consulta de bases de datos y respositorios de información como la del Institute of Electrical and Electronics Engineers (IEEE) (www.ieee.org) o la de la asociación ISACA (relativa a Sistemas y Tecnologías de la Información) (www.isaca.org), o el uso de elementos de infraestructura física como el ordenador personal o la conexión a Internet empleados. El coste que suponen estos elementos se muestra en la tabla siguiente a título ilustrativo, pero no se cuantifican en el coste final, dado que su uso no ha sido exclusivo para la elaboración del presente PFC.

En la siguiente tabla se desglosa el coste del PFC:

Elementos involucrados en la elaboración del PFC			Coste
Trabajo para elaboración del proyecto			
	Duración proyecto: 125 horas	*Coste ingeniero UC3M: 2.694,39 € / mes trabajo (1 mes trabajo = 131,25 horas)*	**2.566,09 €**
Acceso a BBDD y Repositorios de Información			
	Base Datos IEEE	*Provisto por la UC3M*	**-**
	Base Datos ISACA	*135 US$ →118 € (cambio 15/10/2015)*	**-**
Elementos Infraestructura Física empleados			
	Ordenador Personal (amortizable a 6 años)	*Uso durante 6 meses: 108 €*	**-**
	Conexión a Internet	*Uso durante 6 meses: 150 €*	**-**
		TOTAL	**2.566,09 €**

Tabla A-I – Desglose Coste PFC

PÁGINA INTENCIONADAMENTE EN BLANCO