



Oportunidades Industria 4.0 en Galicia

Convenio de
colaboración entre el
Instituto Gallego de
Promoción
Económica, la Alianza
Tecnológica
Intersectorial de
Galicia y los centros
integrantes de esta
alianza para la
detección y análisis de
oportunidades
sectoriales para las
empresas industriales
gallegas en el ámbito
de la industria 4.0



ÍNDICE

1. INTRODUCCIÓN	4
1.1 BIG DATA	5
1.1.1 Definición/Descripción	5
1.1.2 Breve historia.....	7
1.1.3 Ventajas y limitaciones.....	10
1.1.4 Tendencias.....	11
1.1.5 Principales tecnologías en la industria.....	11
1.1.6 Herramientas Big Data	12
1.1.7 Big Data: casos de uso	26
1.2 DATA ANALYTICS	29
1.2.1 Breve historia.....	29
1.2.2 Definición/Descripción	30
1.2.3 Análisis avanzado de datos	32
1.2.4 Ventajas y limitaciones.....	46
1.2.5 Tendencias.....	46
1.2.6 El binomio Big Data- Data Analytics	49
1.3 CLOUD COMPUTING	55
1.3.1 Breve Historia	55
1.3.2 Definición y características	55
1.3.3 Ventajas y limitaciones.....	58
1.3.4 Tendencias y casos de uso	59
2. APLICACIONES POR SECTOR	63
2.1 AGROALIMENTACIÓN Y BIO	63
2.1.1 Proyectos de I+D	64
2.2 AUTOMOCIÓN	65
2.2.1 Proyectos I+D.....	69
2.3 MADERA Y FORESTAL	69
2.4 NAVAL	70
2.5 TEXTIL/MODA	72
2.5.1 Proyectos de I+D	73
2.6 AERONÁUTICA	73
2.6.1 Proyectos de I+D	74
2.7 TICS	74

2.7.1	Proyectos I+D.....	75
2.8	ENERGÍAS RENOVABLES.....	75
2.9	PIEDRA NATURAL.....	78
2.10	METALMECANICO.....	79
2.10.1	Proyectos I+D.....	80
3.	CONCLUSIONES / IMPACTO EN LA INDUSTRIA	81
3.1	RETOS.....	81
3.2	PERSPECTIVAS A MEDIO Y LARGO PLAZO.....	85
3.3	CONCLUSIONES.....	89
3.3.1	Políticas De Apoyo	89
3.3.2	Impacto Económico.....	90
3.3.3	Impacto Industrial.....	92
4.	BIBLIOGRAFÍA.....	98

1. INTRODUCCIÓN

La capacidad actual para producir información (**datos**) ha crecido exponencialmente respecto a años anteriores. La enorme cantidad de datos a disposición hace necesario el desarrollo de herramientas que permitan el procesado y análisis de los mismos, para identificar y extraer la información relevante.

En relación a la industria, Big Data está jugando y seguirá jugando un papel principal en la llamada **Cuarta Revolución Industrial** [1]. La primera revolución industrial que tuvo lugar a finales del S.XVIII hasta el inicio del S.XX, consistió en la mecanización con agua y energía de vapor. La segunda revolución (inicio del S.XX hasta la década de los años 70) introdujo la producción en masa gracias al uso de la energía eléctrica, y desde los años 70 hasta la actualidad, la tercera revolución industrial se caracterizó por el uso de la electrónica y de la informática para automatizar más la producción. La cuarta revolución industrial, denominada “Industria 4.0” se fundamenta en el uso de los grandes volúmenes de datos obtenidos a través de los sistemas ciberfísicos (CPS) y el Internet de las cosas (IoT), para conseguir una industria de factorías inteligentes en las que las máquinas y los distintos recursos se comunican como en una red social.

Para que las factorías y productos sean verdaderamente **inteligentes**, grandes cantidades de datos deben ser recogidos, analizados e interpretados en tiempo real [2].

El principal objetivo del uso de Big Data en la industria es conseguir procesos industriales libres de defectos, ágiles, flexibles y a un coste eficiente. McKinsey¹ sugiere que la Cuarta Revolución Industrial conseguirá **descensos del 50% del coste de fabricación** y ensamblaje de productos a través del uso de Big Data. Los datos enviados por los dispositivos inteligentes pueden ayudar a la industria a identificar con exactitud las preferencias de los consumidores y por lo tanto diseñar nuevos productos ajustados a las necesidades de los clientes.

La eficiente captura y análisis de los datos redundará en la **competitividad de las empresas**, desde la fabricación, cadena de suministro, hasta los sistemas de gestión, procesos que pueden ser mejorados a través del uso de Big Data.

Para dar sentido al término Big Data y que la industria disponga de mecanismos para la toma de mejores decisiones, conocer mejor su negocio, generar posibles oportunidades de negocio y verificar o refutar teorías y modelos existentes, es necesario incorporar dos términos más, necesarios, para el desarrollo de Big Data:

1. **Big Data Analytics:** *Ciencia de examinar datos en bruto con el propósito de sacar conclusiones sobre esa información.* Implica aplicar un proceso algorítmico o mecánico para obtener conocimiento. Por ejemplo, aplicar un proceso para buscar correlaciones significativas entre varias series de datos.
2. **Cloud Computing:** *Modelo para permitir el acceso conveniente y bajo demanda a un conjunto compartido de recursos computacionales configurables (por ejemplo, redes, servidores, almacenamiento, aplicaciones y servicios), que se pueden aprovisionar y liberar rápidamente con un esfuerzo mínimo de gestión o una interacción entre el proveedor de servicios.*

¹ <http://www.mckinsey.com/>

1.1 BIG DATA

1.1.1 Definición/Descripción

Al igual que todos los nuevos términos que surgen de los grandes avances tecnológicos, no existe un consenso claro sobre **cómo definir el término Big Data**. Muchas definiciones están centradas exclusivamente en el volumen de datos y otras tienen en consideración factores específicos como el tiempo o incluso la industria. Sin embargo, existen algunas definiciones que han capturado la esencia de lo que la mayoría entiende por dicho término.

Desde la presentación del término por el **MGI (McKinsey Global Institute)** en junio de 2011 [6] han existido diversos intentos de acotación del concepto. MGI define Big Data como:

“Los conjuntos de datos cuyo tamaño está más allá de la habilidad de las herramientas software de base de datos para capturar, almacenar, gestionar y analizar.”

La definición de Big Data dada por McKinsey tiene en consideración la variable temporal para construir la definición Big Data. No se establece un límite mínimo del tamaño del conjunto de datos que constituirá Big Data, es decir, el término variará en el tiempo y algunos conjuntos de datos que actualmente son considerados Big Data no lo serán en el futuro, es decir, el término cambiará con el tiempo de acuerdo al avance de la tecnología.

La empresa multinacional de auditoría Deloitte define el término como: *“El término que se aplica a conjuntos de datos cuyo volumen supera la capacidad de las herramientas informáticas de uso común, para capturar, gestionar y procesar datos en un lapso de tiempo razonable”*.

Del mismo modo otros autores [7,8] proporcionan diferentes definiciones de Big Data que hacen referencia principalmente al **tamaño de los conjuntos de datos**. Así, Kord Davis y Doug Patterson [9] afirman:

“Big Data son datos demasiado grandes para ser manejados y analizados por protocolos de bases de datos tradicionales como SQL”.

Sin embargo, en la literatura encontramos autores que hacen referencia a otros aspectos de Big Data. Es el caso de Edd Dumbill [10] que da la siguiente definición:

“Big Data son datos demasiado grandes, se mueven demasiado rápido o no encajan las restricciones de sus arquitecturas de bases de datos”.

Otra definición que hace referencia a la **multidimensionalidad de Big Data** es la dada por la consultora Gartner:

“Big Data son activos de información caracterizados por su alto volumen, velocidad y variedad que demandan formas innovadoras y rentables de procesamiento de información para mejorar la compresión y la toma de decisiones”.

CARACTERÍSTICAS

El volumen del conjunto de datos no es la principal característica que nos permite definir Big Data. Varios autores han caracterizado Big Data mediante lo que se denomina como **las tres V's: Volumen, Variedad y Velocidad**. Sin embargo, cada vez más son los autores que incorporan más

V's al conjunto de características de Big Data para completar dicha definición. Por ejemplo, IBM introdujo Veracidad y, actualmente se consideran, también Valor y Visualización como dos características fundamentales de Big Data.

Volumen: Gran cantidad de datos generados. En el año 2013 se generaron en el mundo en torno a 4.4 zettabytes y se estiman que en el año 2020 se alcancen los 44 zettabytes (ZB). Las organizaciones se enfrentan a volúmenes masivos de datos. Hay una tendencia en las principales compañías de almacenar vastas cantidades de datos de múltiples tipos (redes sociales, datos de salud, datos financieros, datos de sensores, etc.), por lo que surge la necesidad de procesar toda esta cantidad de información, siendo posible gracias a las técnicas de análisis de Big Data.

Variedad: Erróneamente se asocia Big Data con fuentes de datos no estructurados. Sin embargo, las fuentes de datos pueden ser de cualquier tipo:

- **Datos estructurados:** La mayoría de las fuentes de datos tradicionales son datos estructurados. Son datos que disponen de un esquema o formato fijos. Son los datos provenientes de bases de datos relacionales, hojas de cálculo o archivos con un formato fijo.
- **Datos semiestructurados:** Son datos que no presentan un formato fijo pero contienen etiquetas o marcadores que permiten formatear o separar los elementos. Ejemplos típicos de datos semiestructurados son los registros Web, logs de las conexiones a Internet, o cualquier otro texto etiquetado de XML o HTML.
- **Datos no estructurados:** Son datos sin tipos predefinidos sobre los que se tiene poco o ningún control. Se almacenan en objetos o documentos sin ninguna estructura. Datos de texto, fotografía, vídeo o audio, son datos no estructurados. Ejemplos típicos son: imágenes digitales, mensajes de texto con formato libre, como son el caso de SMS, correos electrónicos, mensajes de WhatsApp, tweets u otros mensajes de redes sociales, etc. El continuo crecimiento de este tipo de datos sin formato con un análisis complejo ha dado lugar a la aparición de tecnologías tales como MapReduce, bases de datos NoSQL, o herramientas como Apache Hadoop para su correcto y eficiente procesamiento.

Velocidad: Rapidez con la que se generan y se mueven los datos. Existe un aumento creciente de los flujos de datos en las organizaciones, de la frecuencia de la creación de registros, de las actualizaciones en las grandes bases de datos y de la disponibilidad en el acceso y entrega de datos. Este incremento del flujo de información requiere de un almacenamiento, procesamiento y análisis adecuados, especialmente cuando lo que se necesita es una gestión en tiempo real.

Valor: Capacidad de extraer valor, es decir, información o conocimiento de los datos, que en definitiva es el fin último de la tecnología Big Data. Así, la International Data Corporation define las arquitecturas Big Data como:

“diseñadas para económicamente extraer valor a partir de grandes volúmenes de datos permitiendo la captura, el descubrimiento y análisis rápido”.

Veracidad: Es lo que se ajusta a la verdad o al hecho, es decir, datos precisos y certeros. Establecer la fiabilidad de los datos, o lo que es lo mismo, métodos para tratar la incertidumbre de los mismos es uno de los retos de Big Data. Dicha incertidumbre puede ser causada por

inconsistencias, aproximaciones, ambigüedades, latencia, o incluso duplicación. IBM afirma que “Uno de cada tres líderes de negocios (directivos) no confía en la información que utiliza para tomar decisiones”.

Visualización: Importancia de proporcionar buenas herramientas que permitan comprender y analizar los resultados obtenidos tras los análisis Big Data.

Por tanto, en función de sus características, se podría definir Big Data como **gran volumen, alta velocidad y gran variedad de datos** que requieren de un procesado eficiente y poco costoso, para obtener conocimiento o valor que permita tomar decisiones fiables. Sin embargo, es importante resaltar que no todas las características (V's) tienen que estar presentes siempre. Existen, además, otras características importantes de Big Data que no se citan normalmente por caer fuera del grupo de las V's, como es el caso de la tolerancia a fallos.

1.1.2 Breve historia

En los últimos 20 años las mejoras en las tecnologías de adquisición y almacenamiento de datos han originado un incremento exponencial de los datos disponibles en diferentes campos. Según un informe del **International Data Corporation** (IDC), en 2011, la creación y copia de datos en el mundo de manera global fue de 1.8ZB ($\approx 1021B$), lo que significó un incremento de casi nueve veces en cinco años y se considera que esta cifra se duplicará al menos cada dos años en el futuro cercano [3].

El comienzo de la **sobrecarga de información** podría datarse en 1880 cuando el censo de los Estados Unidos tardaba 8 años en tabularse. Con el fin de acortar los tiempos de tabulado se inventó la máquina tabuladora de Holreith (tarjetas perforadas). El boom demográfico de los años 30 agravó este aumento de información. En 1940 las bibliotecas tuvieron que adaptar sus métodos de almacenamiento para responder al rápido aumento de la demanda de nuevas publicaciones e investigación. Es en esta década cuando los científicos empiezan a utilizar el término “explosión de la información”. Término que aparece por primera vez en el periódico Lawton Constitution en el año 1941.

En 1951 el concepto de **memoria virtual** es desarrollado por el físico alemán Fritz-Rudolf Güntsch, como una idea que trataba el almacenamiento finito como infinito.

En la década de los 60 se desarrollan los primeros sistemas informáticos para la automatización de los inventarios y en 1970, Edgar F. Codd, publicó un artículo en el que se explicaba la forma en la que podía accederse a la información almacenada en bases de datos de gran tamaño, sin saber cómo estaba estructurada la información, o dónde residía dentro de la base de datos. Es el comienzo de las **bases de datos relacionales**

A mediados de la década de 1970, los sistemas de **Planificación de Necesidades de Material (MRP)** se diseñaron como herramienta para las empresas de fabricación para organizar y planificar su información.

En los años 80 los avances tecnológicos permitieron a todos los sectores beneficiarse de nuevas formas de organizar, almacenar y generar datos. La expansión del sector de las comunicaciones supone, de nuevo, un enorme crecimiento de la información.

Tras el auge de los sistemas de MRP iniciales se introduce la **planificación de recursos de fabricación (MRP II)** en la década de 1980. MRP II incluía áreas tales como la gestión del área de producción y la distribución, la gestión de proyectos, las finanzas, los recursos humanos y la ingeniería. No fue hasta mucho después de adoptar esta tecnología cuando otros sectores comenzaron a tener en cuenta, y posteriormente adoptar, la tecnología ERP.

En el año 1985, Barry Devlin y Paul Murphy definieron una arquitectura para los informes y análisis de negocio en IBM [4] que se convirtió en la base del almacenamiento de datos.

A finales de la década de los 80 y principios de los 90, se popularizan **sistemas de Planificación de Recursos Empresariales (ERP)**, siendo más flexibles e integrables con otros departamentos de la empresa: producción, la distribución, la contabilidad, las finanzas, los recursos humanos, la gestión de proyectos, la gestión de stocks, el servicio y el mantenimiento, y logística.

En 1989, Howard Dresner amplió el popular término genérico "**Business Intelligence (BI)**" o Inteligencia empresarial, inicialmente acuñado por Hans Peter Luhn en el año 1958, definiéndolo como *"los conceptos y métodos que mejoran la toma de decisiones de negocio mediante el uso de sistemas de apoyo basados en datos reales"*.

En 1992, Crystal Reports creó el primer informe de base de datos sencillo con Windows. Estos informes permitían a las empresas crear un informe sencillo a partir de diversos orígenes de datos con escasa programación de código.

En la década de 1990 se produjo un **crecimiento tecnológico explosivo**, y los datos de la inteligencia empresarial comenzaron a registrarse en forma de documentos de Microsoft Excel. El almacenamiento digital empieza a ser más rentable que el papel para almacenar los datos. Empiezan a emerger, entonces, las plataformas de BI. Además, este crecimiento tecnológico trajo consigo la necesidad de rediseñar los ERP, personalizándolos y rompiendo los límites de titularidad antes impuestos. Los nuevos proveedores de ERP deben adaptarse a un nuevo negocio en colaboración con el cliente, diseñando ERP adaptados.

En 1997 el término "Big Data" se empleó por primera vez en un artículo de los investigadores de la NASA Michael Cox y David Ellsworth. Ambos afirmaron que el ritmo de crecimiento de los datos empezaba a ser un problema para los sistemas informáticos actuales: "**El problema del Big Data**".

En 1999, el término "**Internet de las cosas**" o **IoT**, por sus siglas en inglés, fue acuñado por el emprendedor británico Kevin Ashton, cofundador del Auto-ID Center del MIT, durante una presentación en la que la identificación por radiofrecuencia a lo largo de la cadena de suministro se unía con el mundo de Internet.

En 2002 los principales proveedores de sistemas ERP, como SAP, PeopleSoft, Oracle y JD Edwards, comenzaron a centrarse en el uso de servicios web para enlazar sus propios conjuntos de aplicaciones, y facilitar a los clientes la creación de aplicaciones nuevas a partir de datos de varias aplicaciones utilizando XML.

En 2005 el enfoque se centra en la **usabilidad del usuario final**. Las empresas de SaaS (del inglés, Software as a Service) entraron en escena para ofrecer una alternativa a Oracle y SAP más centrada en la usabilidad del usuario final.

En 2006 nace **Apache Hadoop** como solución de código abierto para gestionar la explosión de los datos en internet. Apache Hadoop es un framework de código abierto para almacenar y procesar los datos que “permite el procesamiento en paralelo distribuido de enormes cantidades de datos en servidores estándar del sector, económicos, que almacenan y procesan los datos, y que pueden escalar sin límite”, según la definición de Mike Olson. A partir de junio de 2008 el término Big Data empieza a utilizarse con más frecuencia en artículos tecnológicos.

En 2010 aparecen los **ERP en la nube**, son las empresas Netsuite y Lawson Software, las primeras que adoptaron estas tecnologías; ofreciendo a medianas empresas soluciones de sistemas ERP ligeros, flexibles y asequibles.

En 2011, las principales tendencias emergentes de Inteligencia empresarial fueron los **servicios en la nube, la visualización de datos, el análisis predictivo y el Big Data**.

En junio de 2012, se produce el Lanzamiento Mundial de **IPv6**. El Internet Protocol versión 6 (IPv6), versión más reciente del protocolo de Internet, proporciona un sistema de identificación y localización para dispositivos dentro de una red y su enrutamiento a través de Internet. El IPv6 fue desarrollado por la Internet Engineering Task Force (IETF) para resolver el problema del agotamiento de direcciones IPv4.

Noviembre 2014 se convierte en el año del **Internet de las cosas (IoT)**. El IoT se ha convertido en un potente habilitador para la transformación de negocios. Su enorme impacto afectará en los próximos años a todos los sectores y todas las áreas de la sociedad. Existen enormes redes de objetos físicos dedicados (cosas) que incorporan tecnología para detectar o interactuar con su estado interno o medio externo.

En 2015 se populariza el término de **smart city**. Una ciudad inteligente (smart city) hace uso del análisis de información contextual en tiempo real para mejorar la calidad y el rendimiento de los servicios urbanos, reducir costes, optimizar recursos e interactuar de forma activa con los ciudadanos.

Probablemente, la informática de Big Data sea la mayor innovación informática de la última década. A día de hoy, tan solo hemos visto el potencial que tiene para recopilar, organizar y procesar los datos en todos los aspectos de nuestras vidas [5].

2016 What happens in an INTERNET MINUTE?

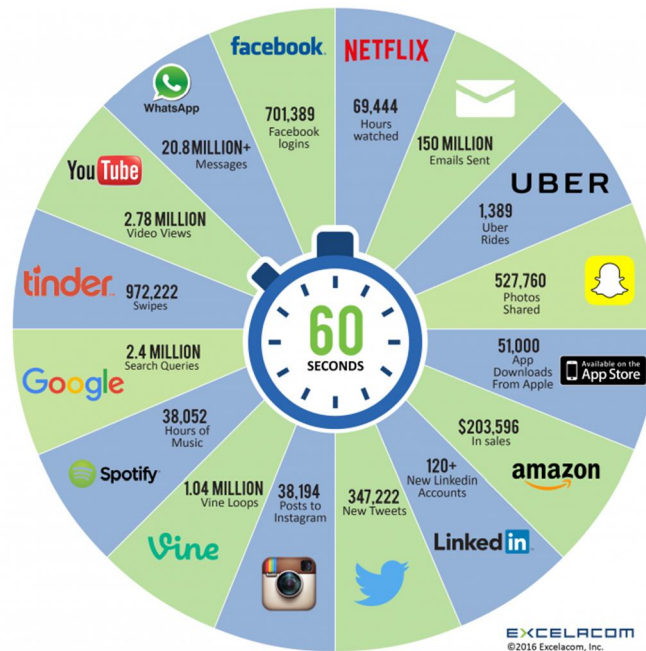


ILUSTRACIÓN 1. EL RÁPIDO CRECIMIENTO DE LOS DATOS.” WHAT HAPPENS IN AN INTERNET MINUTE? HOW TO CAPITALIZE ON THE BIG DATA EXPLOSION”. EXCELACOM 2015

1.1.3 Ventajas y limitaciones

La principal ventaja de Big Data es la **obtención de información y conocimiento**. Esta ventaja puede ser aplicada a todos los sectores y entornos, de forma que en el entorno industrial y sin ser exhaustivos, podríamos traducir esta gran **ventaja** en las siguientes:

- Búsqueda de **nuevas oportunidades** de negocio.
- Mayor **precisión** en la segmentación de mercados.
- **Optimización** de los canales de ventas.
- **Anticipación** a problemas. Sistemas predictivos.
- **Optimización** de los procesos en base a **datos históricos**.
- Soporte a la toma de decisiones en base a **algoritmos automáticos**.
- Mejoras en la **seguridad** del negocio.
- **Ahorro de costes**.
- **Ahorro de tiempos**.

Entre los principales **inconvenientes** de Big Data, podemos destacar los siguientes:

- Necesidad de elaborar una **estrategia en la organización** para la implantación de Big Data.
- **Fiabilidad** de los Datos.
- **Seguridad** de los Datos.
- **Formación** específica para el tratamiento de los datos.
- **Portabilidad** de los datos.
- **Escasez de recursos técnicos formados** para el manejo de las plataformas y análisis de

los datos, y que además dispongan de experiencia funcional en el negocio.

- **Barreras organizacionales**, como la necesidad de elaborar una estrategia de la organización para la implementación de Big Data, el seguimiento del impacto empresarial de las actividades de análisis de los datos, diseñar una estructura apropiada para dar soporte a los datos y a la analítica de datos, entre otras.
- **Costes de implantación.**

1.1.4 Tendencias

Dada la magnitud y complejidad de los desarrollos, métodos, técnicas, etc. que actualmente se incluyen bajo la denominación Big Data, dividiremos las tendencias en **Tecnologías Big Data, Herramientas Big Data y Bases de datos a Gran Escala**, de forma que podamos extraer en cada una de ellas las tendencias más significativas:

1.1.5 Principales tecnologías en la industria

A medida que el gran mercado de análisis de datos se expande, las grandes compañías (Google o Facebook) comienzan rápidamente a adoptar las principales tecnologías existentes en Big Data. El informe TechRadar: Big Data, Q1 2016, de Forrester Research [11], evalúa la madurez y la trayectoria de las diferentes tecnologías a lo largo de todo el ciclo de vida de los datos. De dicho estudio se extrae que las empresas consideran clave las siguientes tecnologías relacionadas con Big Data:

- **Análisis predictivo:** Soluciones de software y/o hardware que permiten a las empresas descubrir, evaluar, optimizar e implementar modelos predictivos analizando grandes fuentes de datos para mejorar el rendimiento del negocio o mitigar el riesgo.
- **Bases de datos NoSQL:** Sistemas de Bases de Datos alternativos a los tradicionales sistemas de Bases de Datos Relacionales, donde el modo de consulta de los datos difiere de la ejecución de sentencias SQL. Fundamentalmente se clasifican entre los siguientes cuatro tipos: Documentales, orientados a columnas, clave-valor y basados en grafo.
- **Search and Knowledge Discovery:** Herramientas y tecnologías para apoyar la extracción de autoservicio de información y nuevos conocimientos a partir de grandes repositorios de datos estructurados y no estructurados que residen en múltiples fuentes como sistemas de archivos, bases de datos, flujos, API (del inglés Application Programming Interface) y otras plataformas y aplicaciones.
- **Stream Analytics:** Software que puede filtrar, agregar, enriquecer y analizar con un alto rendimiento, datos de múltiples fuentes, dispares y en cualquier formato.
- **In-memory data Fabric:** Tecnologías de acceso de baja latencia y procesamiento de grandes cantidades de datos mediante la distribución de datos a través de la memoria dinámica de acceso aleatorio (DRAM), Flash o SSD de un sistema de computadora distribuido (Apache Ignite).
- **Almacenamiento Distribuido:** Red de ordenadores donde los datos se almacenan en más de un nodo, a menudo de forma replicada, para redundancia y rendimiento.
- **Virtualización de Datos:** Tecnología que proporciona información de varias fuentes de datos, incluyendo grandes fuentes de datos como Apache Hadoop y almacenes de datos distribuidos en tiempo real y tiempo cuasi-real.

- **Integración de Datos:** Herramientas para la orquestación de datos a través de soluciones como Amazon Elastic MapReduce (EMR), Apache Hive, Apache Pig, Apache Spark, MapReduce, Couchbase, Hadoop y MongoDB.
- **Preparación de datos:** Software que alivia la carga de aprovisionamiento, configuración, limpieza y distribución de conjuntos de datos diversos y desordenados para acelerar la utilidad de los datos para el análisis.
- **Calidad de los datos:** productos que realizan la limpieza y el enriquecimiento de datos en conjuntos de datos grandes y de alta velocidad, utilizando operaciones paralelas en almacenes de datos distribuidos y bases de datos.
- **Machine learning o aprendizaje automático:** herramientas para realizar de manera automática predicciones o sugerencias calculadas basadas en grandes cantidades de datos. Algunos de los ejemplos más comunes de aprendizaje de máquina son los algoritmos de Netflix para hacer sugerencias de películas basadas en películas que has visto en el pasado o los algoritmos de Amazon que recomiendan libros basados en libros que has comprado antes.

1.1.6 Herramientas Big Data

Durante mucho tiempo hablar de Big Data era hablar de Apache Hadoop, sin embargo en la actualidad han aparecido nuevas herramientas que completan o mejoran la captura, procesado, almacenaje y análisis de datos. Estas **nuevas herramientas** surgen en forma de ecosistema con múltiples componentes intercambiables en función de las necesidades del negocio. Además, estos nuevos ecosistemas permiten su utilización de forma combinada, lo que hace compleja su descripción puesto que depende de los componentes elegidos por el usuario. Teniendo en cuenta la existencia de múltiples combinaciones posibles y la rapidez en la realización de nuevos desarrollos también combinables, las **herramientas con mayor capacidad** en la actualidad son:

APACHE HADOOP

El framework **Apache Hadoop** nació en 2004 de la mano de Doug Cutting y Mike Cafarella, como la mejor solución para manejar grandes volúmenes de datos no estructurados y apoyar la distribución del motor de búsqueda Nutch.

Se trata de un proyecto Apache (organización no lucrativa) creada para dar soporte a proyectos de software bajo la propia denominación Apache. Apache Hadoop es un framework de código abierto para el almacenamiento y procesamiento de grandes volúmenes de datos de forma distribuida. Está licenciado bajo la licencia Apache 2.0.

El framework de Apache Hadoop se compone de los siguientes módulos:



ILUSTRACIÓN 2. COMPONENTES HADOOP. FUENTE: ELABORACIÓN PROPIA

Los principales componentes de la primera generación de Apache Hadoop, Apache Hadoop 1.x, son el Sistema de Archivos Distribuidos (HDFS) y el paradigma de procesamiento paralelo MapReduce. Ambos son proyectos de código abierto, inspirados en las tecnologías Google MapReduce y Google File System (GFS) creadas por Google.

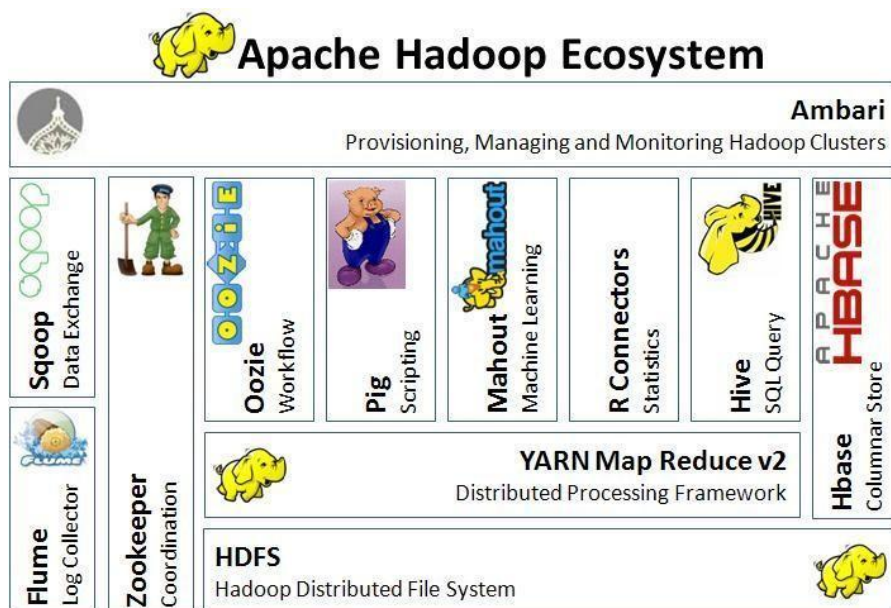


ILUSTRACIÓN 3. ECOSISTEMA HADOOP. FUENTE: HADOOP.APACHE.ORG

Sistema de ficheros distribuido de Hadoop (HDFS)

HDFS es un sistema de ficheros ejecutado en un clúster de máquinas de hardware básico y diseñado para almacenar grandes ficheros proporcionando acceso continuo a los mismos. Es decir, se asume que normalmente se realizan operaciones que implican leer gran parte o todo

el fichero, por lo tanto es más importante el tiempo de lectura de todo el conjunto de datos que la latencia de leer el primer registro. No ofrece acceso aleatorio. Por otro lado, Hadoop HDFS **no requiere un hardware costoso y de grandes prestaciones**, sino que está pensado para ejecutarse en un clúster de máquinas con un hardware básico y para las cuales los fallos de hardware son comunes. Así, HDFS está pensado para ofrecer tolerancia a fallos manteniendo el servicio sin interrupciones. En caso de caída de un nodo, HDFS tiene la capacidad de repartir sus bloques a otro nodo.

Al igual que un disco duro u otros sistemas de ficheros sobre un único disco, la unidad mínima de lectura y escritura de HDFS es el bloque. Sin embargo, el tamaño de bloque de HDFS es mucho mayor que en el caso de otros sistemas de ficheros. En HDFS el tamaño de bloque por defecto es de 128MB mientras que en otros sistemas de ficheros los bloques son típicamente unos pocos kilobytes. Al igual que en un sistema de ficheros corriente, HDFS divide los ficheros en bloques que se almacenan como unidades independientes. A diferencia de los sistemas de ficheros sobre un único disco, en HDFS un fichero que ocupe menos que el tamaño de bloque no ocupará el bloque completo. Los bloques HDFS son grandes en comparación con los bloques de disco con el objetivo de minimizar el costo de las búsquedas. Si el bloque es suficientemente grande, el tiempo que tarda en transferir los datos desde el disco puede ser significativamente más grande que el tiempo para buscar el inicio del bloque.

Para garantizar la disponibilidad de los bloques en caso de corrupción de bloques o fallos en las máquinas, cada bloque es replicado por defecto tres veces. Así, si un bloque no está disponible, se lee una copia de otra ubicación de forma transparente al cliente.

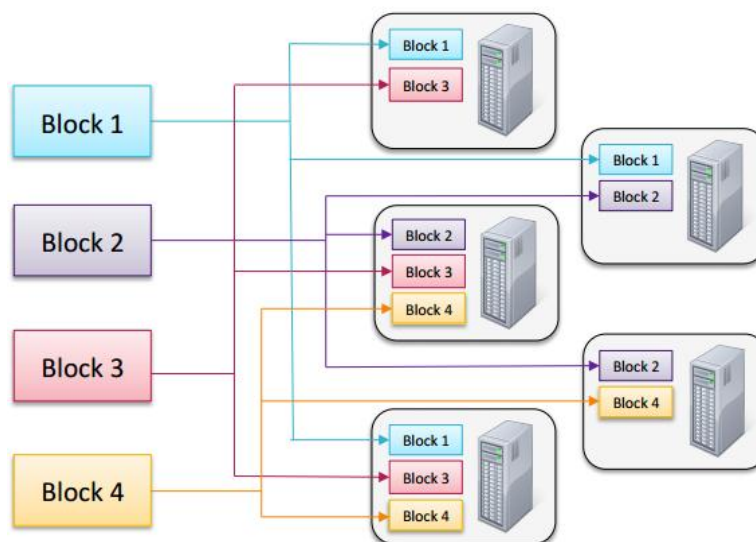


ILUSTRACIÓN: 4 DISTRIBUCIÓN DE LOS BLOQUES EN HDFS. FUENTE: WWW.CLOUDERA.COM

Un **clúster HDFS** está constituido por dos tipos de nodos sobre una arquitectura cliente-servidor o maestro-trabajador: un *namenode* (el maestro) y un número de *datanodes* (trabajadores). El namenode gestiona el espacio de nombres del sistema de ficheros, mantiene el árbol del sistema de ficheros, y los metadatos y directorios que constituyen el árbol. Esta información se almacena en el disco local en dos ficheros: la imagen del espacio de nombre y un fichero de log con los cambios realizados. Además, el namenode conoce también la localización física de los bloques

que constituyen el fichero, es decir, en qué datanodes están almacenados dichos bloques. No obstante, esta información no está persistida en disco localmente, sino que se obtiene de los datanodes.

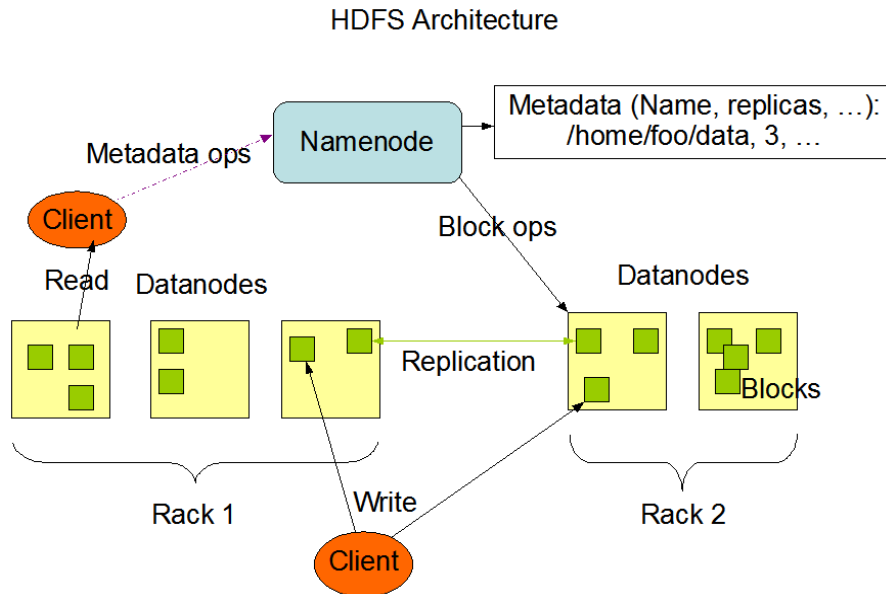


ILUSTRACIÓN: 5 ARQUITECTURA HDFS. FUENTE: HADOOP.APACHE.ORG.

Otro componente fundamental de HDFS es el **cliente**, que permite acceder al sistema de archivos HDFS y comunicarse con el namenode y los datanodes. Así, los datanodes son los encargados de almacenar y recuperar los bloques que son solicitados por los clientes o por el namenode. Además envían periódicamente al namenode la lista de bloques que están almacenando.

Paradigma de programación MapReduce

MapReduce es un modelo de programación para procesamiento de datos. Un trabajo MapReduce es una tarea que un cliente desea realizar y que está constituido por: los datos de entrada a analizar o procesar, un programa MapReduce, e información de configuración. MapReduce trabaja dividiendo el procesado en dos fases: la fase *map* y la fase *reduce*. Cada fase tiene pares clave-valor como entrada y salida. Las funciones map y reduce son especificadas por el cliente. Este modelo de programación evita el problema de lecturas y escrituras a disco, transformándolo en un problema de computación sobre conjuntos de clave-valor.

Los programas MapReduce son inherentemente paralelos, esto es, MapReduce **ayuda a resolver problemas** para los cuales el conjunto de datos puede subdividirse en pequeñas partes que son procesadas de manera independiente. Así, el sistema divide los datos de entrada en varios bloques, a cada uno de los cuales se le asigna la tarea *map* que puede procesar los datos en paralelo. Cada tarea map lee la entrada como un conjunto de pares clave-valor y produce un conjunto transformado de pares clave-valor como la salida. Finalmente, en la tarea *reduce* se

mezclan y ordenan las salidas intermedias de las tareas map y se agrupan produciendo un conjunto de pares clave-valor finales.

Hadoop, mediante su modelo de programación MapReduce, permite a usuarios ordinarios que no disponen de conocimientos de computación paralela, resolver problemas complejos sin necesidad de preocuparse de otros detalles tales como la comunicación entre las máquinas del clúster.

YARN (Yet Another Resource Negotiator)

Apache Hadoop YARN es un gestor de los recursos de un clúster introducido en Hadoop 2 con el objetivo de mejorar la implementación de MapReduce, aunque soporta otro tipo de paradigmas de programación distribuida. YARN proporciona una API para solicitar y trabajar con los recursos disponibles en un clúster, sin embargo el usuario normalmente no trabaja directamente contra dicha API sino que utiliza otros frameworks de computación distribuida a más alto nivel, como MapReduce o Apache Spark.

El objetivo de YARN es **independizar la gestión de recursos**, la planificación y monitorización de tareas de MapReduce, para ello dispone de dos procesos: *ResourceManager* (uno por clúster) para gestionar todos los recursos de todo el clúster y *NodeManager* (uno en cada nodo del clúster) encargado de lanzar y monitorizar los *Containers* que ejecutarán una tarea específica con los recursos (memoria, CPU,..) asignados.

En la evolución de Hadoop en el tiempo coexisten dos versiones, Hadoop 1 y Hadoop 2, es en la segunda en la que se incluye MapReduce 2 incluyendo YARN, proporcionando las siguientes ventajas: Escalabilidad (10000 nodos y 100000 tareas), disponibilidad, compatibilidad con MapReduce, utilización mejorada del clúster y soporte para modelos de programación distintos a MapReduce.

APACHE SPARK

Apache Spark es un framework de computación distribuida para procesar grandes volúmenes de datos. Originalmente fue desarrollado por AMPLab en la Universidad de Berkeley en California en 2009. Los creadores fundaron Databricks para comercializar Spark. Más tarde se convertiría en proyecto open source de Apache que incluiría a representantes de varias organizaciones incluyendo a Databricks, UC Berkeley, Cloudera, Yahoo e Intel.

Al contrario que muchos frameworks relacionados con Hadoop, no utiliza el motor MapReduce, utiliza su propio motor de ejecución distribuido para ejecutar trabajos en un clúster. Sin embargo está muy integrado con Hadoop, puede ejecutarse sobre YARN y trabajar sobre el sistema de ficheros HDFS.

Spark es conocido por su **capacidad para mantener en memoria grandes conjuntos de datos** entre trabajos. Esta capacidad permite a Spark superar el rendimiento respecto a MapReduce.

Los algoritmos iterativos (una función se aplica repetidamente sobre un conjunto de datos mientras se cumple una condición de salida) y el análisis interactivo (cuando el usuario solicita información del conjunto de datos), se ven beneficiados con Spark. Spark proporciona APIs para

escribir programas en cuatro lenguajes: Scala, Java, Python y R; y usa Resilient Distributed Dataset (RDD) que es la abstracción de una colección de objetos de sólo lectura distribuidos en varias máquinas de un clúster.

Spark ha demostrado ser una buena plataforma para construir herramientas de análisis. Para ello ofrece módulos de machine learning (MLlib), procesamiento de grafos (GraphX), streaming (Spark Streaming) y SQL (Spark SQL).

Ofrece dos tipos de operaciones sobre RDD: **transformaciones y acciones**. Una transformación genera un nuevo RDD a partir de otro, mientras que una acción desencadena una computación en un RDD cuyos resultados se pueden procesar, devolver al usuario o almacenar en almacenamiento externo. Las acciones tienen un efecto inmediato, pero las transformaciones no se ejecutan hasta que otra acción es ejecutada.

Spark usa el concepto de “job” (trabajo), al igual que MapReduce. Sin embargo en Spark este concepto es más general, consiste en un grafo acíclico dirigido (DAG) de fases equivalentes a las fases map y reduce. Un job se ejecuta en un contexto de aplicación, representado por una instancia SparkContext, que sirve para agrupar y compartir variables. Existen dos maneras de ejecutar un trabajo de Spark, de manera interactiva mediante una sesión de spark-shell o mediante spark-submit.

OTRAS HERRAMIENTAS

Apache Avro: Sistema de serialización de datos independiente del lenguaje. Fue creado por Doug Cutting para abordar el principal inconveniente de Hadoop Writables: la falta de portabilidad del lenguaje. Los datos de Avro se describen usando un esquema que está siempre presente en tiempo, lectura y escritura.

Cuando los datos de Avro se almacenan, su esquema se almacena con él, de modo que los archivos pueden ser procesados más tarde por cualquier programa. Los esquemas de Avro son escritos normalmente en JSON y los datos son codificados en formato binario. Existe un lenguaje de más alto nivel llamado Avro IDL, cuyo objetivo es permitir a los desarrolladores crear esquemas en un lenguaje más parecido a los lenguajes comunes de programación como Java, C++ o Python.

HBase: Se trata de una base de datos columnar que se ejecuta sobre HDFS. Aunque originalmente se ha definido como “column-oriented database”, quizás es más correcto describirla como una “column-family-oriented database” porque las especificaciones de almacenamiento se realizan a nivel de columna familiar. Las aplicaciones almacenan datos en tablas etiquetadas. Las tablas de HBase están constituidas por filas y columnas, donde las celdas son versionadas. Por defecto, la versión es un timestamp autoasignado en el momento de la inserción. El contenido de la celda es un array de bytes. Las filas son ordenadas por una clave, la clave primaria de la tabla, que también es un array de bytes.

Las columnas se agrupan en familias de tal forma que todos los elementos (todas las filas) de la familia de columnas se almacenan juntos en el sistema de ficheros. Esto es distinto a las bases de datos relacionales orientadas a filas, donde todas las columnas de una fila dada son almacenadas en conjunto. Facebook utiliza HBase en su plataforma desde noviembre del 2010.

Apache Hive: Constituye uno de los principales elementos de la plataforma de información de Facebook dirigida al procesamiento y generación de información. Es una infraestructura de data warehouse que se desarrolló por la necesidad de administrar e interpretar grandes volúmenes de datos generados diariamente en la red social Facebook y almacenados en HDFS. Hive fue creado para permitir a analistas con altos conocimientos de SQL, ejecutar consultas sobre grandes volúmenes de datos almacenados en HDFS. Así Hive se ejecuta en el propio nodo y convierte las consultas SQL en procesos MapReduce ejecutados en un clúster de Hadoop. Para ello define un lenguaje SQL propio, similar a MySQL, denominado Hive Query Language (HQL).

Apache Sqoop: Herramienta diseñada para transferir de forma eficiente datos en masa entre Apache Hadoop y otros almacenes de datos estructurados como bases de datos relacionales. Este proceso puede realizarse mediante programas MapReduce o mediante otras herramientas de alto nivel como Hive (es posible usar Sqoop para mover datos estructurados a HBase). Además, una vez procesados los datos en Apache Hadoop, Apache Sqoop puede exportar los resultados obtenidos en el análisis de nuevo a los almacenes estructurados para su consumo por otros clientes.

Apache Flume: La tarea principal de Apache Flume es establecer un canal para dirigir los datos desde una fuente de datos a otra. El destino usual de los datos de Apache Hadoop es HDFS pero podrían ser otros sistemas como HBase, Solr, etc. En Apache Flume existen tres entidades principales como se puede ver en la imagen: *sources*, *decorators* y *sinks*. Un source es básicamente cualquier fuente de datos, sink es el destino de una operación específica y un decorator es una operación dentro del flujo de datos que transforma esa información de alguna manera, como por ejemplo comprimir o descomprimir los datos o alguna otra operación en particular sobre los mismos. Los ejemplos típicos de uso de Apache Flume son recolectar los ficheros de log de los servicios web de bancos u otras organizaciones o los tweets de la red social Twitter y almacenarlos en HDFS para su posterior procesamiento.

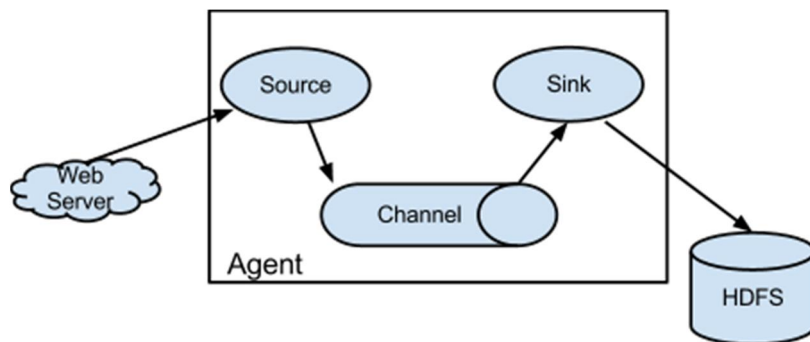


ILUSTRACIÓN 6: APACHE FLUME: COMPONENTES. FUENTE: FLUME.APACHE.ORG

Apache Kafka: Es una plataforma distribuida de streaming que ofrece las siguientes capacidades:

- Publicar y suscribirse a los flujos de datos. Es decir, es similar a una cola de mensajes o sistema de mensajería.
- Almacenar flujos de datos de forma tolerante a fallos.
- Procesar flujos de datos a medida que ocurren.

Kafka se ejecuta como un clúster en uno o más servidores. El clúster de Kafka almacena los flujos de datos en categorías llamadas “topics”, donde cada registro consta de una clave, un valor y una marca de tiempo.

Como se puede observar en la imagen que se muestra a continuación, Kafka consta de cuatro APIs principales:

- **API Producer:** aplicación para publicar un flujo de datos en uno o más topics Kafka.
- **API Consumer:** permite que una aplicación se suscriba a uno o más topics y procese el flujo de datos producido por ellos.
- **API Streams:** permite que una aplicación actúe como un procesador de flujo, consumiendo un flujo de entrada de uno o más topics y produciendo un flujo de salida a uno o más topics de salida.
- **API Conector:** permite crear y ejecutar productores o consumidores reutilizables que conectan topics de Kafka a aplicaciones o a sistemas de datos existentes. Por ejemplo, un conector a una base de datos relacional puede capturar cada cambio en una tabla.

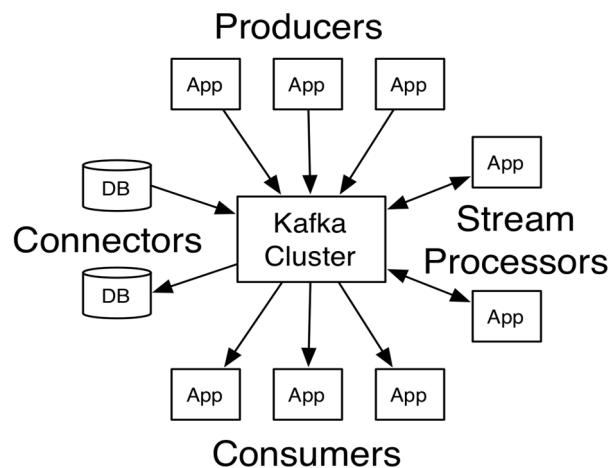


ILUSTRACIÓN 7: APACHE KAFKA. FUENTE: [KAFKA.APACHE.ORG/INTRO](https://kafka.apache.org/intro)

Apache Oozie: Sistema planificador de flujo de trabajo para administrar los trabajos de Apache Hadoop. El flujo de trabajo en Oozie se define como un grafo acíclico dirigido (DAG) de acciones. Es decir, sólo hay un punto de entrada y de salida y todas las tareas y dependencias parten del punto inicial al punto final, sin puntos de retorno.

Oozie está integrado con Hadoop soportando Java MapReduce, Streaming MapReduce, Apache Pig, Apache Hive, Apache Sqoop.

Apache Pig: Se trata de una plataforma de alto nivel para el análisis de grandes conjuntos de datos. Consiste en un lenguaje de alto nivel para expresar los programas de análisis de datos, llamado Pig Latin, y en la propia infraestructura para ejecutar dichos programas. La característica más destacada de los programas de Pig es que su estructura es susceptible de una gran paralelización, lo que a su vez les permite manejar conjuntos de datos muy grandes.

ZooKeeper: Apache ZooKeeper es un servicio de coordinación de alto rendimiento para aplicaciones distribuidas. Es un servicio centralizado para mantener información de configuración, proporcionar sincronización distribuida y servicios de grupo. Se puede usar para

implementar el consenso, la administración del grupo, la elección del líder y los protocolos de presencia. Ofrece una serie de herramientas para construir aplicaciones distribuidas que puedan de forma segura manejar fallos parciales.

Apache Mahout: El principal objetivo de Apache Mahout es proporcionar un entorno para crear de forma rápida, sencilla y robusta aplicaciones de aprendizaje máquina escalables. Su nombre representa su íntima relación con Apache Hadoop, a pesar de esto ofrecen algoritmos que no dependen de Hadoop, y su intención es proveer en el futuro implementaciones para plataformas más adecuadas para el aprendizaje máquina, como Apache Spark.

Apache Lucene: Es un proyecto de Apache bastante popular para realizar búsquedas sobre textos. Lucene provee de librerías para indexación y búsqueda de texto independientemente del formato del archivo. Ha sido utilizado principalmente en la implementación de motores de búsqueda aunque debe considerarse que no incluye funciones de "crawling" ni análisis de documentos HTML. El concepto a nivel de arquitectura de Lucene es simple, básicamente los documentos son divididos en campos de texto generando un índice sobre ellos. Apache Lucene comúnmente se implementa junto a Nutch, un robot y motor de búsqueda que forma parte del mismo proyecto. Nutch es una tecnología desarrollada en Java, y basa su arquitectura en la plataforma Hadoop.

ElasticSearch: Servidor de búsqueda basado en Lucene. Provee un motor de búsqueda de texto completo, distribuido y con capacidad de servir a múltiples clientes mediante servicios Web. Elasticsearch, permite búsqueda escalable, con ejecución casi en tiempo real. Desarrollado en Java y publicado como código abierto bajo las condiciones de la licencia Apache, ofrece un conector para trabajar con Hadoop llamado Elasticsearch-Hadoop para búsquedas en tiempo real en entornos Big Data.

Apache Zeppelin: Interfaz Web que permite realizar un análisis interactivo de datos de forma simple y visual sobre grandes volúmenes de datos gestionados a través de Apache Spark. Está en el concepto web notebook, introducido por iPython, que permite trabajar sobre un interfaz web en lugar de sobre un shell. Zeppelin está pensado para poder integrarse perfectamente con Spark y Hadoop.

Apache Flink: Framework de procesamiento distribuido pensado inicialmente para aplicaciones de streaming. Este framework trata los procesamientos por lotes como un caso especial de flujo de datos, de forma que se puede considerar un framework de procesamiento tanto por lotes como en tiempo real. La herramienta incluye una serie de APIs, una API de streaming para Java y Scala, una API de datos estáticos para Java, Scala, y Python y una API de consultas "SQL-like" para código embebido en Java y Scala. Dada su orientación al flujo de datos su aplicación puede ser preferible sobre Apache Spark en dominios en los que se procesan flujos de datos en tiempo real.

Apache Storm: Sistema de computación distribuida en tiempo real, cuyas aplicaciones están diseñadas como grafos acíclicos dirigidos. Storm está diseñado para procesar fácilmente flujos sin límites y puede utilizarse con cualquier lenguaje de programación. Es altamente escalable y proporciona garantías en los trabajos de procesamiento. Apache Storm se puede utilizar para análisis en tiempo real, aprendizaje máquina distribuido, y muchos otros casos, especialmente los de alta velocidad de datos. Storm puede funcionar en YARN e integrarse en los ecosistemas

de Hadoop, proporcionando a las implementaciones existentes una solución para el procesamiento de flujo en tiempo real.

BASES DE DATOS A GRAN ESCALA

Hasta hace unos años, los vastos conjuntos de datos eran tratados en **Sistemas Data Warehouse**, mediante procesos en lotes, para obtener datos agregados que aportarían información a un sistema de Business Intelligence (BI). Pero en la actualidad surge la necesidad de obtener muchos de estos datos agregados de manera inmediata (tiempo real) y no es asumible el tiempo que supone la ejecución de dichos procesos.

En este nuevo contexto, se extiende el uso del término **NoSQL (Not only SQL)**, que se refiere a los Sistemas de Gestión de Base de Datos que difieren del modelo clásico de implementación relacional (RDBMS), fundamentalmente, en que no utilizan SQL como lenguaje principal para realizar consultas, no permiten JOINS (uniones entre diferentes conjuntos de datos), la mayoría no siguen el principio ACID (Atomicity, Consistency, Isolation and Durability) y son escalables horizontalmente. Este término fue acuñado por Carlo Strozzi en 1998 y fue reutilizado por Eric Evans en 2009, y el mismo sugiere la utilización del término Big Data para esta nueva generación de Sistemas de Gestión de Bases de Datos.

SISTEMAS DE GESTIÓN DE BASES DE DATOS RELACIONALES

Los **Sistemas de Gestión de Base de Datos Tradicionales**, basados en el modelo relacional, no han de ser descartados para un contexto Big Data, son una buena opción siempre que se gestionen con técnicas adecuadas de distribución y escalado. Las mayores ventajas de los Sistemas de Gestión de Bases de Datos Relacionales son la integridad inherente de los datos y la atomicidad de las operaciones, frente a la penalización que esto supone en el rendimiento en la inserción de datos. Por el contrario, entre sus mayores desventajas se encuentra la lentitud y otros inconvenientes que se presentan al realizar cambios de estructura de los datos y el alto coste económico de la escalabilidad en este modelo. Entre los principales Sistemas de Gestión de Bases de Datos destacan: PostgreSQL, MySQL, Oracle y SQL Server.

Las técnicas más relevantes para abordar el escalado en el modelo relacional son:

- **Actualización Hardware:** Consiste en el escalado vertical, es decir, mejorar las características Hardware de la máquina sobre la que se ejecuta el Sistema de Gestión de Bases de Datos. Entre las mejoras posibles que afectarían al rendimiento en las consultas, destacan:
 - **Soporte de Almacenamiento:** Es posible cambiar la tecnología de almacenamiento desde un disco duro tradicional mecánico a un disco duro de estado sólido para el que los accesos aleatorios y la transferencia sostenida en lectura y escritura es mucho mayor. También existe en este aspecto, la posibilidad de utilizar redundancia mediante los diferentes niveles de RAID. Esto permite, según el nivel elegido, que los accesos aleatorios a los datos y la transferencia media de lectura o escritura sea más rápida.
 - **CPU:** Mediante el incremento de la velocidad del procesador, y el número de

cores o procesadores, es posible aplicar los cálculos de las funciones y las combinaciones de los JOINS, de manera mucho más ágil, sobre los datos que se encuentran fundamentalmente en memoria.

- **Memoria RAM:** Con mayor cantidad de memoria RAM es posible tener una caché más grande, lo que permite acelerar en mayor medida las diferentes consultas que se realicen con más frecuencia.
- **Vistas Materializadas:** Las vistas tradicionales permiten en SQL introducir una capa de abstracción, que permite ocultar la forma de recoger los datos. Por ejemplo, si para obtener ciertos datos necesitáramos hacer una serie de JOINS, mediante una vista, podríamos ocultar estas operaciones, ofreciendo la vista como si se tratase de una única tabla. Precisamente, los JOINS, debido a la inherente combinatoria de los mismos, incrementan de manera exponencial el tiempo de consulta en función del número de filas de cada tabla y del número de ellos que se realicen en la misma consulta. Es aquí, donde las vistas materializadas, permiten disminuir los tiempos de consulta. Esto se consigue ejecutando una consulta a una vista tradicional y almacenando el resultado. De esta manera cada vez que se consulte la vista materializada, se recogerán los datos de la misma, en el mismo tiempo que si se tratase de una única tabla. Para garantizar la coherencia entre la vista materializada y las tablas que la nutren, existen fundamentalmente dos operativas:
 - Actualización de la vista en el momento de actualización en tabla.
 - Actualización de la vista mediante procesos batch.
- **Partitioning:** A medida que aumenta el número de filas en una tabla, aumenta el tiempo de consulta sobre la misma, y además afecta de manera exponencial en los JOINS. Al trabajar con magnitudes elevadas, que superan el orden de millones de filas, el tiempo de ejecución de las consultas se convierte en excesivo. Realizando la partición de una tabla en función del valor para una columna determinada, se dispone de subtablas con menos filas. De esta manera, ejecutando consultas SQL, que aprovechan esta segmentación de los datos, los tiempos de las mismas serán mucho menores al reducir las combinaciones en el momento de realizar los JOINS.
- **Clustering:** Consiste en hacer uso del escalado horizontal, para distribuir las operaciones del Sistemas de Bases de Datos entre diferentes nodos. Existen diferentes arquitecturas, que dependen del modo en que se particionan los datos y se replican entre los diferentes nodos de un clúster. El Clustering puede ofrecer una distribución en el pool de conexiones entre nodos de manera que las consultas de diferentes clientes sean atendidas por distintos nodos de manera distribuida.

SISTEMAS DATA WAREHOUSE

Data Warehouse, es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Puede considerarse una evolución de los sistemas de bases de datos relacionales, pero es un proceso en sí mismo y no un producto. Data warehouse es un término que data de 1988, acuñado por los investigadores de IBM Barry Devlin y Paul Murphy, no obstante es William Harvey Inmon el que se considera el padre de los Data Warehouse y cuya definición dice:

“Una colección de datos que sirve de apoyo a la toma de decisiones, organizados por temas, integrados, no volátiles y en los que el concepto de tiempo varía respecto a los sistemas tradicionales”

Los Data Warehouse fueron creados en la década de los 90 y son un conjunto de datos que las organizaciones utilizan como apoyo a la toma de decisiones y que pueden ser consultados mediante las tecnologías de los Data Mining. Las principales características que definen los Data Warehouse son:

- **Organizado por temas:** La organización por temas hace referencia al hecho de que los datos se organizan de acuerdo a su semántica, independientemente de que aplicación los utilice. De este modo, una compañía podría tener datos organizados por clientes, proveedores, productos...
- **La integración:** Un Data Warehouse se compone de los datos que se obtienen de las diversas fuentes de información de una organización, lo cual implica “poner en común” toda esta información. La principal ventaja que se deriva del proceso de integración se centra en la agrupación única de la estructura de la información evitando así el problema que surge cuando cada una de las fuentes de datos de una organización disponga de sus propios modelos estructurados, sus propias políticas de asignación de nombres a campos, seguridad, y un sin fin de diferencias.
- **No volatilidad:** La principal función de un Data Warehouse es dar soporte a la toma de decisiones. Para ello, Data Warehouse, no realiza actualizaciones de los datos, sino que mantiene las diferentes versiones de los datos a lo largo del tiempo, permitiendo recuperar el estado de los mismos en una organización en cualquier instante.
- **Temporalidad:** Los datos del Data Warehouse tienen un horizonte temporal de vida de entre 5 y 10 años. Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones. En los sistemas de gestión, los datos con los que se trabaja son los datos actuales, mientras que los datos del Data Warehouse pueden verse como una serie de “snapshots” o fotografías tomados en un momento del tiempo, que no sufren actualizaciones.

En base a lo anterior podemos destacar innumerables ventajas de los sistemas Data Warehouse en el apoyo en la toma de decisiones de la empresa a cualquier nivel jerárquico, ya que consigue proporcionar mejores productos a través de la **optimización de tiempos de producción y toma de decisiones**. También permite analizar información relevante para la empresa con carácter diario permitiendo agilizar la toma de decisiones que puedan afectar el desempeño o proyección de la empresa.

No obstante, también existen ciertas desventajas a tener en cuenta como el **alto coste** que conlleva la implementación del mismo. Desde la mera puesta en marcha del almacén de datos, hasta el coste de mantenimiento pasando por los costos de adaptación de la empresa, formación, mantenimiento, coste del Software y Hardware.

SISTEMAS DE BASES DE DATOS NOSQL

Los Sistemas de Bases de Datos NoSQL aportan: una gran flexibilidad en el cambio dinámico de las estructuras de los datos, carácter descentralizado, y un coste económico muy inferior al de los sistemas tradicionales para realizar el escalado en el volumen de datos, dado que soporta escalabilidad horizontal (crecimiento en número de máquinas). Su principal hándicap, es que no todos los Sistemas de Bases de Datos NoSQL ofrecen la atomicidad de instrucciones y la integridad de los datos. Otras pequeñas limitaciones serían la falta de estandarización y la poca usabilidad en algunas de las herramientas de administración de estos sistemas.

De entre los diferentes tipos de Sistemas de Bases de Datos NoSQL, podemos distinguir principalmente las siguientes 4 clases:

- **Sistemas de Bases de Datos Documentales.**

En las bases de datos documentales los registros se consideran documentos con un esquema flexible que depende de los atributos de cada documento. Suelen emplear codificación JSON, BSON o XML. Se podría decir que son los Sistemas de Bases de Datos NoSQL más versátiles. Las implementaciones de este tipo de Sistemas de Bases de Datos, más comunes son:

- **MongoDB:** Es probablemente uno de los Sistemas de Bases de Datos NoSQL más utilizado actualmente. MongoDB es usado por compañías como: Cisco, Bosch, Ebay, Forbes IBM, Windows Azure, McAfee, el periódico The Guardian, el periódico New York Times, el periódico Le Figaro, el CERN...
- **CouchDB:** Es el sistema de bases de Datos NoSQL de Apache. Una de sus características más interesantes es que los datos son accesibles a través de la API REST.
- **Otros Ejemplos:** SimpleDB, RavenDB, BaseX, djondb, eXist, IBM Lotus Domino, Terrastore, Riak.

- **Sistemas de Bases de Datos orientados a Columnas/Tabulares.** El origen de este tipo de sistemas reside en posibilitar la realización de consultas y agregaciones sobre cantidades enormes de datos. Su funcionamiento es similar al de los Sistemas de Bases de Datos Relacionales, pero el almacenamiento se realiza por columnas, en lugar de por filas (registros). En esta categoría encontramos los siguientes Sistemas NoSQL:

- **Cassandra:** Mantiene un esquema híbrido entre Orientación a Columnas y Clave-Valor.
- **HBase:** Desarrollada sobre Java y mantenida por el proyecto Hadoop de Apache.
- **Otras:** Apache Accumulo, BigTable, Hypertable, Mnesia, OpenLink Virtuoso, LevelDB (Versión abierta de BigTable).

- **Sistemas de Bases de Datos Clave-Valor.** Son quizás, las que presentan un esquema más sencillo, almacenando la información en un esquema simple de clave-valor, lo que se traduce en que no se requiere de un modelo de datos fijo y se puede almacenar en una estructura de datos u objeto de cualquier lenguaje de programación. Las implementaciones de este tipo de Sistemas de Bases de Datos, más comunes son:

- **DynamoDB:** Desarrollada por Amazon, exponiendo un modelo de datos similar y que deriva del esquema de Dynamo. Es una opción de almacenamiento de AWS (Amazon Web Services).
- **Apache Cassandra:** Dispone de un esquema mixto entre orientación a columnas y Clave-Valor.

- **Redis:** Desarrollada en lenguaje C y orientada para trabajar con en el conjunto total de datos residiendo en memoria RAM.
- **Otras:** Hibari, Freebase, Scalaris, MemcacheDB, Berkeley DB, Voldemort, Tokyo Cabinet, KAI.
- **Sistemas de Bases de Datos en Grafo.**

En estos sistemas la información reside sobre en un grafo donde las entidades son representadas por nodos y las relaciones por las aristas de los mismos. Esto permite hacer uso de la teoría de grafos para realizar las consultas sobre la Base de Datos. Las implementaciones de este tipo de Sistemas de Bases de Datos, más comunes son:

- **InfiniteGraph:** Desarrollado en Java por la compañía Objectivity. Los desarrolladores hacen uso de este sistema para descubrir conexiones ocultas entre conjuntos de datos Big Data, con un grado de conectividad elevado.
- **Neo4j:** Sistema de Base de Datos de código abierto, desarrollado en Java por la compañía Neo Technology. Es empleada por compañías como HP, Infojobs o Cisco.
- **Otras:** DEX/Sparksee, AllegroGraph, OrientDB, Sones GraphDB, InfoGrid, HyperGraphDB.

Además de estas tipologías, debemos resaltar otros dos sistemas cuyas características hacen que no puedan incluirse en las tipologías anteriores:

- **Sistemas de Bases de Datos Multivalor.** Se distinguen de los Sistemas de Bases de Datos tradicionales, fundamentalmente, en que pueden hacer uso de atributos que almacenan una lista de valores en lugar de un valor único. Son ejemplos de estas bases de datos: *Northgate Information Solutions, Extensible Storage Engine, jBase, Rocket U2, OpenInsight, OpenQM, Reality, InterSystems Caché, D3 Pick database, InfinityDB.*
- **Sistemas de Bases de Datos Orientados a Objetos.** Son los Sistemas de Bases de Datos que combinan las funcionalidades propias de este tipo de sistemas con las capacidades de un lenguaje orientado a objetos. *Ejemplos: db4o, Eloquera, GemStone/S, InterSystems Caché, JADE, NeoDatis ODB, ObjectDatabase++, ObjectDB, Objectivity/DB, ObjectStore, ODABA, Perst, OpenLink Virtuoso, Versant Object Database, Wakanda, ZODB.*

En el marco de las Ciencias de la Computación, y en especial en el contexto NoSQL, dado su carácter distribuido, tiene una importante relevancia el **Teorema de CAP**, también conocido como **Teorema de Brewer**, el cual enuncia que no es posible para un sistema de cómputo distribuido garantizar simultáneamente más de dos de las siguientes tres condiciones:

- **Consistencia** (Consistency): Todos los nodos han de poder ver la misma información de manera simultánea.
- **Disponibilidad** (Availability): Cada petición a un nodo tiene la garantía de recibir siempre una respuesta de confirmación o de error.
- **Tolerancia al Particionado** (Partition Tolerance): El sistema continúa operativo aunque se hayan perdido o retrasado un número arbitrario de mensajes entre los nodos de la arquitectura.

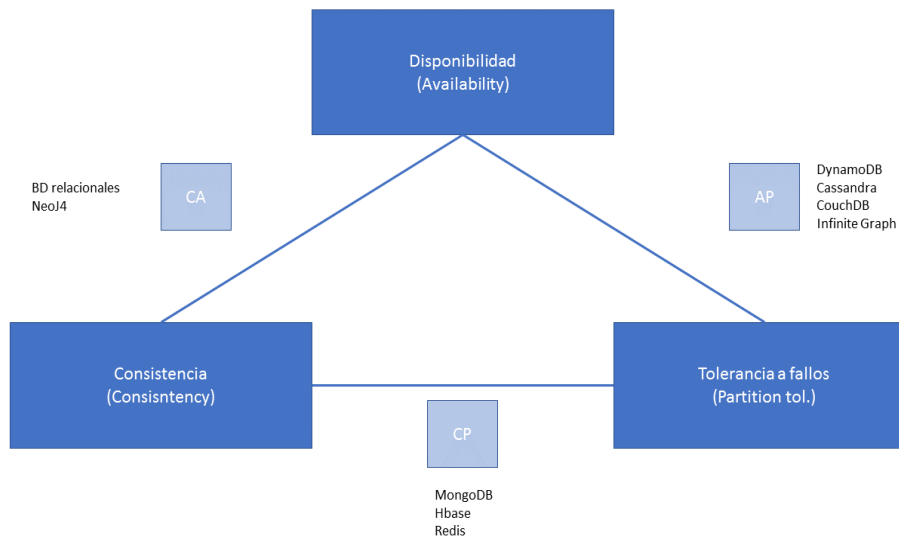


ILUSTRACIÓN 8. SITUACIÓN DE ALGUNAS BASES DE DATOS RESPECTO AL TEOREMA CAP. FUENTE: ELABORACIÓN PROPIA

De esta manera, los Sistemas de Bases de Datos NoSQL y los Sistemas Tradicionales Relacionales, pueden ser clasificados según cumplan o no las condiciones anteriores:

- **AP (Availability & Partition Tolerance):** Cassandra, CouchDB, SimpleDB, DynamoDB, Voldemort, Tokyo Cabinet, KAI, Riak.
- **CP (Consistency & Partition Tolerance):** MongoDB, Redis, BigTable, HyperTable, HBase. Terrastore, Scalaris, Paxos, Berkeley DB, MemcacheDB.
- **CA (Consistency & Availability):** Relational Database Management Systems (RDBMS), Aster Data, Greenplum, Vertica.

1.1.7 Big Data: casos de uso

Las aplicaciones y usos de Big Data son múltiples y dependen de cada sector. Los sectores donde más se ha expandido y donde su futuro crecimiento es más prometedor son [12]:

- **Telecomunicación:** Las compañías de telecomunicaciones están entre los precursores de la adopción de Big Data. La explosión de datos, impulsada por aplicaciones tales como registros de llamadas, monitoreo del tráfico de red, contenido digital, la gestión de activos, hacen que este sector sea uno de los pioneros en el uso de estas tecnologías.
- **Energía:** La introducción de medidores inteligentes, sensores de todo tipo, etc. ha incrementado la cantidad de datos disponibles, por lo que se prevé que el uso de Big Data tendrá un incremento exponencial en los próximos años con el objetivo de un mejor desempeño empresarial.
- **Servicios financieros:** El uso de Big Data focalizado en la identificación de perfiles conductuales, identificación y evaluación del riesgo, o en la identificación de las oportunidades de inversión, son algunos de los intereses más claros para el sector.
- **Fabricación:** En este sector la adopción de Big Data presenta diferencias entre los distintos subsectores. Es en subsectores como la automoción o aeronáutica donde la tecnología es más aceptada. Sin embargo, los beneficios que puede aportar Big Data aplicado a la industria, junto con el apoyo de los estados en la aplicación de estas

tecnologías, harán que su crecimiento se fortalezca en los próximos años.

- **Salud:** En la actualidad, se están desarrollando utilidades del Big Data en múltiples ámbitos de la salud: genómica, investigación clínica, epidemiología, monitorización y seguimiento de enfermos crónicos, operativa clínica, farmacología,
- **Otros:** Con menor intensidad Big Data empieza a abrirse camino en sectores como el transporte, o los servicios profesionales. En el caso del transporte las grandes compañías serán las primeras en adoptar esta tecnología, mientras que las pequeñas compañías parecen ser más reacias.

Como ejemplos de casos de uso reales de Big Data en distintos sectores podemos identificar:

BELK

Aplicación: Aumentar el número de clientes multicanal, optimizar el stock, optimizar el formato de la tienda, y las decisiones de horario

Resultado: Analizar a millones de consumidores a través de sus bases de datos con censo de cliente, etnia, emigrantes, etc.



Smarter cities

Aplicación: Optimizar las inversiones de las ciudades en infraestructura de transporte mejorando el flujo de tráfico.

Resultado: Se almacenan datos dispares como tráfico por carril, características de los vehículos individuales, acciones realizadas por los vehículos. El crecimiento del volumen de datos es continuo y en tiempo real.



CARGO SMART

Aplicación: Logística, diagnóstico y pronóstico. Mejora de las rutas de las buques para ahorro de combustible, optimización de la tripulación y gastos portuarios

Resultado: CargoSmart gestiona datos sobre la velocidad de los buques, la ubicación en el puerto y en el mar, los tiempos de tránsito totales, rutas, eventos de atraque, eventos de excepción de 5500 buques - TIBCO



Rio Tinto

Aplicación: Minería y Extracción. La mina del futuro: Automatización y análisis en la minería de hierro en Australia

Resultado: RTV "VirtualEYES" Herramienta de modelado 3D, controlando el sistema de automatización de las minas en tiempo real. Más de 200 sensores embebidos en camiones, alimentando



80.000 piezas de datos por segundo a operadores remotos. Inspección de minas a través de UAV.

STUBHUB

Aplicación: Procesado y analizado de grandes volúmenes de datos en tiempo real para fortalecer la experiencia del cliente, y mejorar la detección y prevención del fraude



Resultado: Información analizada proveniente de 25 recursos de datos en un solo data warehouse. Permitiendo analizar a 180 millones de clientes y desarrollando tickets promocionales y campañas de promoción específicas.

1.2 DATA ANALYTICS

1.2.1 Breve historia

El término “**Data Science**” o Ciencia de los Datos surge para dar sentido a una nueva disciplina cuyo fin último es hacer inteligibles los grandes volúmenes de datos que se encuentran dentro de Big Data.

En 1962 John W. Turkey [13] escribe “El futuro del Data Análisis” y en 1977 publica el artículo “Exploratory Data Analysis”, en el que se argumenta que la importancia radica en el uso de los datos para sugerir hipótesis que permitan testear y explorar los mismos, permitiendo extraer conclusiones veraces. Ese mismo año la International Association for Statistical Computing (IASC) establece que su misión principal consistirá en unir la estadística tradicional, la tecnología informática y el conocimiento experto, para convertir los datos en información y conocimiento.

En 1989 se organiza en Estados Unidos la primera conferencia anual sobre Descubrimiento de conocimiento y minería de datos (KDD, del inglés Knowledge Discovery in Databases).

En 1996, se utiliza por primera vez el término **Data Science** en la cumbre bienal de la International Federation of Classification Societies, en una conferencia de título “*Data Science, classification and related methods*”

En 1997 se lanza la revista “*Data Mining and Knowledge Discovery*” (Minería de datos y Descubrimiento del Conocimiento), como puede verse se invierte el orden de las denominaciones en el título, respecto a 1989, lo que refleja la importancia creciente de la Minería de Datos como el método más popular para extraer información y conocimiento de las grandes bases de datos.

En septiembre de 2005 The National Science Board publica “*Long lived Digital Data Collections: Enabling Research and Education in the 21st*” En el informe se define a los Científicos de Datos como: “*informáticos, ingenieros y programadores de bases de datos y de software, expertos en estadística, bibliotecarios, y otros, cruciales para el éxito de la gestión de una colección de datos digitales*”.

En 2009 se publica el informe de la Interagency Working Group on Digital “*Harnessing the power of Digital Data for Science and Society*”. En él se establece que se necesita identificar y promover la aparición de nuevas disciplinas y especialistas expertos en abordar los retos complejos y dinámicos de la preservación digital, el acceso sostenido a los datos, y la reutilización de datos.

Podríamos decir que los primeros análisis de datos se ejecutaron con las primeras hojas de cálculo en los años 50 para la apoyar la toma de decisiones en las empresas. Estas hojas de cálculo van evolucionando de la mano de las compañías de IT y SAS, pero trabajando siempre con datos estructurados. Es, hace una década cuando, en Silicon Valley, empiezan a emerger las aplicaciones para tratar información desestructurada y poder manejar los enormes volúmenes de datos existentes hoy en día.

Data Analytics está íntimamente ligado a Big Data y en muchos informes aparece la expresión Big Data Analytics, entendiéndose por ello el conjunto de herramientas que permiten explotar los datos de las enormes bases de datos conocidas como Big Data. En este sentido podríamos hablar de: analítica predictiva, minería de datos, análisis estadístico, etc. Incluso en algunos informes

[14] se habla de visualización de datos, inteligencia artificial y bases de datos capaces de soportar análisis como MapReduce, in-memory databases o almacenes de datos columnares.

1.2.2 Definición/Descripción

“Data Analytics es la ciencia de examinar datos en bruto con el propósito de sacar conclusiones sobre esa información.”

Data Analytics implica aplicar un proceso algorítmico o mecánico para obtener conocimiento; por ejemplo, aplicar un proceso para buscar correlaciones significativas entre varias series de datos. Las técnicas usadas tradicionalmente en productos **BI (Business Intelligence)** para analizar y generar conocimiento de los datos en bruto están pensadas para trabajar con datos estructurados. Estas técnicas no son suficientes para manejar Big Data, que engloba tanto datos estructurados como semi estructurados y no estructurados. Por ello aparece un nuevo término: “Big Data Analytics”.

“Big Data Analytics es el proceso de examinar grandes conjuntos de datos para descubrir patrones ocultos, correlaciones desconocidas, tendencias de mercado, preferencias de los clientes u otra información de negocio útil. Los resultados analíticos pueden conducir a una comercialización más efectiva, detección de nuevas oportunidades de ingresos, un mejor servicio al cliente, una mayor eficiencia operativa, ventajas competitivas sobre las organizaciones rivales y otros beneficios empresariales.

El enfoque de Data Analytics reside en la inferencia, que es el proceso de derivar conclusiones que se basan únicamente en lo que el investigador ya conoce.”

Así, el fin último del **Big Data Analytics** es proporcionar a las organizaciones y empresas un mecanismo para tomar mejores decisiones, conocer mejor su negocio, generar posibles oportunidades de negocio y verificar o refutar teorías y modelos existentes.

El término Data Analytics se refiere a un conjunto de aplicaciones que van desde la inteligencia empresarial básica o Business Intelligence (BI), la elaboración de informes y el procesamiento analítico en línea (OLAP, del inglés Online Analytical Processing) hasta diversas formas de análisis avanzado como machine learning o data mining. Sin embargo, en la literatura, es común encontrar el término Data Analytics referido específicamente a análisis avanzado de datos, tratando las técnicas de BI como una categoría separada.

Según la consultora Gartner, *“El análisis avanzado (Advanced Analytics) es el examen autónomo o semiautónomo de datos a través de técnicas y herramientas sofisticadas más allá del Business Intelligence (BI) tradicional, con el objetivo de descubrir conocimiento más detallado, hacer recomendaciones y generar predicciones. Las técnicas de análisis avanzado incluyen data/text mining o minería de datos, machine learning o aprendizaje máquina, pattern matching o reconocimiento de patrones, forecasting o predicción, visualización, análisis semántico, análisis de sentimientos, análisis de redes y clusters, estadística multivariante, análisis de gráficos, simulación, procesamiento de eventos complejos y redes neuronales.”*

Los sistemas Big Data han provocado un aumento considerable en la cantidad de información que se puede procesar y extraer información de valor. Este hecho ha producido, a su vez, un

avance sustancial en algunas disciplinas de análisis avanzado como el **análisis predictivo, la minería de datos y el análisis de texto.**

Análisis Predictivo: Es una forma de analítica avanzada que usa datos nuevos e históricos para predecir una actividad, comportamiento o tendencia. El análisis predictivo implica realizar análisis estadísticos y/o aplicar técnicas de machine learning para realizar modelos predictivos que permitan valorar la posibilidad de que un determinado evento ocurra.

Aunque existe una cierta tendencia a pensar que con el uso de Big Data, los modelos predictivos son más precisos, hoy en día existen trabajos que indican que el aumento de datos a procesar no implica la obtención de mejores resultados.

Minería de datos: Es el proceso que intenta descubrir patrones y tendencias en grandes volúmenes de datos. Para conseguir dicho objetivo la minería de datos se apoya en el análisis matemático, estadística, inteligencia artificial, aprendizaje máquina y sistemas de bases de datos. La minería de datos requiere cálculos intensivos; por ello, el uso de plataformas con acceso eficiente a los datos, se hace indispensable para optimizar dichos procesos.

Análisis de texto: El análisis del texto permite obtener información potencialmente valiosa a partir de contenido textual (documentos de texto, correos electrónicos y publicaciones en redes sociales como Facebook, Twitter y LinkedIn o foros) utilizando técnicas de procesado de lenguaje natural.

El principal problema que existe al realizar procesamiento del lenguaje natural son las inconsistencias en el mensaje, principalmente los mensajes de foros o redes sociales. Existen múltiples matices en el lenguaje natural que pueden ser fácilmente reconocidos por un lector o interlocutor, sin embargo no es tan sencillo diseñar un algoritmo capaz de reconocer ambigüedades del lenguaje como bromas, sarcasmos, ironía o argots que pueden ser comúnmente utilizados en el lenguaje natural.

Muchas de las disciplinas de análisis avanzado de datos se basan en la aplicación de técnicas de análisis estadístico, machine learning o deep learning.

Análisis estadístico: Es el análisis que emplea técnicas estadísticas para interpretar datos, ya sea para ayudar en la toma de decisiones o para explicar los condicionantes que determinan la ocurrencia de algún fenómeno. El análisis estadístico, y muy particularmente el análisis multivariante, es un conjunto de técnicas estadísticas que permiten detectar patrones de comportamiento ocultos y, basándose en los mismos, establecer predicciones e identificar tendencias.

Machine learning: El aprendizaje máquina o aprendizaje automático es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Estas técnicas son capaces de sacar conclusiones y predecir comportamientos futuros a partir de grandes cantidades de datos actuales o históricos cuyo análisis sería inabordable por una persona.

Deep learning: Los algoritmos de Deep Learning son una clase de algoritmos de machine learning basados en la aplicación de una cascada de capas no lineales para la extracción y transformación de características, de tal forma que cada capa usa la salida de la capa anterior como entrada. Estas técnicas se basan en el aprendizaje de múltiples niveles de representación de los datos que se corresponden con diferentes niveles de abstracción.

Varias arquitecturas de aprendizaje profundo, como redes neuronales profundas, redes neuronales profundas convolucionales, y redes de creencia profundas, han sido aplicadas con éxito en campos como visión por computador, procesamiento de lenguaje natural o reconocimiento de audio.

1.2.3 Análisis avanzado de datos

Para comprender el significado de la analítica avanzada de datos debemos partir del entendimiento del procedimiento para la **extracción e interpretación de los datos**. Como se observa en la siguiente imagen, un enfoque clásico para enfrentarse a un problema de extracción de conocimiento puede dividirse en tres pasos principalmente (Entrada de información, Análisis de datos y salida de información).

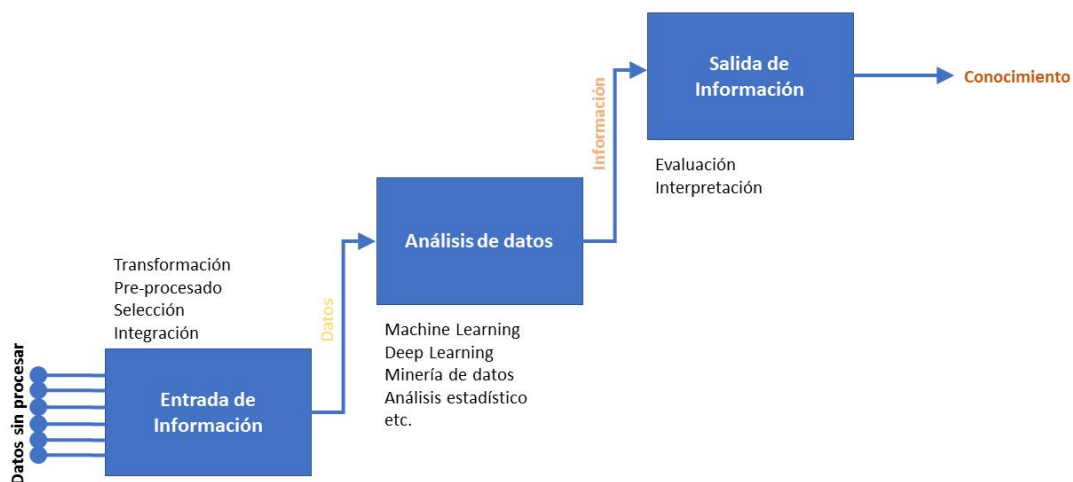


ILUSTRACIÓN 9: UN ENFOQUE CLÁSICO PARA ENFRENTARSE A UN PROBLEMA DE EXTRACCIÓN DE CONOCIMIENTO. FUENTE: ELABORACIÓN PROPIA.

Cada uno de estos pasos tienen sus particularidades dependiendo del problema al que nos enfrentemos, por ejemplo, en el reconocimiento de patrones a través de imágenes el preprocesado y extracción de características cobra mucha importancia por el contrario inferir patrones de consumo energético no necesita tanta preparación del conjunto de datos.

El proceso de análisis avanzado de datos comienza con la integración de datos de una o diversas fuentes. Es aquí donde los analistas de datos identifican y tratan la información necesaria para una aplicación particular. Es posible que sea necesario combinar datos de diferentes orígenes y formatos, transformarlos y almacenarlos.

Una vez se han recopilado los datos, el siguiente paso es realizar un **preprocesado** de los mismos, esto es, eliminar duplicidades, errores, datos de mala calidad u outliers que puedan desvirtuar el análisis, adecuando todos los datos para asegurar un conjunto coherente.

Una vez el conjunto de datos se considera que está completo comienza la fase de análisis de datos donde se aplican técnicas de **machine learning o deep learning**, etc. Para ello se construye

un modelo analítico utilizando herramientas de análisis predictivo u otro software analítico y este modelo se ejecuta inicialmente contra un subconjunto de los datos que sea representativo de todo el conjunto. En este conjunto se prueba y se adapta varias veces el modelo hasta que se obtengan los resultados esperados.

Finalmente, el modelo se ejecuta sobre el conjunto de datos completo **una única vez** para responder a una necesidad de información concreta, o bien de forma continua a medida que se actualizan los datos. En caso de hacerlo de manera continua lo más normal es configurar la tarea de análisis de forma automática para que se ejecute con una determinada frecuencia.

Finalmente, el último paso del análisis será **comunicar los resultados obtenidos** por los modelos analíticos a los decisores y a otros usuarios finales para la ayuda en la toma de decisiones (Evaluación e interpretación). Sin embargo, lo ideal es que esta tarea no se realice de forma manual sino que los resultados generados se almacenen de forma que puedan ser leídos por otras aplicaciones que generen uno o varios paneles de control. Estos paneles de control ofrecen en una sola pantalla gráficos y otras representaciones de los datos diseñados para entender fácilmente los mismos obteniendo conocimiento y valor que permitan ayudar en el proceso de toma de decisión.

ENTRADA DE DATOS

Una vez recogido los datos sin procesar de las diferentes fuentes de información o sensores que van a ser utilizados como conjunto de datos para el análisis avanzado es necesario realizar **diferentes transformaciones** para conseguir que dicho conjunto sea significativo al problema que queremos modelar. Por tanto, el propósito fundamental de la preparación de los datos es manipular y transformar los datos en crudo (del inglés raw data, sin procesar) para fácilmente poder inferir los patrones que contienen.

Una de las primeras técnicas a implementar cuando nos enfrentamos a un problema es el **preprocesado** de los datos. Estas técnicas son muy diversas y dependen tanto del problema como de la fuente de datos. Algunas de las técnicas de preprocesado que se pueden realizar son la recuperación información incompleta, eliminación de outliers (valores atípicos), resolución de conflictos, etc. Estas técnicas generan datos de mayor calidad de los cuales se podrán obtener patrones/reglas de mayor calidad.

También como técnica de preprocesado puede ser considerada la recolección e integración (data collecting and integration) de datos de diferentes fuentes de datos que permiten crear conjuntos más homogéneos resolviendo los problemas de representación y codificación. Data cleaning engloba todas aquellas técnicas que permiten resolver conflictos entre los datos, eliminar valores atípicos y resuelve problemas de ruido y valores perdidos. Por último existen técnicas para transformar los datos que realizando operaciones de agregación y sumariación de los datos permiten obtener un conjunto de datos más cómodo para las siguientes fases del aprendizaje.

Por otro lado tenemos las técnicas de extracción de características como por ejemplo la **reducción de la dimensionalidad** del problema. La reducción de la dimensionalidad es el proceso de reducir el número de variables o parámetros que se consideran en el análisis. En todos los problemas de análisis avanzado de datos se aplican técnicas para reducir la dimensionalidad con

el objetivo de simplificar el problema y reducir con ello el tiempo de computación, al mismo tiempo que se incrementa la precisión del método. La reducción de la dimensionalidad puede dividirse en:

- **Selección de características:** Los métodos de selección de características tratan de encontrar un subconjunto de las variables originales.
- **Extracción de características:** Las técnicas de extracción de características transforman los datos en un espacio de alta dimensión a otro espacio con menos dimensiones.
- **Discretización:** Es el proceso mediante el cual los valores se incluyen en almacenes de datos para que haya un número limitado de estados posibles.

En el caso del análisis de grandes volúmenes de datos, aplicar técnicas de reducción de dimensionalidad es casi obligatorio puesto que la cantidad de variables a estudiar puede ser extremadamente alto y algunas de ellas puede que no aporten información relevante al problema. Se presentan a continuación un conjunto de técnicas de reducción de dimensionalidad aceptadas en un entorno Big Data [15]:

- **Missing Values Ratio:** Método de selección de características consistente en eliminar aquellos parámetros para los cuales faltan muchos valores, esto es, eliminar muestras que no presentan valor para un determinado parámetro.
- **Low Variance Filter o Maximum relevance:** Método de selección de características donde los parámetros con pocos cambios en los datos se consideran que aportan poca información y por tanto no son considerados representativos del conjunto.
- **High Correlation Filter o Minimum redundancy:** Método de selección de características donde los parámetros que presentan alta correlación entre ellos se supone que llevan información muy similar por lo tanto no sería necesario incluirlos en proceso de análisis, sería suficiente alimentar el modelo con uno de ellos.
- **Random Forests o Ensemble Trees:** Los árboles de decisión además de ser efectivos como clasificadores, han demostrado ser útiles como selector de características en conjuntos de datos con un gran número de dimensiones. Existen varios métodos de selección de características basadas en random forests.
- **Principal Component Analysis (PCA):** Método de extracción de características estadístico que transforma ortogonalmente las n coordenadas originales de un conjunto de datos en un nuevo conjunto de n coordenadas llamadas componentes principales. Como resultado de la transformación obtenemos un conjunto de componentes que están ordenados de acuerdo a la varianza y son ortogonales, es decir, no hay correlación entre ellos. Con ésta técnica se reduce la dimensionalidad de los datos conservando la mayor parte de la información

Hay que tener en cuenta que con las técnicas de extracción como PCA, las variables originales se transforman por lo que se pierde la interoperabilidad de los resultados.

- **Backward Feature Elimination.** Esta técnica consiste en entrenar el modelo con todas las características de entrada original y sucesivamente eliminar una característica en cada iteración y entrenar el modelo con las $n-1$ restantes. La característica que produce un incremento mayor del error, se elimina. De esta forma quedarían $n-1$ características en las que aplicaríamos de nuevo el proceso. Cada iteración k produce un modelo entrenado sobre características $n-k$ y una tasa de error $e(k)$. Seleccionando la tasa de error tolerable máxima, definimos el menor número de características necesarias para

alcanzar ese rendimiento de clasificación con el algoritmo de aprendizaje seleccionado.

- **Forward Feature Construction:** Este es el proceso inverso al método anterior, por tanto, se empieza con una característica y se va incrementando progresivamente el número de características añadiendo aquella que produce un aumento en el rendimiento del modelo. Hay que destacar que tanto éste como el método anterior son lentos y computacionalmente caros por lo que es recomendable su aplicación para conjuntos de datos donde el número de características de entrada es muy bajo.

ANÁLISIS DE DATOS

Inferencia estadística

El objetivo principal del análisis estadístico es identificar tendencias en el conjunto de datos a evaluar. Por ejemplo, un negocio minorista podría hacer uso del análisis estadístico para encontrar patrones no estructurados y semi-estructurados a partir de los datos de los clientes que dispone, pudiendo de esta manera, mejorar la experiencia cliente, y esto traducirse en un aumento en las ventas.

El objetivo de la inferencia estadística es mejorar el conocimiento sobre la población a partir de un conjunto representativo de miembros (muestra). Los métodos principales de inferencia paramétrica son: estimación de los parámetros de la población y contrastes de hipótesis. Ambos métodos se basan en el conocimiento de la distribución de probabilidad de un estadístico muestral que se utiliza como estimador de los parámetros poblacionales a inferir.

La estimación de los parámetros consiste en inferir los valores de los parámetros o características de la población, desconocidos a partir de los valores de la muestra. Esta estimación está sujeta a un error por ello se construye un intervalo de confianza, esto es, un rango de valores al que pertenece el parámetro poblacional con un determinado valor de confianza o probabilidad.

Por otro lado, los métodos de contraste de hipótesis tienen como objetivo comprobar si un determinado supuesto referido a un parámetro poblacional, es compatible con la evidencia empírica contenida en la muestra.

Antes de la realización de cualquier proceso de inferencia estadística paramétrica es necesario conocer la distribución de los datos, para ello se realiza una exploración de los mismos de la que se obtendrá una serie de información como:

- Tablas de frecuencia.
- Gráficos e histogramas.
- Estadísticos resumen:
 - Medidas de posición: media, mediana, moda, cuantiles (cuartiles, deciles, percentiles...).
 - Medidas de variabilidad: varianza, desviación típica, rango intercuartil.
 - Medidas de apuntamiento y curtosis.

Métodos lineales para la regresión

En estadística la **regresión lineal o ajuste lineal** es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y , las variables independientes X_i y un término aleatorio ε . Este modelo puede ser expresado como

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

donde:

- Y_t : variable dependiente, explicada o regresando.
- X_1, X_2, \dots, X_p : variables explicativas, independientes o regresores.
- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$: parámetros, miden la influencia que las variables explicativas tienen sobre la variable dependiente. β_0 es la intersección o término "constante", las β_i ($i > 0$) son los parámetros respectivos a cada variable independiente, y p es el número de parámetros independientes a tener en cuenta en la regresión.
- ε : es un término de error de media cero que refleja la perturbación aleatoria y recoge todos aquellos factores de la realidad no controlables u observables y que por tanto se asocian con el azar, y es la que confiere al modelo su carácter estocástico.

La primera forma de regresión lineal documentada fue el método de los mínimos cuadrados que fue publicada por Legendre en 1805, no obstante, Gauss publicó un trabajo en donde desarrollaba de manera más detallada el método de los mínimos cuadrados [16] y en donde se incluía una versión del teorema de Gauss-Márkov.

El término **regresión** se utilizó por primera vez en un estudio que comparaba la estatura de padres e hijos, donde resultó que los hijos cuyos padres tenían una estatura muy superior al valor medio, tendían a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendían a reducir su diferencia respecto a la estatura media; es decir, "regresaban" al promedio. La constatación empírica de esta propiedad se vio reforzada más tarde con la justificación teórica de ese fenómeno.

El término **lineal** se emplea para distinguirlo del resto de técnicas de regresión, que emplean modelos basados en cualquier clase de función matemática. Los modelos lineales son una explicación simplificada de la realidad, mucho más ágiles y con un soporte teórico mucho más extenso por parte de la matemática y la estadística.

Para poder crear un modelo de regresión lineal es necesario que la relación entre las variables sea lineal, los errores en la medición de las variables explicativas sean independientes entre sí, los errores dispongan varianza constante (Homocedasticidad), los errores tengan una esperanza matemática igual a cero (los errores de una misma magnitud y distinto signo no son equiprobables) o el error total sea la suma de todos los errores.

Vecinos más próximos

Los métodos de vecinos más próximos nacen como una forma de búsqueda por proximidad. Consiste en el problema de optimización de encontrar un punto de un determinado conjunto que sea el más cercano (o el más similar) a otro punto dado. La distancia, en general, puede ser cualquier medida métrica, aunque la distancia Euclídea es la elección más común. La proximidad se expresa habitualmente en términos de una función de disimilitud, es decir, cuanto menos parecidos sean ambos objetos, más alto será el valor devuelto por la función. Por tanto, la optimización de este problema consistirá en minimizar esta función de disimilitud.

Formalmente, el problema de la búsqueda del vecino más cercano se define de la siguiente forma: dado un conjunto de puntos S en un espacio M y un punto q perteneciente a M , encontrar el punto más cercano en S a q . Este problema fue formulado por Donald Knuth en el libro *"The Art of Computer Programming"* [17], donde escribió sobre el Problema de la Oficina Postal, refiriéndose a la asignación de la oficina postal más cercana a cada residencia.

El algoritmo más conocido para el atajo de este problema es la búsqueda del **kNN (k-Nearest Neighbor, K-Vecinos más Cercanos)**. Este algoritmo identifica un número K de vecinos más cercanos al consultado. La técnica se usa habitualmente en Análisis Predictivo para la estimación o clasificación de un punto basado en el contexto de sus vecinos. Los gráficos de kNN muestran cada punto conectado a sus K vecinos más cercanos. Se puede utilizar tanto para clasificación como para regresión.

Además del kNN, también existen otras variantes:

- **Vecino más cercano aproximado:** En algunas ocasiones puede ser aceptable una "buena suposición" del vecino más cercano. Se puede usar un algoritmo que no garantice la devolución del vecino más cercano real en todos los casos.
- **Vecino más cercano con distancia ponderada:** En este caso se pondera la contribución de cada vecino según la distancia entre él y el ejemplar a ser clasificado, dando un mayor peso a los vecinos más cercanos.
- **Vecinos más cercanos por ratio fijo:** En el caso de necesitar encontrar eficientemente todos los puntos dados en un espacio euclídeo dentro de una distancia fija desde un punto específico. La estructura de datos trabajará en una distancia fija de que el punto a buscar sea aleatorio.
- **Todos los vecinos más cercanos:** En algunas ocasiones, se tienen N puntos de datos y se desea conocer cuál es el vecino más cercano para cada uno de los N puntos. Esto se podría conseguir ejecutando el kNN una vez por cada punto, sin embargo, este algoritmo tiene en cuenta la información redundante entre estas diferencias para producir una búsqueda más eficiente.

Árboles de decisión

Los árboles de decisión son una herramienta de apoyo a la decisión que utiliza una estructura semejante a un árbol o un modelo de decisiones y sus posibles consecuencias, incluyendo los resultados de posibles eventos, los costes de recursos y su utilidad. A diferencia del resto de métodos que necesitan un alto componente de conocimiento para comprender el resultado este tipo de técnicas son muy visuales y por tanto guían al lector en el proceso mental de decisión seguido por el algoritmo.

Los árboles de decisión se usan habitualmente en operaciones de investigación, concretamente en el análisis de decisiones, para ayudar a identificar la estrategia con mayor probabilidad de alcanzar una meta, pero también son una herramienta popular en el aprendizaje automático.

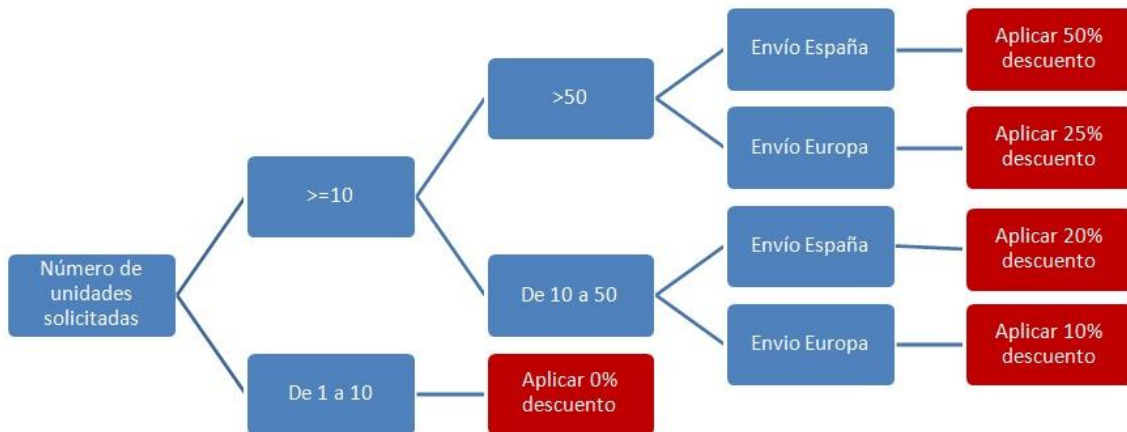


ILUSTRACIÓN 10: ÁRBOL DE DECISIÓN: EJEMPLO PARA OBTENCIÓN DE DESCUENTOS SEGÚN VOLÚMENES DE PEDIDO. FUENTE: WIKIPEDIA

En muchas ocasiones se aplican los árboles de decisión para el aprendizaje, utilizándolos como un modelo predictivo que mapea ítems con sus resultados sobre su valor objetivo final. Dentro de este tipo de árboles, se denominan Árboles de Clasificación si su variable de destino puede resultar en un conjunto finito de valores. Si la variable de destino puede resultar en valores continuos se denominan Árboles de Regresión.

La construcción de estos árboles se realiza a partir de **tuplas de entrenamiento**, donde cada una de ellas va etiquetada con su correspondiente clase. Estos árboles de decisión se forman de una forma, en cierto modo, similar a un diagrama de flujo. Cada nodo interno (no hoja) se corresponde con una prueba en un atributo, cada una de las ramas denota el resultado de una prueba, y finalmente cada nodo terminal (nodo hoja) tendría una etiqueta de clase.

Algunos de los algoritmos específicos más importantes para árboles de decisión son:

- **ID3:** Dado un conjunto de ejemplos, su uso se engloba en la búsqueda de hipótesis. Consta de nodos, arcos y hojas. La elección del mejor atributo se realiza mediante la entropía, eligiendo aquel que proporciona una mejor ganancia de información.
- **C4.5:** Se trata de una evolución realizada a partir del algoritmo ID3, especialmente utilizado para clasificación. Algunas de las mejoras son el manejo de atributos con costos diferentes, manejo de datos con valor faltante o, podado de árboles después de su creación.
- **MARS:** El algoritmo MARS (Multivariate Adaptive Regression Splines) es una forma de análisis regresivo introducida por J. H. Friedman en 1991. Se trata de una forma de regresión no paramétrica y puede ser vista como una extensión de modelos lineales. Por norma general son más flexibles que los modelos de regresión lineal, además de ser más simples de entender e interpretar.

Redes de neuronas artificiales

Las redes neuronales surgieron como una herramienta de aprendizaje automático para la clasificación, mejorando sustancialmente métodos de clasificación convencionales. La primera

definición formal de red neuronal fue dada en 1943 por McCulloch y Pitts como una máquina binaria con varias entradas y salidas [18].

La principal ventaja que aportan las redes neuronales reside en que son **métodos autoadaptativos** impulsados por datos que pueden ajustarse sin ninguna especificación del modelo subyacente. Son considerados aproximadores funcionales universales ya que las redes neurales pueden aproximar cualquier función con exactitud. Las redes neuronales generan modelos no lineales, lo que permite aproximaciones más precisas de las relaciones al mundo real.

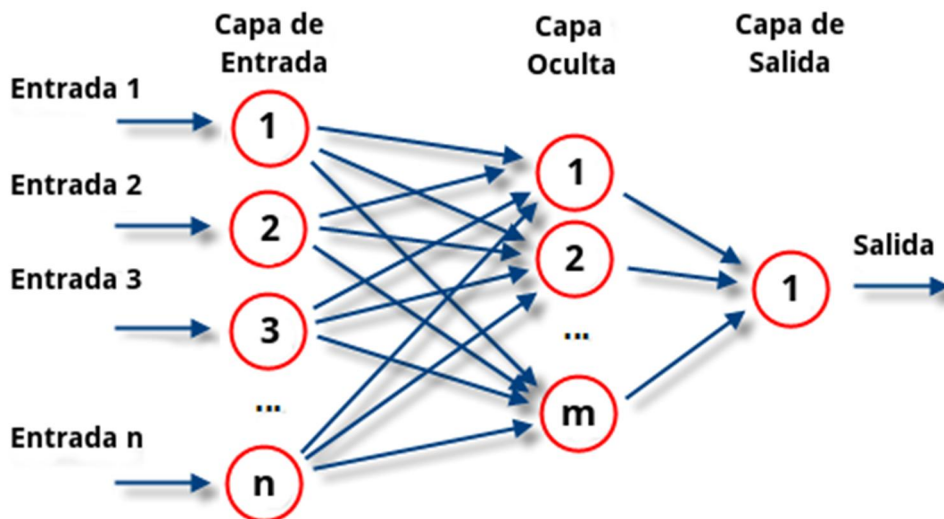


ILUSTRACIÓN 11: EJEMPLO DE UNA RED DE NEURONAL MULTICAPA (PERCEPTRÓN MULTICAPA). FUENTE: WIKIPEDIA

La eficacia como **algoritmo de clasificación** ha sido probada en numerosos problemas tanto en la industria, el comercio y la ciencia. Por ejemplo en aplicaciones de predicción de impagos, reconocimiento de caracteres, reconocimiento del habla, diagnóstico médico, etc.

Las redes neuronales suelen consistir en una o varias capas donde la información se propaga desde las capas de entrada hacia las de salida a través de las neuronas que componen la capa o capas ocultas.

Todas las neuronas disponen de una función de propagación, la cual normalmente es la suma ponderada de las entradas multiplicadas por los pesos. En esta función se interpreta como un regulador de las señales que se emiten entre neuronas al ponderar las salidas que entran a la neurona. Las funciones de activación más comunes son:

- **Lineal:** Algunas redes neuronales usan esta función de activación como el Adeline por su eficiencia y facilidad.
- **Escalón:** Esta función es la más usada para redes neuronales binarias ya que no es lineal y es muy simple. Algunas redes que usan esta función son el Perceptrón y Hopfield.
- **Hiperbólicas o tangenciales:** Las redes con salidas continuas, como el Perceptrón multicapa con retropropagación, usan esta función ya que su algoritmo de aprendizaje necesita una función derivable.

La modelización de las redes neuronales consta de una fase de entrenamiento donde se utiliza un subconjunto de datos o patrones de entrenamiento para determinar los pesos de cada una de las neuronas que componen la red. De manera iterativa se calculan los pesos con el objetivo de minimizar el error cometido entre la salida obtenida por la red neuronal y la salida esperada. Posteriormente se pasa a la fase de prueba donde el modelo generado se prueba con otro conjunto de datos para determinar si este se ha ajustado demasiado a las particularidades de los datos de entrenamiento.

Máquinas de soporte vectorial

Las máquinas de soporte vectorial (**SVM, del inglés Support Vector Machine**) son algoritmos de aprendizaje máquina que principalmente se utilizan para resolver problemas de clasificación o regresión. Son especialmente útiles en casos donde la separación entre clases no es lineal (no existe una línea recta virtual que separe los elementos de las distintas clases). El método de aprendizaje de las SVM consiste en determinar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para conseguir un margen máximo a cada lado del hiperplano. Los ejemplos de cada clase que caen en la frontera y que definen el hiperplano se denominan vectores soporte.

Si bien las redes neuronales también permiten resolver problemas de clasificación no lineal, el conjunto de vectores resultado al entrenar con SVM resulta más fácil de observar y comprender que una red neuronal entrenada, puesto que el número de parámetros necesarios normalmente es sensiblemente menor en el caso de las SVM. Una gran desventaja de los SVMs es su complejidad computacional, que crece de forma no lineal con el tamaño del conjunto de datos utilizado para el entrenamiento. Para afrontar esta desventaja se puede hacer uso de algoritmos como MapReduce y dividir el entrenamiento de la SVM entre los distintos nodos de un clúster Hadoop combinando finalmente los vectores soporte resultantes en cada nodo.

Ensembles: boosting y bagging

Los métodos ensemble se describen como algoritmos de aprendizaje estadístico que alcanzan un mejor rendimiento predictivo construyendo conjuntos de clasificadores, generalmente pequeños árboles de decisión, y luego clasificando nuevos puntos de datos tomando un voto ponderado de las predicciones.

Este mismo concepto se aplica al **machine learning**, esto es, un algoritmo de predicción puede obtener mejores resultados mediante la combinación de varios modelos simples. La evaluación de la predicción producida por un ensemble tiene un mayor coste computacional que la predicción realizada por un modelo simple, por lo que en cierto modo podemos pensar en los Ensembles como una manera de compensar la baja tasa de aprendizaje de los algoritmos realizando una gran cantidad de computación extra. Existen diversos métodos de Ensemble, de los cuales destacan:

- **Bagging:** método comúnmente utilizado en la práctica, que implica tener cada modelo en el ensemble calificado/votado con el mismo peso. Con el fin de promover la varianza del modelo, el bagging entrena a cada modelo en el conjunto usando un subconjunto elegido al azar del conjunto de entrenamiento. El método bagging también reduce la

varianza y ayuda a evitar el sobreentrenamiento. Como ejemplo, el algoritmo Random Forest combina colecciones de árboles de decisión aleatorios con bagging para lograr una precisión de clasificación muy alta.

- **Boosting:** implica la construcción incremental de un ensemble entrenando cada nueva instancia del modelo para enfatizar las instancias de entrenamiento que los modelos anteriores hayan clasificado mal. El boosting comienza ajustando un clasificador simple en los datos originales, dando a cada observación el mismo peso. Después de ajustar el primer modelo, se calculan los errores de predicción y se vuelven a ponderar los datos para dar a ejemplos previamente mal clasificados un peso mayor que los ejemplos correctamente clasificados. Las observaciones son incrementalmente re-ponderadas en cada paso para dar mayor peso a las observaciones que el modelo anterior clasificó erróneamente. Mientras que el bagging trabaja reduciendo la varianza, el boosting trabaja refinando incrementalmente el límite de la decisión del clasificador agregado (aunque bagging también mejora el límite de la decisión, boosting lo hace de forma más directa).

En algunos casos, se ha demostrado que el boosting proporciona una mejor exactitud predictiva que el bagging, pero también tiende a ser más probable que se produzca sobreentrenamiento. La implementación más común de boosting es el algoritmo AdaBoost (algoritmo adaptativo de refuerzo).

- **Blending:** es la manera más simple e intuitiva de combinar diferentes modelos. Implica tomar las predicciones de varios modelos compuestos e incluirlos en un modelo más grande, tal como una segunda etapa de regresión lineal o logística. Blending puede utilizarse en cualquier tipo de modelos compuestos, pero es más apropiado cuando los modelos compuestos son menos y más complejos que en el boosting o bagging. Las propiedades de blending dependen de cómo se combinen los modelos.

Selección del modelo de análisis

La elección del método de análisis de datos a aplicar para un determinado problema no es directa. Konoshi et al. [19] afirman que "La mayoría de los problemas de inferencia estadística pueden ser considerados problemas relacionados con el modelo estadístico", mientras Sir David Cox [20] apunta "La traducción del problema al modelo estadístico es a menudo la parte más crítica del proceso de análisis". Ambas afirmaciones reflejan la importancia y la dificultad de seleccionar el algoritmo de análisis más adecuado para resolver el problema dentro de un conjunto de candidatos.

Una vez elegido el conjunto de modelos o algoritmos candidatos, el análisis estadístico nos permite seleccionar el mejor de estos modelos. Sin embargo, lo que se entiende por mejor es controvertido; una buena técnica de selección de modelos equilibrará la bondad del ajuste con la sencillez. Los modelos más complejos serán más capaces de adaptar su forma a los datos (por ejemplo, un polinomio de quinto orden puede ajustarse exactamente a seis puntos), pero los parámetros adicionales no representarán nada útil. Así, dados varios candidatos con una capacidad de explicación y predicción similares, a menudo el modelo más simple será la mejor opción.

Existen **múltiples medidas estadísticas** que se pueden calcular para determinar la precisión y el ajuste de un modelo a los datos. Dependiendo del tipo de problema a resolver se seleccionarán unas u otras. En la mayoría de los problemas se utiliza el error cuadrático medio² (MSE, del inglés Mean Squared Error), el promedio de los errores al cuadrado, esto es, la diferencia entre el resultado esperado y el resultado dado por el modelo. Sin embargo, dependiendo de la naturaleza del problema, se utilizan también otras medidas estadísticas más apropiadas. Por ejemplo, para problemas de clasificación o segmentación binaria se utiliza: la sensibilidad y especificidad³, precisión y recall⁴ o F1 score o F-score⁵.

El método más simple para seleccionar el modelo es el **método de retención o holdout** que consiste en dividir el conjunto de datos en dos conjuntos predefinidos: entrenamiento y test. Los algoritmos candidatos se entrenan sobre el mismo conjunto, el conjunto de entrenamiento, y se prueban sobre el conjunto de test calculando la medida estadística correspondiente, es decir, el valor cuantitativo que nos permitirá comparar los distintos algoritmos y seleccionar el más preciso. Sin embargo, este método no es el más adecuado puesto que es altamente dependiente del conjunto de datos. Por ejemplo, puede ocurrir que exista un modelo mejor que otro en el conjunto original completo pero en el conjunto de test se comporte peor. Para solucionar este problema, han surgido distintos métodos. El más utilizado de ellos es el conocido como **cross-validation** o validación cruzada

Cross-validation consiste en dividir el conjunto de datos original en múltiples particiones en las que se aplica el método de retención (entrenamiento y test en cada partición), y calcular la media aritmética de los resultados en todas las particiones. De esta forma, la decisión no estará desvirtuada por el conjunto de test utilizado y además obtenemos información adicional de cómo se comporta el modelo en distintos conjuntos, es decir, de su variabilidad. Existen otras variantes del método de cross-validation:

- **k-fold cross-validation:** el conjunto de datos original es dividido aleatoriamente en k subconjuntos de igual tamaño. Uno de los subconjuntos se utiliza como conjunto de test para validar el modelo y el resto (k-1) constituyen el conjunto de entrenamiento. Este proceso de validación se repite k veces. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado que determine la precisión del modelo.
- **Leave-one-out:** consiste en repetir el proceso de validación tantas veces como observaciones existan en el conjunto de entrada. En cada iteración, se selecciona una única observación para validar el modelo y el resto constituirá el conjunto de entrenamiento. De esta forma, todas las observaciones del conjunto de entrada formarán parte del conjunto de test pero una única vez. La validación leave-one-out es equivalente a k-fold cross-validation cuando k es igual al número de observaciones del conjunto de entrada.

² https://es.wikipedia.org/wiki/Error_cuadr%C3%A1tico_medio

³ [https://es.wikipedia.org/wiki/Sensibilidad_y_especificidad_\(estad%C3%ADstica\)](https://es.wikipedia.org/wiki/Sensibilidad_y_especificidad_(estad%C3%ADstica))

⁴ https://es.wikipedia.org/wiki/Precisi%C3%B3n_y_exhaustividad

⁵ <https://es.wikipedia.org/wiki/Valor-F>

Data analytics: Frameworks

Son muchos los frameworks y bibliotecas disponibles actualmente en el mercado que ofrecen analítica de datos y aprendizaje automático en general. A continuación presentamos una lista que recoge los más destacados en este ámbito, incluyendo en la misma tanto aquellos productos orientados desde su nacimiento al ámbito del Big Data, como aquellos más tradicionales pero que de alguna forma pueden ser escalables a este tipo de entornos.

- **Apache Singa⁶**: Se trata de una plataforma distribuida de Deep Learning sobre grandes conjuntos de datos. Está diseñado con un modelo de programación intuitivo basado en la abstracción de capas. Soporta una amplia variedad de modelos de Deep Learning como son las redes de convolución (CNN), modelos de energía como Restricted Boltzman Machine (RBM), y redes recurrentes (RNN).
- **Apache Mahout⁷**: Este proyecto ha nacido con la intención de construir un entorno para facilitar la creación de aplicaciones de aprendizaje máquina escalables basadas fundamentalmente en clustering, clasificación y filtrado colaborativo. (técnica usada por sistemas recomendadores). Apache Mahout ofrece tres características principales:
 - Un entorno de programación simple y extensible y un framework para la construcción de algoritmos escalables.
 - Una amplia variedad de algoritmos predefinidos para Scala + Apache Spark, H2O y Apache Flink.
 - Samsara, un entorno de experimentación matemática vectorial con una sintaxis similar al lenguaje del entorno R⁸ y que funciona a escala.
- **Amazon Machine Learning⁹**: Es un servicio pensado para facilitar a los desarrolladores el uso de la tecnología de aprendizaje automático. Amazon Machine Learning provee herramientas de visualización y asistentes para guiar al desarrollador durante el proceso de creación de los modelos Machine Learning. Utiliza datos almacenados en Amazon S3, Redshift, o RDS, y puede ejecutar clasificación binaria, multiclase o regresión.
- **Azure Machine Learning Studio¹⁰**: Machine Learning Studio permite a los usuarios de Microsoft Azure probar e implementar soluciones de análisis predictivo, para posteriormente publicarlos como servicios web y poder utilizarlo en otras aplicaciones o herramientas de BI como por ejemplo Excel.
- **Caffe¹¹**: Es un framework de Deep Learning desarrollado por el Berkeley Vision and Learning Center y por los contribuidores de la comunidad. Caffe ha sido liberado bajo la licencia BSD 2-Clause. Una de sus principales ventajas es su sencillez puesto que los modelos se definen mediante configuración, esto es, sin necesidad de codificación. Además ofrece al usuario la posibilidad de elegir de forma sencilla dónde realizar el procesamiento, en la CPU o la GPU. Otra ventaja de Caffe es la velocidad de procesamiento, puede procesar más de 60 millones de imágenes por día con una sola

⁶ Apache Singa, <http://singa.apache.org/docs/overview.html>

⁷ Apache Mahout, <http://mahout.apache.org/>

⁸ R project, <https://www.r-project.org/>

⁹ Amazon Machine Learning, <https://aws.amazon.com/machine-learning/>

¹⁰ Azure ML Studio, <https://studio.azureml.net/>

¹¹ Caffe, <http://caffe.berkeleyvision.org/>

GPU Nvidia K40.

- **H2O**¹²: Es una plataforma open source de Inteligencia Artificial. Permite que cualquier usuario pueda fácilmente aplicar algoritmos de machine learning y análisis predictivo. Ofrece una interfaz web sencilla y entornos de desarrollo para los lenguajes de programación R, Python, Java, Scala y JSON. Permite controlar el entrenamiento de los modelos de manera visual. H2O está escrito en Java y se integra perfectamente con Apache Hadoop y Spark. Ofrece soporte también para diferentes fuentes de datos como HDFS, S3 y SQL y bases de datos NoSQL.
- **Massive Online Analysis (MOA)**¹³: Es el framework open source más popular para la minería de datos con una comunidad muy activa y en crecimiento. Incluye múltiples algoritmos de machine learning como: clasificación, regresión, clustering, detección de valores atípicos, análisis predictivo y sistemas de recomendación. Está relacionado con el proyecto Weka¹⁴, una conocida librería de aprendizaje máquina escrita en Java. Al igual que Weka, MOA también está escrito en Java.
- **MLlib (Apache Spark)**¹⁵: MLIB es la biblioteca de machine learning escalable de Apache Spark. Incluye algoritmos de clasificación (Logistic Regression, Naive bayes,...), regresión, clustering (K-means, GMM,...), árboles de decisión, random forests, y técnicas de reducción de dimensionalidad (PCA, SVD,..). Ofrece la posibilidad de construir y guardar pipelines, es decir, cadena de fases de procesamiento para definir el flujo de trabajo. Se integra con NumPy¹⁶, una librería de computación científica de Python y con las librerías del entorno R¹⁷. Es compatible con cualquier fuente de datos de Hadoop (HDFS, HBase,...).
- **Mlpack**¹⁸: Es una biblioteca de machine learning basada en C++ que vio la luz en 2011 con un diseño orientado a la escalabilidad, velocidad y facilidad de uso. Mlpack ofrece la posibilidad de ejecutar los algoritmos mediante programas de línea de comandos o a través de una API en C++ que puede integrarse con otras soluciones de aprendizaje máquina a gran escala.
- **Pattern**¹⁹: Módulo de minería web para el lenguaje de programación Python. Proporciona herramientas de minería de datos sobre distintas fuentes (Google, API's de Twitter y Wikipedia, web crawler y HTML DOM parser), procesamiento de lenguaje natural (part-of-speech taggers, búsqueda de n-grams, análisis de sentimientos, WordNet), machine learning (clustering, SVM), análisis de red y visualización.
- **Scikit-Learn**²⁰: Es una librería open source de Python para minería y análisis de datos construidos sobre las librerías NumPy, SciPy y Matplotlib. Ofrece soporte para

¹² H2O, <http://www.h2o.ai/>

¹³ Massive Online Analysis, <http://moa.cms.waikato.ac.nz/>

¹⁴ Weka, <http://www.cs.waikato.ac.nz/ml/weka/>

¹⁵ MLlib, <http://spark.apache.org/mllib/>

¹⁶ NumPy, <http://www.numpy.org/>

¹⁷ R project, <https://www.r-project.org/>

¹⁸ Mlpack, <http://mlpack.org/>

¹⁹ Pattern, <http://www.clips.ua.ac.be/pattern>

²⁰ Scikit-Learn, <http://scikit-learn.org/stable/>

clasificación (SVM, nearest neighbors, random forest,..), regresión, clustering (k-means, mean-shift,...), reducción de dimensionalidad, selección de modelos (cross-validation, grid search,..) y preprocesado.

- **Shogun**²¹: Es una de las bibliotecas más antiguas de machine learning, creada en 1999 y escrita en C++. Sin embargo, no está limitada a C++; gracias a la librería SWIG, se puede utilizar en otros lenguajes como Java, Python, C#, Ruby, R, Lua, Octave y Matlab.
- **TensorFlow**²²: Se trata de una librería open source para la computación numérica mediante grafos de flujo de datos. Los nodos de los grafos representan operaciones matemáticas, mientras que las aristas representan los arrays de datos multidimensionales (tensores). Su arquitectura flexible permite ejecutar los modelos en distintos dispositivos: una o más CPUs o GPUs en un escritorio, servidor o dispositivo móvil. TensorFlow fue desarrollado originalmente por investigadores e ingenieros de Google con el objetivo de investigar en técnicas de machine learning y deep learning con redes neuronales, sin embargo la librería es lo suficientemente genérica para ser aplicada en otros dominios.
- **Theano**²³: Theano es una biblioteca de Python que permite definir, optimizar y evaluar expresiones matemáticas, especialmente aquellas con arrays multi-dimensionales. Esta biblioteca es capaz de alcanzar velocidades que rivalizan con implementaciones realizadas en C pensadas para problemas que implican datos a gran escala.
- **Torch**²⁴: Framework de computación científica con soporte para algoritmos de machine learning mediante GPU. Es de uso fácil y eficiente gracias al sencillo lenguaje LuaJIT y a la implementación subyacente en CUDA C. El objetivo de Torch es ofrecer la máxima flexibilidad y velocidad para construir algoritmos científicos de forma muy simple. Además existe una amplia gama de paquetes desarrollados por la comunidad de machine learning, visión artificial, procesamiento paralelo, procesamiento de señales entre otros.
- **Veles**²⁵: Plataforma distribuida para aplicaciones de deep learning escrita en C++ aunque utiliza Python para la coordinación entre nodos. Los conjuntos de datos pueden ser analizados y normalizados automáticamente antes de ser enviados al clúster. Además proporciona una API REST que permite que el modelo entrenado sea utilizable inmediatamente en producción. Se centra en el rendimiento y la flexibilidad, y ofrece soporte para todas las topologías de red ampliamente conocidas como redes totalmente conectadas, redes convolucionales, redes recurrentes, etc. Veles incluye Mastodon, un subproyecto que permite la integración con cualquier aplicación Java. Gracias a este puente se facilita la integración de Veles con Hadoop.
- **Keras**²⁶: Las librerías Theano y TensorFlow, mencionadas anteriormente, son las más utilizadas para implementar sistemas deep learning en Python. Sin embargo, su

²¹ Shogun, <http://www.shogun-toolbox.org/>

²² TensorFlow, <https://www.tensorflow.org/>

²³ Theano, <http://deeplearning.net/software/theano/>

²⁴ Torch, <http://torch.ch/>

²⁵ Veles, <https://velesnet.ml/>

²⁶ Keras, <https://keras.io/>

utilización directa puede resultar complicada. Para solventar este problema nace Keras, una biblioteca de Python que se construye sobre Theano y TensorFlow y permite que la implementación de los modelos de aprendizaje profundo sea considerablemente más sencilla. Además ofrece soporte para procesamiento sobre CPU o GPU.

1.2.4 Ventajas y limitaciones

La principal ventaja de **Data Analytics** es la posibilidad de extraer conocimiento oculto de un conjunto de datos y en particular en el caso de Big Data al disponer de grandes volúmenes de datos. Data Analytics permite explotar nuevas ventajas competitivas para las empresas ofreciendo información que antes no era posible procesar. El desarrollo de Data Analytics posibilitará también el desarrollo de la tecnología Business Intelligence necesaria para las empresas hoy en día. Data Analytics proporciona, a día de hoy, la mejor forma de realizar BI en las empresas:

- **Segmentación** del cliente.
- Ahorro de **costes**.
- **Mejoras** en la fabricación.
- Realización de **previsiones** y planificación.
- Cuantificación del **riesgo**.
- Detección del **fraude**.
- Mejora en la identificación del **cliente objetivo**.
- Mejora de la **experiencia del cliente**.
- Ventas **personalizadas**.
- Establecimiento de **precios dinámicos**.

Sin embargo, la aplicación de Data Analytics dentro de la empresa puede presentar **dificultades** principalmente centradas en:

- Inadecuada formación o **falta de habilidades**.
- **Costes** en la implantación de los nuevos sistemas.
- Inversión elevada para adecuar la infraestructura necesaria para ejecutar algoritmos de Data Analytics.
- **Problemas de escalabilidad**, y capacidad en las bases de datos actuales de las empresas.
- Obtención de consultas cuyos procesos corren demasiado lentos.
- **Falta de capacidad** para manejar ecosistemas de Big Data, como el mencionado Hadoop u otros.
- **Ineficaces sistemas de visualización** de los datos extraídos.

1.2.5 Tendencias

El potencial de crecimiento de Data Analytics irá de la mano de la mejora en el conocimiento de las bases de datos de Big Data, la mejora en la capacidad de visualización de datos y la obtención de conocimiento en tiempo real.

¿Qué datos pueden ser explotados por los programas de análisis?

- **Datos no explotados** por los programas convencionales de Business Intelligence (BI).

- **Registros de servidores web** (por ejemplo datos de clics de Internet).
- **Contenido e informes de actividad de redes sociales.**
- **Texto de correos electrónicos de clientes** y respuestas de encuestas.
- **Datos de máquinas** capturados por sensores IoT.

El análisis de datos está íntimamente relacionado con el concepto Big Data y comparten campos y sectores de aplicación.

Podemos decir que las tendencias y aplicaciones están íntimamente relacionadas con el apartado **2.2.3.Tendencias y casos de uso de BIG DATA**, de este informe. Con el fin de no ser repetitivos incluimos en este apartado algunas de las aplicaciones de Data Analytics más desarrolladas hoy en día:

- **Asistencia sanitaria:** Uno de los factores más importantes en la asistencia sanitaria es la reducción de costes sin perder o incluso aumentar la calidad del servicio. Para ello, cada vez más se están utilizando los datos recogidos por equipos de medida así como también datos estáticos procedentes de las historias clínicas con el objetivo de mejorar los diagnósticos tanto en tiempo como en precisión, optimizar el tratamiento y el uso de equipos en los hospitales, con el objetivo de optimizar los flujos de pacientes y la calidad de la atención.
- **Viajes:** El principal desafío de las agencias de viaje es realizar una buena segmentación del cliente y mejorar su experiencia de compra a través de recomendaciones que se adecuen a su perfil. Para ello se pueden analizar los datos de los registros Web o los datos de redes sociales.
- **Juegos:** En la industria del juego se utiliza Data Analytics para recopilar datos de los jugadores, tanto a nivel de las relaciones de los jugadores como de sus gustos o aversiones. Con esta información es posible mejorar significativamente el gasto que se realiza, ya sea dentro del juego (in-game purchase) o en la compra del propio juego.
- **Gestión energética:** La mayoría de las empresas están utilizando el análisis de datos para la mejora de la gestión energética, incluida la gestión inteligente de la red, la optimización de la energía, la distribución de energía y la automatización de edificios en empresas de servicios públicos. Los contadores de suministro inteligentes generan una ingente cantidad de datos cuyo análisis permite detectar patrones de consumo que posibilitan mejorar los ratios de suministro energético
- **Telecomunicación:** La explosión de datos, impulsada por aplicaciones tales como registros de llamadas, monitoreo del tráfico de red, contenido digital, la gestión de activos, obliga a su tratamiento para extraer conocimiento útil para el desarrollo de negocio.
- **Fabricación:** En este sector la adopción de Big Data presenta diferencias entre los distintos subsectores. Es en subsectores como la automoción o aeronáutica donde la tecnología es más aceptada. Sin embargo, los beneficios que puede aportar Big Data aplicado a la industria, junto con el apoyo de los estados en la aplicación de estas tecnologías, harán que su crecimiento se fortalezca en los próximos años.

Las **herramientas de Web analytics** son usadas principalmente por los negocios online y empresas de marketing para recoger y medir el tráfico web, con el fin de obtener un mayor entendimiento del usuario y mejorar el sitio web. Las métricas más usadas para realizar el análisis del sitio web son: número de visitas, visitas únicas, tiempo en la página, localización de

los usuarios, porcentaje de salida y el porcentaje de conversiones. Obtener métricas y comprobar el comportamiento del usuario es crucial para saber qué elementos del sitio funcionan y cuáles no.

Estas herramientas están orientadas a conocer los hábitos de navegación de los usuarios a través de la web, y de manera más específica en los distintos portales de negocio online. Así, es posible obtener aspectos tales como el tipo de productos que despiertan mayor interés a diferentes escalas y profundidades, la caracterización de los usuarios, el coste de adquisición de cliente o venta, etc. En definitiva, la aplicación de estas herramientas en este tipo de negocios es de sumo interés ya que proporcionan información de alto valor para el diseño de estrategias de venta.

Por otro lado, **el análisis social**, término introducido por Martin Wattenberg en 2005 [21], es una forma de análisis donde las personas trabajan de forma colaborativa para dar sentido a los datos. En la actualidad este término ha tenido gran relevancia debido a su estrecha relación con el Big Data. El análisis de los datos sociales se centra principalmente en los datos generados en redes sociales tipo Facebook, Twitter, etc. o por otras aplicaciones donde los usuarios generan contenidos.

Como ejemplos de casos de **uso reales** de Data Analytics en distintos sectores podemos identificar:

BELK

Aplicación: Aumentar el número de clientes multicanal, optimizar el stock, optimizar el formato de la tienda, y las decisiones de horario

Resultado: Desarrollar modelos de tipología de cliente en función del nivel de gasto, histórico para identificar y dirigirse a clientes de alto valor añadido. Mejora del merchandising en tienda y optimización de la colocación del producto en tienda mediante el análisis de los datos obtenidos de los clientes.



Smarter cities

Aplicación: Optimizar las inversiones de las ciudades en infraestructura de transporte mejorando el flujo de tráfico

Resultado Los datos de tráfico se diseccionan y analizan por carril, por acción e incluso hasta el vehículo individual.



CARGO SMART

Aplicación: Logística, diagnóstico y pronóstico. Mejora de las rutas de las buques para ahorro de combustible, optimización de la tripulación y gastos portuarios

Resultado: CargoSmart gestiona datos sobre la velocidad de los buques, la ubicación en el puerto y en el mar, los tiempos de tránsito totales, rutas, eventos de atraque, eventos de excepción de 5500 buques – TIBCO.

Visualización en tiempo real, alerta y optimización de la velocidad y ruta del buque.



Identificación de patrones para monitorizar situaciones meteorológicas que influyen en retrasos y accidentes.

STUBHUB

Aplicación: Procesado y analizado de grandes volúmenes de datos en tiempo real para fortalecer la experiencia del cliente, y mejorar la detección y prevención del fraude



Resultado: Información analizada proveniente de 25 recursos de datos en un solo data warehouse. Permitiendo analizar a 180 millones de clientes y desarrollando tickets promocionales y campañas de promoción específicas.

KREDITECH

Aplicación: Ofrecer crédito a personas sin historial de crédito



Resultado: Desarrollo de un sofisticado modelo de puntuación basado en autoaprendizaje. Obtención de datos dinámicos a través de redes sociales, uso del móvil, localización, e-commerce.

Mercado del petróleo / gas

Aplicación: Inversión de capital basada en datos

Resultado: Modelo de simulación de la producción “Monte Carlo”.

Modelo de optimización que maximiza el retorno de la inversión y el flujo de caja dado las restricciones de presupuesto y la tolerancia al riesgo.

1.2.6 El binomio Big Data- Data Analytics

Si bien ambas tecnologías pueden presentarse por separado es su comunión la que presenta un potencial realmente disruptivo. Como consecuencia surge el concepto Big Data Analytics.

Según el estudio realizado por la consultora Gartner en 2015 “*Big Data Industry Insight*” [22] las aplicaciones de Big Data Analytics son múltiples y diferentes para cada tipo de sector.

De esta forma podremos encontrar aplicaciones en los siguientes sectores:

- Industria
- Media y Comunicaciones
- Servicios
- Administración pública
- Educación
- Venta al por menor
- Banca
- Seguros
- Salud
- Transporte

De la misma forma, en dicho estudio se clasifican los procesos de negocio en los que Big Data Analytics puede tener una implicación directa de mejora:

- Fortalecimiento de la experiencia del consumidor
- Eficiencia en la operativa de los procesos
- Segmentación de mercado
- Gestión del riesgo
- Desarrollo de nuevos productos
- Seguridad
- Cumplimiento de la regulación

En el mapa de calor resultado de este estudio que se muestra en la Ilustración 12, se puede observar cómo la mejora de la **experiencia del consumidor (Enhanced customer experience)** es el proceso de negocio en el que se ha priorizado más el uso de Big Data en todos los sectores. A este proceso lo siguen la eficiencia en la operativa de los procesos y la segmentación de mercado.

	Manu & N. Res.	Media/ Comm	Svcs	Gov.	Edu	Retail	Banking	Insurance	Health-care	Transportation	Utilities
Enhanced customer experience	52%	78%	66%	43%	76%	83%	77%	77%	73%	69%	44%
Process efficiency	45%	33%	35%	49%	65%	43%	41%	50%	73%	69%	78%
More targeted marketing	43%	89%	53%	17%	41%	78%	66%	58%	-	38%	17%
Cost reduction	42%	33%	35%	37%	35%	30%	41%	31%	45%	56%	61%
Improved risk management	14%	22%	29%	29%	35%	22%	52%	58%	55%	31%	61%
New products	23%	67%	37%	14%	24%	35%	27%	50%	-	19%	33%
Developing information products	26%	33%	44%	31%	12%	22%	23%	19%	9%	19%	11%
Enhanced security capabilities	17%	22%	21%	34%	29%	13%	27%	27%	9%	19%	28%
Regulatory compliance	11%	22%	18%	23%	18%	9%	25%	23%	27%	31%	44%
n=	65	9	62	35	17	23	44	26	11	16	18

ILUSTRACIÓN 12: PRIORIZACIÓN DEL USO DE BIG DATA POR SECTORES. FUENTE: GARTNER

Por otro lado y según el estudio *“Business opportunities: Big Data”* realizado para la UE en Julio de 2013 [23], las oportunidades de negocio y de mejora del negocio, que surgirán en el futuro pueden clasificarse de la siguiente manera:

- **Procesos de negocio horizontales** o intersectoriales
- **Procesos de negocio verticales**

Big Data Analytics tiene aplicaciones directas para el apoyo a los procesos de negocio industriales, no obstante muchos de los desafíos ante los que nos encontramos, no son sólo desafíos tecnológicos, sino desafíos organizacionales que se ven claramente afectados por las nuevas tecnologías.

Procesos horizontales en los que Big Data Analytics tienen y tendrán una aplicación directa:

- Gestión de relaciones con el cliente: mejora de la experiencia del cliente.

- Ventas y marketing: micro segmentación del cliente, análisis de los datos de redes sociales y móviles en tiempo real, venta cruzada de productos, gestión dinámica de los precios de venta al cliente, comercialización basada en localización.
- Cadena de suministro: optimización de la distribución y logística, gestión de la demanda y cadena de suministro, gestión y optimización de stocks.
- Producción y operación: producción inteligente.
- Administración (finanzas, contabilidad, recursos humanos...): sistemas de apoyo a la toma de decisiones, optimización de la planificación, detección de fraude.
- Investigación y desarrollo: Desde la generación de ideas hasta la gestión del ciclo de vida del producto.
- Gestión de tecnologías de la información y comunicación de la empresa.
- Gestión del riesgo.

Aunque como vemos, podemos identificar oportunidades en cada proceso, no debemos pensar en Big Data Analytics de forma estanca. Un uso efectivo de esta tecnología se expande y entrecruza entre los procesos de negocio mencionados, uniendo la investigación y desarrollo, la producción, operaciones, ventas, cadena de suministro, relaciones con el cliente, etc. para aportar valor al negocio.

En la siguiente imagen se representan algunas de las oportunidades de negocio detectadas para cada proceso empresarial horizontal.

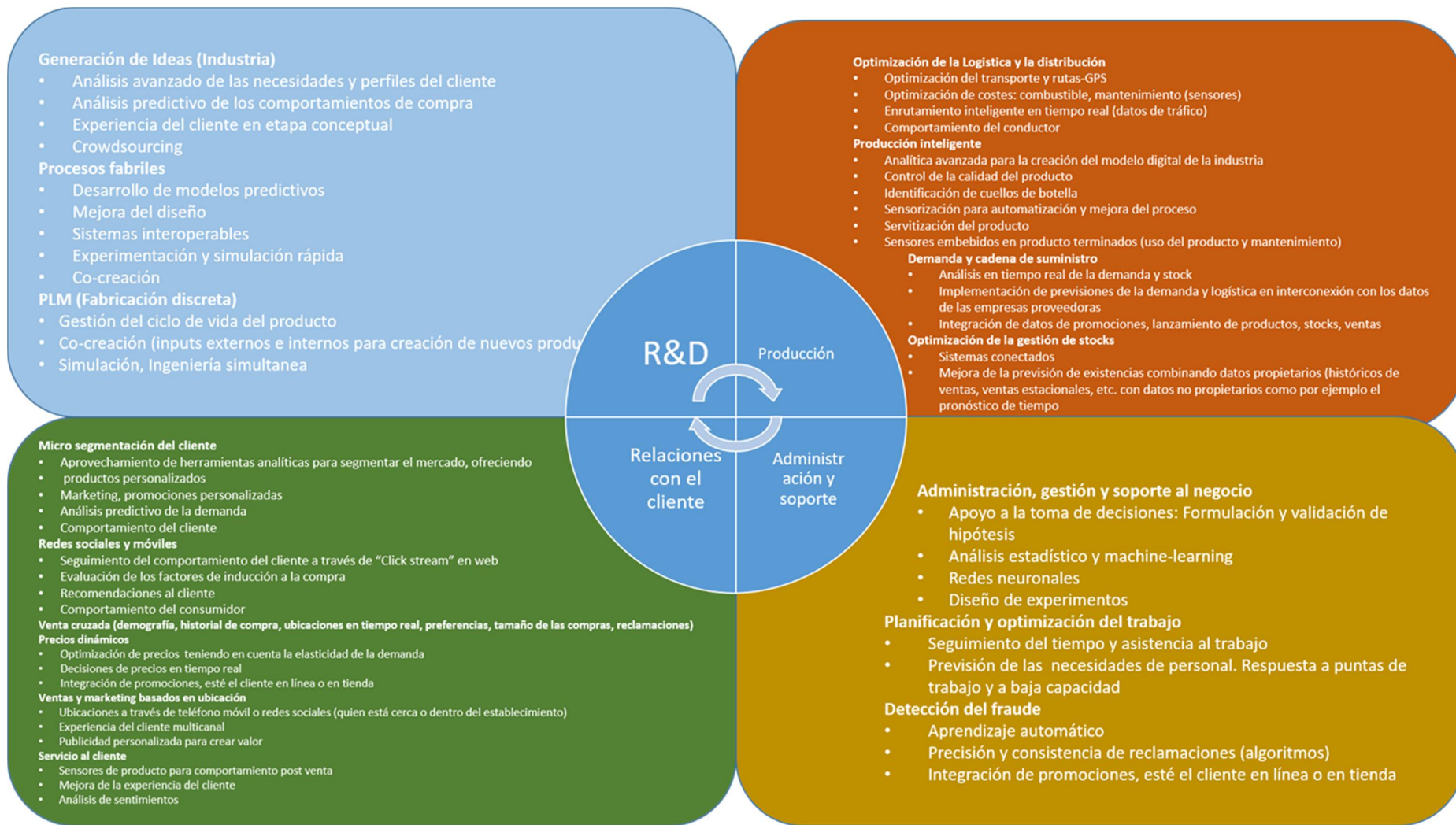


ILUSTRACIÓN 13: OPORTUNIDADES DE BIG DATA ANALYTICS POR PROCESOS DE NEGOCIO INTERSECTORIALES. FUENTE: ELABORACIÓN PROPIA.

Además de las oportunidades identificadas para procesos intersectoriales, las oportunidades específicas están relacionadas con los procesos, aplicaciones y fuentes de datos característicos de cada sector. Algunos ejemplos de oportunidades potenciales son las reflejadas en la siguiente tabla:

SERVICIOS FINANCIEROS

Negociación algorítmica	Prevención y detección de fraude en banca y seguros	Evaluaciones predictivas de daños en la industria de seguros Análisis de siniestros en seguros Modelado de catástrofes
Análisis de clientes a partir de la integración de datos transaccionales (de CRM, pagos con tarjeta de crédito, transacciones de cuentas) y feeds de medios sociales no estructurados	Evaluación de la cartera y la exposición al riesgo	Perfil de clientes, orientación y optimización de la venta cruzada
Centro de atención al cliente y eficiencia del centro de llamadas	Análisis del sentimiento y reputación de la marca	Gestión del valor del cliente

TELECOMUNICACIONES

Optimización de red	Retención de clientes basada en registros de llamadas y actividad en redes de abonados	Perfil de clientes, orientación y optimización de la venta cruzada
Centro de atención al cliente y eficiencia del centro de llamadas Prevención del fraude	Servicios basados en localización utilizando datos GPS y análisis geoespacial	Asignación de ancho de banda basada en patrones de uso

MEDIA

Ranking de clientes. Asignación de ancho de banda basada en patrones de acceso para secuencias de software de video, música y juegos	Prevención del fraude. Gestión de activos digitales	Gestión de la propiedad intelectual. Gestión de contenidos digitales
---	--	---

INSTALACIONES. PETRÓLEO Y GAS

Monitorización de señales inteligentes, análisis de patrones de uso en tiempo	Análisis predictivo, previsión de carga de distribución y programación	Monitorización de sensores para mantenimiento predictivo y basado en condiciones
---	--	--

real para optimizar el consumo de energía y el establecimiento de precios.		
Optimización inteligente de la red, patrón meteorológico, uso y distribución en tiempo real.	Modelado operativo.	Pronóstico y programación de la carga de distribución.
Gestión de desastres y apagones.	Exploración de recursos naturales en la industria del petróleo y el gas.	Procesamiento de datos sísmicos.
Vigilancia y optimización de perforaciones.		

TRANSPORTE

Optimización logística.	Análisis basado en la ubicación utilizando datos GPS.	Análisis de clientes y marketing de fidelización.
Mantenimiento predictivo.	Optimización de precios.	

VENTA AL POR MAYOR Y DETALLE

Optimización de la cadena de suministro.	Seguimiento RFID.	Optimización de precios y precios dinámicos.
Análisis del comportamiento de los clientes.	Conocimiento del cliente, micro-segmentación.	Análisis de fidelización y promociones.
Venta cruzada y venta en el punto de venta.	Patrones reales de compra de los clientes.	Análisis de la cesta de la compra basada en datos demográficos.
Optimización de mercancías.	Detección y prevención de fraude.	

INDUSTRIA

Mantenimiento predictivo.	Análisis de procesos y calidad. Gestión de garantía. Gestión de la calidad basada en los medios sociales.	Automatización de fábricas. Fábrica digital - Fabricación Lean
Monitorización de sensores para el mantenimiento de automóviles, edificios y maquinaria.	Monitorización de contadores inteligentes para optimizar el consumo de energía.	Análisis basado en la ubicación utilizando datos GPS.
Ingeniería concurrente y gestión del ciclo de vida del producto.	Análisis de comentarios de medios sociales para la gestión de calidad de vehículos en automoción.	Previsión de la demanda y planificación de la oferta.
Analítica de operaciones con sensores de datos.	Medios de comunicación social para la comercialización en las industrias de consumo.	Optimización de la distribución.

Es evidente que el despliegue de Big Data/Data Analytics en estas áreas ofrecerá enormes **oportunidades a las empresas** de los distintos sectores para agilizar sus procesos, reducir costes, incrementar su eficiencia, ofrecer mejores y/o nuevos productos y servicios. Al mismo tiempo, y desde el lado de la oferta de servicios y tecnologías, surgen nuevas oportunidades para poder satisfacer las necesidades de los sectores. La especialización será clave entre los ofertantes de estos nuevos servicios tecnológicos.

1.3 CLOUD COMPUTING

1.3.1 Breve Historia

En términos evolutivos, la **computación** ha cambiado significativamente y múltiples veces en su corta historia. En primeras etapas, se utilizaban grandes equipos, llamados mainframes para realizar operaciones altamente costosas, sin embargo a medida que la tecnología ha ido mejorando, esa tendencia se ha visto invertida hacia servidores más pequeños y más asequibles interconectados para construir lo que ahora se conoce como Cloud Computing.

La idea principal detrás de lo que entendemos como computación en la nube no es algo nuevo; ya en la década de los 60 John McCarthy preveía que las instalaciones de computación se proporcionarían al público en general como una utilidad. El término "**nube**" también se ha utilizado en diversos contextos como la descripción de grandes redes ATM en los años noventa. Sin embargo, fue con la llegada del nuevo milenio cuando se popularizó su uso gracias a que el CEO de Google, Eric Schmidt utilizó la palabra para describir el modelo de negocio de prestación de servicios a través de Internet.

La computación en la nube es un paradigma que proporciona servicios de Tecnologías de la Información a través de Internet o una red, y permite aumentar o disminuir dinámicamente la capacidad de la infraestructura para satisfacer las necesidades de uso. Al aprovechar la infraestructura compartida y las economías de escala, el **cloud computing** ofrece un sinfín de ventajas en términos económicos y de competitividad para las pequeñas y medianas empresas. Por ejemplo, permite a los usuarios controlar los servicios informáticos a los que se accede, reduciendo la inversión en los recursos de TI subyacentes ya que estos son compartidos por varios usuarios.

Además, los proveedores se benefician de economías de escala, lo que a su vez les permite reducir los costos de uso individual y centralizar los costos de infraestructura. Los usuarios pagan por lo que consumen, pueden aumentar o disminuir su uso y aprovechar los recursos subyacentes compartidos. Con un enfoque de cloud computing, un usuario puede dedicar menos tiempo a gestionar complejos recursos de TI y más tiempo a su trabajo.

1.3.2 Definición y características

El término **cloud computing** hace referencia a una tecnología que aúna diferentes ideas tan diversas como el almacenamiento de información, las comunicaciones entre ordenadores, la provisión de servicios o las metodologías de desarrollo de aplicaciones, todo ello bajo el mismo concepto: "todo ocurre en la nube".

Al igual que el resto de términos surgidos del vertiginoso desarrollo de la tecnología no existe una definición estandarizada para definir Cloud Computing. Aunque en este caso las diferencias son mínimas, por ejemplo el Instituto Nacional de Estándares y Tecnología americano (NIST) define cloud computing como:

"Un modelo para permitir un acceso conveniente a un conjunto compartido de recursos computacionales configurables (por ejemplo, redes, servidores, almacenamiento, aplicaciones y servicios) bajo demanda que se pueden aprovisionar y liberar rápidamente con un esfuerzo mínimo de gestión o una interacción entre el proveedor de servicios."

Por otro lado, en 2009 Heiser J. definió Cloud Computing como:

"Un estilo de computación altamente escalable donde las capacidades IT son proporcionadas 'como servicio' para el uso de usuarios externos".

La principal razón para la existencia de diferentes percepciones sobre la computación en la nube es que a diferencia de otros términos técnicos, **no es una nueva tecnología**, sino un nuevo modelo de operaciones que reúne un conjunto de tecnologías existentes para operar de una manera diferente. De hecho, la mayoría de las tecnologías utilizadas por la computación en nube como por ejemplo la virtualización, no son nuevas.

Centrándonos en la definición dada por el NIST, se identifican cinco características esenciales del cloud computing: **servicio a la carta, amplio acceso a la red, agrupación de recursos, elasticidad rápida y servicio medido**. El cloud computing tiene diversas formas de despliegue y cada una de ellas ofrece diferentes ventajas a los usuarios que migran sus aplicaciones hacia la nube.

- **Nube privada:** En este caso la infraestructura de la nube es operada únicamente para una organización aunque podría ser administrada por la organización o un tercero.
- **Nube comunitaria:** La infraestructura de la nube es compartida por varias organizaciones o comunidad específica que tienen objetivos comunes y compartidos. Al igual que las nubes privadas pueden ser administradas por las organizaciones o por un tercero.
- **Nube pública:** La infraestructura de la nube se pone a disposición del público en general o de un gran grupo industrial y es propiedad de una organización que vende servicios en la nube.
- **Nube híbrida:** La infraestructura de la nube es una composición de dos o más nubes (privadas, comunitarias o públicas) que siguen siendo entidades únicas, pero están unidas entre sí por una tecnología normalizada o propietaria que permite la portabilidad de datos y aplicaciones (por ejemplo, "cloud bursting" para balanceadores de carga entre distintas nubes).

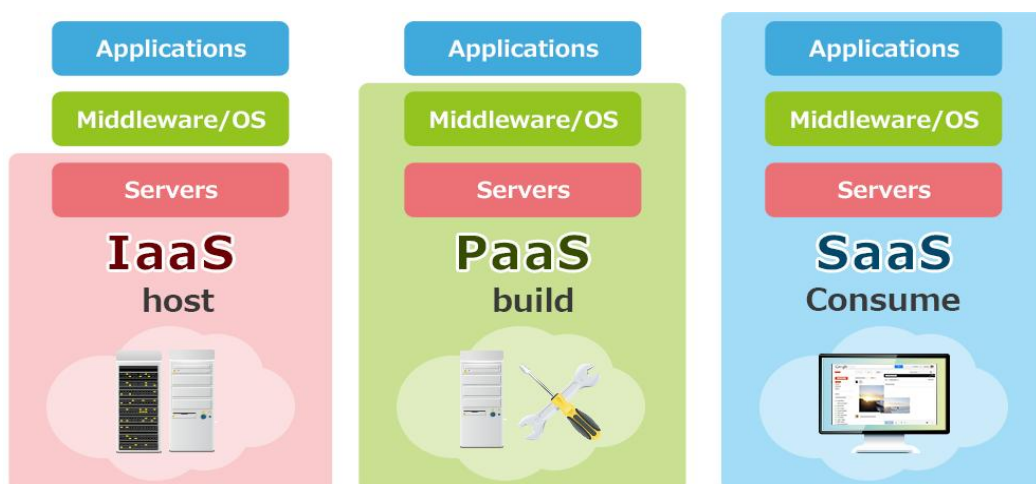


ILUSTRACIÓN 14: "UNDERSTANDING THE CLOUD COMPUTING STACK: SAAS, PAAS, IAAS". FUENTE: [HTTP://WWW.RACKSPACE.COM](http://www.rackspace.com)

En la práctica, los proveedores de servicios de la nube tienden a ofrecer servicios que pueden ser agrupados en tres categorías [24]:

- **Software como Servicio (SaaS):** Los sistemas *SaaS* ofrecen al usuario la capacidad de utilizar las aplicaciones del proveedor que se ejecutan en una infraestructura en la nube. Un ejemplo tradicional de este tipo de aplicaciones son los gestores de correo electrónico como gmail o aplicaciones más modernas como Dropbox, Evernote, etc. Las aplicaciones en la “nube” son accesibles por varios dispositivos del cliente a través de una interfaz sencilla, como puede ser un navegador web. El usuario del servicio no gestiona o controla la infraestructura subyacente del servicio, que incluye la red de comunicaciones, los servidores, los sistemas operativos y el almacenamiento.
- **Plataforma como Servicio (PaaS):** Los servicios *PaaS* ofrecen a los usuarios la capacidad de desplegar en la infraestructura de nube aplicaciones creadas por ellos mismos o adquiridas. El usuario no gestiona ni controla la infraestructura de la nube, incluyendo la red, los servidores, los sistemas operativos o el almacenamiento, pero tiene control sobre las aplicaciones desplegadas y, posiblemente, sobre las configuraciones del entorno de hospedaje de aplicaciones. Existen múltiples proveedores de servicios PaaS, no obstante los principales son Amazon a través de Amazon Web Service, Microsoft con su servicio Azure o Google con Google apps.
- **Infraestructura como servicio (IaaS):** En el caso de los sistemas *IaaS* el principal aporte es dotar al usuario de la capacidad de procesamiento, almacenamiento, redes y otros recursos de computación fundamentales donde se pueda desplegar y ejecutar cualquier software, pudiendo incluir sistemas operativos y aplicaciones. El usuario no gestiona ni controla la infraestructura de la nube, pero sí tiene control sobre el sistema operativo, el almacenamiento, las aplicaciones implementadas y, posiblemente, el control limitado de determinados componentes de red (por ejemplo, firewalls de host).

IaaS un modelo de Cloud Computing que permite utilizar recursos informáticos y el hardware de un proveedor en forma de servicio. Con ello, IaaS permite que los clientes puedan comprar recursos hardware (servidores, sistemas de almacenamiento, conmutadores, routers, etc.) como si se tratara de servicios totalmente externalizados. Con este modelo se pueden ampliar o reducir los recursos informáticos físicos de una empresa, en un periodo de tiempo muy breve.

En la siguiente figura se muestran los recursos administrados por cada categoría y ejemplos de plataformas reales que ofrecen servicios enmarcados en alguna de las categorías.

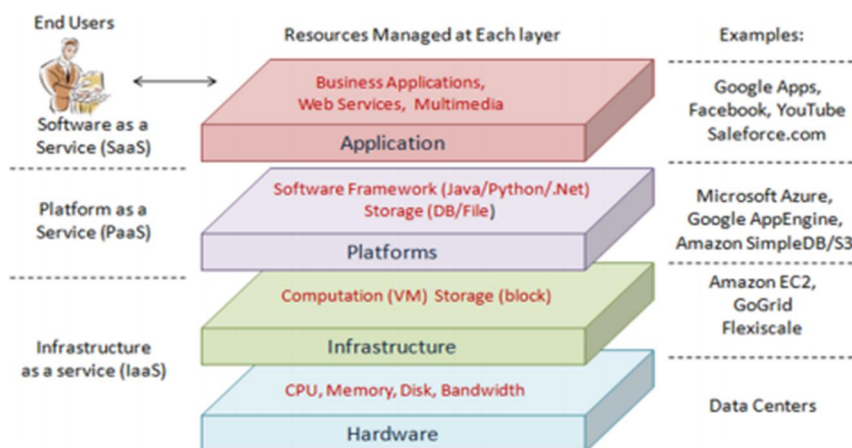


ILUSTRACIÓN 15: RECURSOS ADMINISTRADOS EN CADA CATEGORÍA. FUENTE: INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH. V-3, ISSUE-5, MAY-2012.

1.3.3 Ventajas y limitaciones

Las principales ventajas, a día de hoy de Cloud Computing son:

- Escalabilidad
- Ahorro de costes en equipamiento informático. Las aplicaciones son ejecutadas en la nube, por lo que no se necesitan equipos con altas potencias de procesamiento.
- Menor coste en desarrollo de software
- Eliminación de defectos por mala configuración
- Actualizaciones instantáneas
- Mucha mayor capacidad de almacenamiento
- Acceso desde cualquier parte y desde cualquier dispositivo a los documentos o datos

Sin embargo, la utilización de Cloud Computing también presenta limitaciones:

- Conexión permanente a Internet de alta velocidad
- Mayor facilidad para la vulnerabilidad de los datos
- Integración con otras aplicaciones de la empresa que puedan no estar en la nube.

1.3.4 Tendencias y casos de uso

TENDENCIAS

En la actualidad existen múltiples soluciones de Cloud disponibles para el consumidor. No obstante, se centrará el análisis de este apartado en las tres principales plataformas actuales que responden a las categorías **PaaS e IaaS**:

Amazon Web Services (AWS)

Amazon Web Services (AWS)²⁷ corresponde a la categoría IaaS principalmente. Es un conjunto de servicios en la nube que proporcionan capacidad de cómputo, almacenamiento y otras funciones permitiendo a las organizaciones y a los individuos implementar aplicaciones y servicios bajo demanda. Además, los servicios ofrecidos por Amazon Web Services son accesibles a través de HTTP, utilizando los protocolos REST y SOAP.

Aunque la cartera de productos que ofrece AWS va aumentando progresivamente los productos más utilizados por los usuarios son **Amazon Elastic Compute Cloud (Amazon EC2)** que permite a los usuarios iniciar y administrar instancias de servidor. Las instancias EC2 son máquinas virtuales que se ejecutan en la parte superior del motor de virtualización Xen²⁸. Después de crear e iniciar una instancia, los usuarios pueden cargar el software y configurarlo para posteriormente empaquetarse como una nueva imagen de máquina. Esta imagen puede ser desplegada en diferentes instancias en cualquier momento.

EC2 proporciona la capacidad de desplegar instancias en múltiples ubicaciones a lo largo del mundo, lo que hace que el sistema sea más robusto frente a problemas de infraestructura o ataques en una determinada localización. Los usuarios pueden elegir una región para optimizar la latencia, minimizar los costes, o por razones legales.

Amazon Virtual Private Cloud (VPC) es una conexión segura entre las infraestructuras de la empresa con la nube que ofrece AWS. Amazon VPC permite a las empresas conectar su infraestructura existente a un conjunto de recursos de computación AWS aislados a través de una conexión de red privada virtual (VPN) y extender sus capacidades de administración existentes como servicios de seguridad, firewalls y sistemas de detección de intrusos para incluir sus recursos AWS.

Otro de los productos que ofrece Amazon es **Amazon CloudWatch**, herramienta de gestión de información de los servicios AWS asociados, como Amazon EC2. Ofrece métricas o KPIs en tiempo real que permiten medir los parámetros más importantes como el número de peticiones por minuto, la utilización de la CPU, los bytes de entrada/salida de red, las operaciones de lectura /escritura de disco, etc.

²⁷ Amazon Web Services, <http://aws.amazon.com>

²⁸ XenSource Inc, Xen, <http://www.xensource.com>

Microsoft Windows Azure Platform

La plataforma Windows Azure de Microsoft²⁹ se corresponde principalmente a la categoría PasS. Integra y da acceso a tres subsistemas que proporcionan un conjunto específico de servicios a los usuarios de la infraestructura en la nube:

- Windows Azure proporciona un entorno basado en Windows (aunque actualmente existen versiones bajo Linux) para ejecutar aplicaciones y almacenar datos en servidores en centros de datos. Admite aplicaciones creadas bajo el framework .NET y otros lenguajes comunes soportados en sistemas Windows, como C#, Visual Basic o C++ entre otros.
- SQL Azure proporciona un servicio de almacenamiento en bases de datos relacionales basados en SQL Server,
- NET Services es un servicio que ofrece una infraestructura distribuida a aplicaciones basadas en la nube y locales.

Los desarrolladores pueden crear aplicaciones web utilizando tecnologías como **ASP.NET y Windows Communication Foundation (WCF)**, que se ejecutan como procesos batch independientes o aplicaciones que combinan ambas. Windows Azure permite almacenar datos en blobs, tablas y colas, todas ellas accesibles en un estilo RESTful via HTTP or HTTPS.

SQL Azure está compuesto por la base de datos de igual nombre y el sistema de sincronización de datos "Huron". La base de datos de SQL Azure se basa en Microsoft SQL Server, proporcionando un sistema de gestión de bases de datos (SGBD) en la nube. El acceso a los datos se puede realizar a través de ADO.NET y otras interfaces de acceso de Windows (JDBC, ODBC, etc.). Además, el sistema de sincronización de datos de "Huron" sincroniza datos relacionales entre varios SGBD locales permitiendo trabajar de manera local con la información almacenada en la nube.

Todos los recursos físicos, VMs y aplicaciones desplegadas en la infraestructura son controlados por un software llamado **"fabric controller"**. En función de las necesidades indicadas, *"fabric controller"* decide dónde deben ejecutarse las nuevas aplicaciones, eligiendo servidores físicos para optimizar la utilización del hardware.

Google App Engine

Google App Engine³⁰ se corresponde a la categoría PasS principalmente. Es una plataforma para aplicaciones web tradicionales gestionada por Google que soporta los lenguajes de programación Python, PHP, Node.js, Ruby, Go y Java. Los frameworks que se ejecutan en Google App Engine incluyen Django, CherryPy, Pylons y web2py entre otros, así como un framework personalizado de aplicaciones web de Google similar a JSP o ASP.NET.

Google gestiona el despliegue del software en un clúster gestionando las instancias según sea necesario. Las APIs actuales soportan características tales como almacenar y recuperar datos de una base de datos no relacional BigTable³¹ y el almacenamiento en caché.

²⁹ Microsoft Azure, <https://azure.microsoft.com/es-es/>

³⁰ Google App Engine, <https://appengine.google.com>

³¹ Big Table, <https://cloud.google.com/bigtable/>

La siguiente tabla resume las tres principales **ofertas de servicios de cloud computing** en base a los tipos de aplicación, modelos de computación, almacenamiento y escalado automático.

	AMAZON EC2	WINDOWS AZURE	GOOGLE APP ENGINE
Tipo	IaaS	PaaS	PaaS
Aplicaciones objetivo	Aplicaciones de carácter general	Aplicaciones windows de carácter general	Aplicaciones Web
Computación	Nivel de SO en máquinas virtuales XEN	Microsoft Language Runtime (CLR) VM	Frameworks web predefinidos
Almacenamiento Relacional	RDS	Relational DBs	Google Cloud SQL
Almacenamiento NoSQL/BigData	DynamoDB, EMR, Kinesis, Redshift	Windows Azure HDInsight	Cloud Datastore, Big Query, Hadoop
Autoescalado	Sí	Sí	Sí
Firewall/ACL	Sí	Sí	Sí
IP Pública	Sí	Sí	Sí
Nube híbrida	Sí	Sí	No

CASOS DE USO DE CLOUD COMPUTING

Hoy en día, las pequeñas y medianas empresas están cada vez más convencidas de que aprovechando la potencia que ofrecen los servicios en la nube, pueden obtener acceso rápido a aplicaciones empresariales que permitirán mejorar el rendimiento de los empleados, o aumentar drásticamente la capacidad de cómputo de la empresa mientras se reduce el coste.

Uno de los servicios más demandados por las empresas en la actualidad es la **gestión de grandes volúmenes de datos**. La mayoría de los proveedores de almacenamiento masivo de datos alquilan potentes servidores a los que se puede acceder vía Internet. De esta forma, se puede acceder a la información almacenada a

través de aplicaciones sencillas que permiten sincronizar y acceder a dicha información desde múltiples dispositivos en tiempo real.

Además, Cloud tiene ventajas al ofrecer servicios más escalables y tolerantes a fallos con rendimientos mejores. La computación en la nube puede proporcionar multitud de recursos de computación debido a su alta escalabilidad. Por tanto, las pequeñas empresas pueden hacer uso de estos servicios sin un alto coste, únicamente aumentando los recursos de hardware cuando haya un aumento en la necesidad de computación.

Cloud Computing pretende dotar a las empresas de una infraestructura donde se pueda acceder a la información almacenada, independientemente de los sistemas físicos que se utilizan o de su ubicación real, siempre y cuando se disponga de acceso a Internet. En definitiva:

La información ya no tiene que almacenarse necesariamente en los dispositivos informáticos de la empresa, sino en los sistemas proporcionados por la “nube”. Ya no es necesario instalar aplicaciones informáticas en los sistemas de la organización, sino que éstas se alojan y ejecutan en la nube, lo que permite liberar recursos, tales como la memoria de los ordenadores de la organización o su consumo de energía. Los recursos informáticos dispuestos en red son compartidos por varios usuarios y a través de distintos dispositivos, pudiendo trabajar conjuntamente sobre el mismo contenido.

Como ejemplos de **casos de uso reales** de Cloud Computing en distintos sectores podemos identificar:

Spotify:

Aplicación: Personalizar la experiencia de los usuarios y enriquecerla con datos relacionados con la música que escucha.

Resultado: El uso de Google Cloud Platform ha mejorado el tiempo de ejecución de las consultas usadas para obtener información de interés y ofrecérsela al usuario.



Hoteles NH:

Aplicación: Gestionar la agregación de valoraciones y opiniones de clientes y usuarios en la web.

Resultado: El sistema de análisis utiliza la base de datos en la nube de Google Cloud Platform para almacenar la información de los clientes. También se hace uso de la máquina de traducción que ofrece Google Cloud Platform. La traducción de las opiniones es imprescindible para que el sistema pueda extraer conclusiones que mejoren el posicionamiento web.



Bankinter:

Aplicación: Simulación de crédito-riesgo.

Resultado: Se utiliza la plataforma AWS de Amazon como parte integral de su aplicación de simulación de crédito-riesgo. La aplicación utiliza complejos algoritmos para realizar cinco



millones de simulaciones. Gracias a AWS, Bankinter redujo el tiempo medio de las soluciones de 23 horas a 20 minutos.

2. APLICACIONES POR SECTOR

A continuación, se muestran ejemplos de los beneficios del uso de **BIG DATA** en los distintos sectores objeto de estudio.

Las aplicaciones que se presentan pretenden dar una visión global del potencial de la tecnología y una mayor comprensión de su uso, al mismo tiempo que se observa la aceptación de la misma en distintos sectores para ser trasladada al entorno de una fábrica.

2.1 AGROALIMENTACIÓN Y BIO

A lo largo de la vida de un producto de **agroalimentación y bio** se generan multitud de datos diversos, tanto estructurados como no estructurados, que dan lugar una gran cantidad de nuevas opciones que permiten optimizar tanto los procesos de producción como los de distribución, suministro etc. y mejorar la experiencia cliente.

Cada vez es más común hablar de **trazabilidad y datos** a lo largo de la cadena alimentaria o de monitorización inteligente de la cadena de valor alimentaria (desde la producción, logística y distribución hasta el consumidor). En la base de esta nueva tecnología se encuentra Big Data Analytics, cuyo principal objetivo es la extracción de información y conocimiento de todos estos datos combinados, permitiendo ajustar la producción a la demanda solicitada en el mercado y haciendo los procesos cada vez más eficientes y precisos.

En la actualidad, alguna de las aplicaciones más comunes son las centradas en:

- **Ahorro en costes y mejoras en la productividad**, partiendo del análisis de datos en ámbitos como la previsión de cosechas, gestión de plagas o enfermedades y/u optimización del riego y abonado.
- **Transparencia hacia el consumidor** sobre el proceso productivo y su transformación: huella ecológica, impacto social, características nutricionales, interacción con la salud, recomendaciones culinarias y/o implicación cultural del producto.

En la actualidad existen ya soluciones en el mercado como por ejemplo **BYNSE**, entre otras, que se constituye como solución Big Data para la agricultura de precisión. Proporcionando información valiosa sobre las necesidades actuales y futuras de los cultivos a los gestores agroalimentarios, para mejorar así la gestión, ahorrar costes y mejorar la rentabilidad. [32]. Las tecnologías utilizadas por esta herramienta son entre otras: cloud service, Big Data Cluster, Knowledge Generator Framework, etc.



ILUSTRACIÓN 25. BYNSE BIG DATA. FUENTE: [HTTP://BYNSE.COM/BIG-DATA/](http://bynse.com/big-data/)

2.1.1 Proyectos de I+D

En España, proyectos como **Hortysis – Innterconecta** sobre “Control Remoto de Producción Hortícola en Invernaderos e Integración con Previsiones de Demanda y Sistemas de Comercialización” (<http://www.hispatec.es/proyectos/hortysis-innterconecta-feder/>) busca, a través de la obtención de datos captados de forma automática y continua, con modelos de estimación / predicción climática, conocer cómo afectan las variables meteorológicas a las manifestaciones periódicas o estacionales de las especies y su maduración.

El objetivo principal del proyecto HORTISYS es el control remoto de la producción en invernadero para:

- Manejar los cultivos de forma que se ajusten las producciones a los tiempos óptimos de comercialización en los principales mercados de consumo.
- Maximizar rendimientos productivos de las plantas.
- Planificación biológica de las plantas para el diseño de un modelo predictivo de la producción.
- Diseño de un modelo de estimación de la demanda en función de las temperaturas e histórico de consumo en los mercados de destino.
- Diseño de un modelo de indicadores biológicos que permitan al productor controlar su cultivo y adaptarlo a la potencial demanda calculada a través del modelo de estimación de demanda anteriormente indicado.

Todos estos datos combinados con el modelo de estimación de la demanda, permiten al productor ajustar su producción a las necesidades del mercado.

Desde el punto de vista de la venta al por menor de productos agroalimentarios y bio, la principal fortaleza del uso de Big Data la encontramos en la **mejora de la experiencia del cliente**.

De este modo, podemos encontrar prácticas de Big Data Analytics, orientadas a:

- Aumentar el número de clientes multicanal
- Optimizar el stock
- Optimizar el formato de tienda

- Modificar y flexibilizar los horarios en tiempo real
- Etc.

DataBio Data Driven Bioeconomy [33]. Data Bio propone implementar una plataforma de big data de vanguardia: la plataforma Big DATABIO

Regional crop monitoring and assessment with quantitative remote sensing and data assimilation. [34]. Su objetivo es aplicar técnicas avanzadas de asimilación de datos a múltiples tipos de datos de cultivos, tanto de modelos de crecimiento de cultivos calibrados como de imágenes satelitales, para producir estimaciones mejores de la productividad agrícola de China. De esta forma se podrán evaluar las posibles geografías del futuro estrés agrícola en China.

PLANNING: MIDWEST: Cyberinfrastructure to Enhance Data Quality and Support Reproducible Results in Sensor Originated Big Data. BD Spokes Project [35]. Este proyecto tiene como objetivo crear y fomentar una comunidad multidisciplinaria centrada en la calidad de los datos y la reproducibilidad de los resultados de la investigación para experimentos basados en sensores. El proyecto también dará como resultado avances en el uso de circuitos integrados en el área de aplicaciones agrícolas en relación al uso de sensores y a la calidad de los datos.

14TSB_ATC_IR Optimising Big Data to Drive Sustainable Agricultural Intensification [36]. Desarrollo de aplicaciones móviles y servicios de datos web relacionados para proporcionar a los productores el acceso síncrono y georreferenciado al banco de datos Soil-for-Life (SfL). Los productores podrán consultar datos armonizados por parcela, campo por campo, tanto para operaciones históricas como actuales. Este proyecto proporciona evidencia científica para apoyar sistemas de intensificación sostenible y para mantener la salud del suelo a nivel de campo, granja y empresa.

2.2 AUTOMOCIÓN

Las aplicaciones dentro del vehículo están emergiendo cada vez más rápido (optimización de consumo, sensorica de todo tipo, adaptación de cambios, protección de ciberataques, etc.) y es éste un ámbito que seguirá creciendo a medida que los vehículos empiecen a estar conectados entre sí y con las infraestructuras, y/o se avance hacia la conducción autónoma.

Las aplicaciones de Big Data en la fabricación en el sector automoción se centran en aspectos como:

- **Mantenimiento predictivo de maquinaria**
- **Planificación** y asignación óptima de recursos de producción
- **Predicción de fallos** en cadenas de producción
- **Optimización logística** de suministro y distribución
- Pronóstico de la **demanda**
- **Eficiencia energética** en producción
- Despliegue óptimo de bienes y cadenas de **producción (lay-outs)**
- **Seguridad** en la planta
- Análisis de **riesgos y predicción** de fallos
- **Automóvil personalizado**

Además de la aplicación puramente fabril mencionada anteriormente, el volumen de los datos generados por un futuro automóvil conectado, junto con la conexión a las redes de su conductor o usuarios, abren **nuevos campos y modelos de negocio** para el sector de la automoción, de forma que, poco a poco, ya no sólo nos refiramos a procesos centrados en la fabricación si no a procesos centrados en el cliente, como por ejemplo:

- Señales de emergencia reemplazadas por **notificaciones** enviadas a un teléfono inteligente
- Vehículo cuyos elementos **informen directamente de daños** sufridos a un especialista que tendrá inmediatamente un repuesto listo
- Mantenimiento e inspección técnica llevados a cabo de **forma remota** cuando el coche se encuentra estacionado
- Compañías aseguradoras del sector del automóvil que ven en este gran volumen de datos la posibilidad de lograr un **repositorio mundial con todos los siniestros**, incluyendo todos los datos proporcionados por los sensores y comunicaciones del fabricante, junto con los datos de la incidencia y resolución de las mismas proporcionadas por las aseguradoras
- Etc.

A medida que se avance en la tecnología **Big Data Analytics**, el fabricante dispondrá de los datos necesarios para acercarse más a las necesidades en términos de seguridad, conectividad, y comodidad, de experiencia cliente, vehículo a medida, comportamiento del conductor, gustos, compras que realiza por internet, etc. La propiedad de los datos que pueden ser generados por un conductor o usuario del automóvil, plantea retos todavía, lo que favorece la irrupción en el sector de nuevos actores cuyo principal potencial es la captación e interpretación de los datos.

Los grandes fabricantes del sector automoción han empezado a aliarse con **partners tecnológicos** que operan de forma diferente, tienen una mejor predisposición a la innovación y disponen de ciclos de lanzamiento al mercado menores. Los modelos de negocios de estos nuevos actores son radicalmente diferentes a los de los fabricantes: centrados en los ingresos por servicio y en la venta de información.

Según el estudio **“Connected car report 2016”** [37] elaborado por PWC: *“los ingresos disponibles para la industria del automóvil y su ecosistema están modificándose”*. De forma que en la actualidad y en el futuro próximo podamos ver ingresos en el sector derivados de:

- Ventas de paquetes de conexión incluidos en los vehículos nuevos (Audi, Mercedes-Benz y Tesla...)
- Uso de datos de automóvil conectados para aumentar la eficiencia interna, la calidad y la diferenciación del producto
- Estrategia de diferenciación a través de la utilización de servicios conectados para reforzar la lealtad del cliente de automoción
- Establecimiento de un ecosistema integral de servicios al consumidor, con participación en los ingresos derivado de la venta de datos por parte del usuario
- Creación de sistemas para el uso de datos de clientes, tales como una base de datos de información de clientes, que se monetizarán a través de futuros modelos de negocio (y aún no especificados), especialmente en servicios de movilidad y opciones de transporte multimodal

De acuerdo con estas tendencias, nuevos operadores tecnológicos han irrumpido en el sector ya sea de forma individual o a través de alianzas con los fabricantes, como por ejemplo:

Mobileye: ofrece soluciones completas de sistemas avanzados de asistencia al conductor.

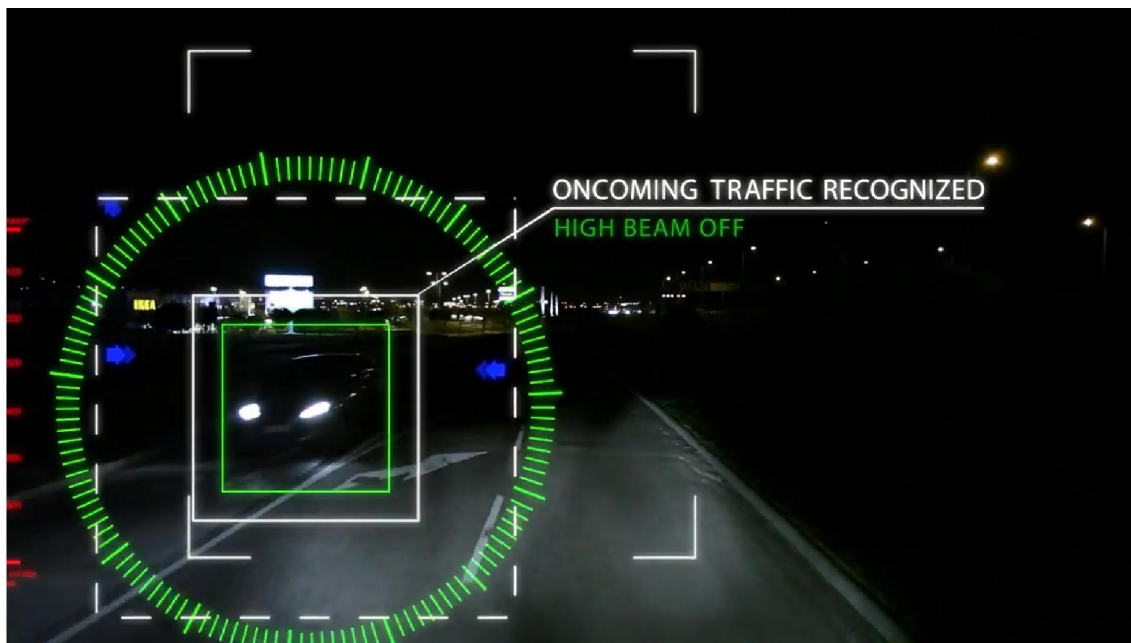


ILUSTRACIÓN 26. INTELLIGEN HIGH BEAM CONTROL. FUENTE: [HTTP://WWW.MOBILEYE.COM/EN-US/TECHNOLOGY/FEATURES/](http://www.mobileye.com/en-us/technology/features/)

Nvidia: que fabrica sistemas de información para cuadro de mando para la conducción y el establecimiento y guiado de mapas autónomo.

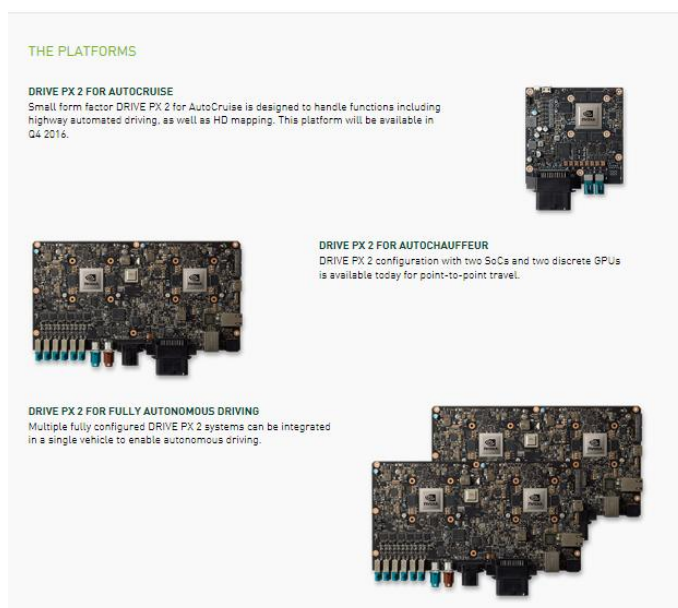


ILUSTRACIÓN 27: PLATAFORMAS NVIDIA PARA AUTOMOCIÓN. FUENTE: [HTTP://WWW.NVIDIA.COM/OBJECT/DRIVE-PX.HTML](http://www.nvidia.com/object/drive-px.html)

Los sistemas de visualización de AUDI son realizados en colaboración con Nvidia. Mercedes-Benz y NVIDIA han anunciado una asociación para traer el automóvil NVIDIA AI al mercado. El trabajo es parte de una colaboración continúa enfocada en el aprendizaje profundo y la inteligencia artificial.

Volvo planea utilizar NVIDIA DRIVE como parte de su proyecto "Drive Me", que pondrá 100 vehículos de prueba en un conjunto definido de carreteras en Gotemburgo.

Baidu y NVIDIA están colaborando en la primera plataforma de inteligencia artificial abierta para todos los fabricantes de automoción. La colaboración combina la plataforma de computación en la nube de Baidu con las soluciones de vehículo autónomo y aparcamiento autónomo de Nvidia, entre otras.

Las GPU NVIDIA alimentan los sistemas de navegación e información del automóvil de toda la línea de automóviles BMW de próxima generación. Todos los BMWs llevan versiones de iDrive, navegación BMW y sistema de información de vehículos de NVIDIA.

Los vehículos Tesla -Modelo S, Modelo X y el próximo Modelo 3- estarán equipados con un "superordenador" integrado a bordo de NVIDIA, que puede proporcionar una capacidad de autoconducción completa.

El consorcio BMW, Audi y Daimler ha adquirido la empresa Here, que se dedica a la fabricación y servicios de GPS y mapas para los vehículos. El consorcio pretende aprovechar el Hardware desarrollado por Here para recopilar datos del vehículo y proporcionar recomendaciones a los conductores en tiempo real: mejores rutas a escoger para llegar a su destino, la necesidad de realizar mantenimiento del vehículo, etc.

Drive Smart. Esta aplicación se instala de forma voluntaria en el SmartPhone de los conductores para permitir la recogida de información relativa a la conducción, esencialmente localización GPS y velocidad. Gracias al procesamiento de estos datos combinados con otros como velocidad máxima de la vía, sentido de circulación de la vía, tráfico o meteorología, es capaz de analizar la conducción del conductor para así poder ofrecerle recomendaciones que le permitan conducir mejor y, al mismo tiempo, identificar el perfil del conductor. Los mejores conductores se benefician de descuentos en seguros de automóvil, gasolineras u otros comercios, que al mismo tiempo se benefician de publicitar estos servicios mediante la aplicación.

Volvo está utilizando Big Data en el contexto de la inspección, el servicio y la venta de camiones Volvo, y facilitando la gestión de riesgos para Volvo Used Trucks EMEA. [63]

Daimler AG busca la manera de maximizar el número de culatas producidas en su fábrica de Stuttgart a través de ajustes de proceso específicos. La compañía quiere aumentar la productividad y acortar los tiempos de fabricación. Con el software IBM® SPSS®, Daimler reúne datos sobre más de 500 factores que incluyen dimensiones, tiempos, temperaturas, herramientas y muchos otros atributos de la producción de culatas en todo el proceso de producción. Los datos se procesan a diario y se evalúan automáticamente de varias formas con el software de análisis predictivo de IBM SPSS. Esto permite una monitorización completa de todos los parámetros del proceso [38].

AutoRose es una compañía de automóviles que busca investigar el valor de mercado potencial de Big Data dentro del espacio de la industria automotriz. El proyecto "A market investigation into alternate monetisation methods for a connected car network infrastructure and how to create a network structure which provides maximum value to all stakeholders" se enfoca en una cantidad de diferentes opciones de utilización de datos y las percepciones relacionadas de seguridad y del conductor hacia ellas.

2.2.1 Proyectos I+D

Cloud-LSVA - Cloud Large Scale Video Analysis [39]. Cloud-LSVA creará Tecnología Big Data para abordar la falta de herramientas software y hardware para la lectura de datos de video a gran escala, orden de magnitud petabyte. Las tecnologías dependen del análisis de video y otros datos de sensores del vehículo. Las anotaciones de objetos, eventos y escenas de la circulación de los automóviles son fundamentales para entrenar y probar las técnicas de visión por computador que son la base de los sistemas ADAS y de navegación. El proyecto busca desarrollar una herramienta comercial que aborde la necesidad de una anotación semiautomatizada y que aproveche la versatilidad de la computación en la nube para reducir costes.

2.3 MADERA Y FORESTAL

Los bosques representan uno de los tipos de ecosistemas terrestres más importantes, tanto por la biodiversidad que albergan como por los servicios ecosistémicos que proporcionan. La multitud de datos estructurados y no estructurados que pueden estar ligados a un ecosistema terrestre es ingente. De ahí que Big Data Analytics tenga en el sector madera y forestal una importancia relevante en aspectos como:

- Cambio climático
- Predicción de plagas,
- Predicción y simulación de incendios
- Predicción de contaminación
- Desarrollo de modelos de supervivencia
- Seguimiento del uso del suelo
- Planificación del territorio
- Gestión de datos georreferenciados
- Patrones de distribución de especies
- Regeneración de especies
- Etc.

El **clima** es uno de los ámbitos científicos que, por sus características, es de los más adecuados a las investigaciones basadas en el Big Data y en la computación en la Nube. Por el volumen, velocidad y variabilidad de los datos generados, Big Data parece la herramienta apropiada para afrontar los enormes retos que plantean la **modelización y el seguimiento del fenómeno del cambio climático**. En este sentido uno de los grandes retos que tienen los Servicios Meteorológicos Nacionales es la gestión de la gran cantidad de información y el desarrollo de herramientas eficientes para la extracción de conocimiento. Entre la información meteorológica se encuentran observaciones in-situ, imágenes de satélites, imágenes de radares meteorológicos, predicciones meteorológicas, avisos de fenómenos adversos, modelos numéricos meteorológicos, modelos climáticos, etc., información que debe ser procesada a diferentes niveles dependiendo de las necesidades del usuario.

Algunos ejemplos de aplicación:

- **Prevención de incendios** a través de una red de nodos inalámbricos desplegados capaces de captar temperatura, humedad, dirección y velocidad del viento, etc. A partir de los datos obtenidos y la

introducción de Big Data Analytics se generan patrones predictivos que podrían ayudar a la prevención contra el fuego, al disponer de información precisa y en el mismo momento en que un incendio se produzca. Como ejemplo de proyectos desarrollados en esta temática destaca el Proyecto de Investigación B105 Electronyc System de la Universidad Politécnica de Madrid, desarrollos llevados a cabo por expertos de la Universidad Politécnica de Valencia y la empresa ISDEFE, entre otros.



ILUSTRACIÓN 28. BOSQUE FORESTAL. FUENTE: IMAGEN LIBRE

- **La elaboración de patrones de Distribución de las especies** [41]. Análisis de los cambios de distribución y a qué pueden ser debidos, para poder elaborar modelos predictivos en función del cambio climático o del uso que el hombre hace de los terrenos forestales.
- **La regeneración de especies** [42]. Integración de datos de inventario y modelos de hábitat para predecir la regeneración de especies leñosas mediterráneas en repoblaciones forestales. Rafael M^a Navarro Cerrillo, Inmaculada Clavero Rumbaó, Astrid Lorenzo Vidaña, José Luis Quero Pérez, Joaquín Duque-Lazo.
- **Los efectos del cambio climático en los ecosistemas forestales** [41].
- **La evolución, mantenimiento y funcionamiento de la diversidad de especies** [41].
- **La facilidad de acceso a grandes datos a través de tecnologías semánticas en agricultura y bosque.** [40]

2.4 NAVAL

Big Data supone una oportunidad para la mejora de la productividad y competitividad del sector naval, tanto en las fases de construcción o reparación como en las de mantenimiento.

La aplicabilidad de Big Data, además, puede abarcar tanto al proceso constructivo o de reparación, como al mantenimiento.

Entre los posibles casos de aplicación de Big Data al sector, destacan tres:

- **Optimización de fabricación de bloques** en buques individuales o series de buques
- **Sistemas autónomos** o semiasistidos de fabricación
- **Análisis modal de fallos** en buques

En el primero de los casos, la obtención de información heterogénea, numérica y subjetiva del proceso y su contexto, relativa a la fabricación de bloques de un mismo buque o de una serie de buques con características similares facilita el análisis de los factores que influyen en la productividad de este tipo de procesos; este enfoque puede aplicarse para análisis de bloques armados en el propio astillero o bien en astilleros externos. La integración de información heterogénea y la aplicación de técnicas de tratamiento basado en análisis clúster o en otras técnicas de analítica avanzada pueden ser abordadas a través de un entorno tecnológico basado en el enfoque Big Data [43].

En el segundo de los casos, un entorno Big Data puede servir para **dotar de inteligencia** a los sistemas semiautomatizados o automáticos de conformado de chapa en caliente contando con información detallada de la secuencia de operaciones de un número elevado de piezas de igual geometría.

En el tercer caso, buques en funcionamiento, el empleo de Big Data e inteligencia artificial es útil para la **optimización de los tiempos** relativos a los procesos de mantenimiento así como a la reducción de fallos y su análisis causal.

Como ejemplo práctico, **Navantia**, con un centro de producción en Ferrol y Fene, está buscando modernizar su astillero gracias al proyecto Astillero 4.0, con el que planea construir al menos 5 fragatas F110 para la Armada Española. Dentro del conjunto de tecnologías que posiblemente serán utilizadas, podrá encontrarse Big Data. El objetivo será la modificación de las cadenas de producción, que pivotarán sobre el astillero inteligente, garantizando una producción segura, rápida y adaptada en tiempo real según las necesidades del mercado, y ofreciendo una mejor relación coste-beneficio y menos errores en la fabricación.

Así mismo se ha creado una **Unidad Mixta de Investigación entre Navantia y la Universidad de A Coruña** [44] para el desarrollo de tecnologías habilitadoras de Industria 4.0 en el sector naval con líneas de investigación como:

- Modelado y simulación de procesos de fabricación de la planta y de los productos desarrollados que permitan un astillero sostenible.
- Aplicación de la automatización y robotización de los procesos y su digitalización (movilidad, realidad aumentada, IoT, Big Data)
- Etc.

Wärtsilä Genius Services [45]

Los productos Wärtsilä Genius - Optimizar, Predecir y Resolver - aplican los datos para optimizar los activos de los clientes en tiempo real, mejorando la previsibilidad y ayudando a resolver problemas a través de soluciones digitales.

Utilizando datos históricos y en tiempo real, Wärtsilä Genius Services está diseñado para optimizar, desde la eficiencia energética de una sola instalación hasta la gestión de toda una flota. Esto último se logra integrando la planificación dinámica avanzada de la travesía, los servicios de asesoramiento sobre la eficiencia de los buques y el análisis energético, así como una amplia vigilancia de la situación de los principales equipos.

El servicio de monitorización de la eficiencia del motor (EEMS) de Wärtsilä funciona con cualquier motor de cuatro tiempos.

AkzoNobel and Tessella [46] han desarrollado una herramienta para la industria naviera de predicción del rendimiento de las tecnologías de recubrimiento. Intertrac Vision utiliza miles de millones de puntos de

datos de rutas de barcos y el riesgo de contaminación y, a través de técnicas analíticas avanzadas, algoritmos y modelos, proporciona evaluaciones precisas y completas de los recubrimientos. Genera un análisis completo de coste-beneficio, y detalla el consumo estimado de combustible, el coste del combustible y las emisiones de CO2, identificando la opción más eficiente en diferentes escenarios.

ABB AG ha desarrollado la herramienta EMMA Advisory Suite [47] que utiliza datos AIS sobre Apache HBase para predecir el comportamiento de la navegación en situaciones de alto tráfico marítimo.

2.5 TEXTIL/MODA

Hasta hace relativamente poco, uno de los ámbitos más conocidos de aplicación de Análisis de Datos en el sector textil era la **medición de flujos de clientes**, tanto dentro como fuera de los propios establecimientos o centros comerciales. Sin embargo, dada la irrupción del e-commerce han surgido numerosas líneas de aplicación de Big Data. Entre otras, Big Data permite detectar potenciales clientes, reducir costes de marketing o personalizar acciones, la utilización de modelos para predecir picos de demanda en función de históricos o tendencias por temporada, la realización de la planificación en función de las predicciones, y la utilización de analítica predictiva para la realización de ofertas en punto de venta, ventas cruzadas, recomendaciones online personalizadas, campañas personalizadas de promoción, etc.

Big Data permite **diseñar prendas adecuadas y ajustadas** a las necesidades del consumidor.

En los procesos de fabricación del sector textil Big Data tiene también claros ejemplos de aplicación:

- Mantenimiento predictivo de maquinaria
- Planificación y asignación óptima de recursos de producción
- Predicción de fallos en cadenas de producción
- Optimización logística de suministro y distribución
- Pronóstico de la demanda
- Eficiencia energética en producción
- Despliegue óptimo de bienes y cadenas de producción (lay-outs)
- Seguridad en la planta
- Análisis de riesgos y predicción de fallos

Como ejemplos prácticos podemos destacar, el **diseño de sistema de fabricación textil basado en big data** [48]. El sistema diseñado realizó un enlace de información efectivo entre la capa de planificación y la capa de producción, proporcionando un nuevo método para la detección en tiempo real de la calidad de la tela. Para el desarrollo del estudio se utilizó tecnología Hadoop, métodos teóricos de evidencia D-S, clustering incremental, y algoritmos, entre otras tecnologías.

Inditex, en función de la aplicación de determinadas técnicas de clusterización, es capaz de predecir las tallas que más se van a vender, en función de la localización de cada tienda.

2.5.1 Proyectos de I+D

SOMATCH - Support IT solution for creative fashion designers by integrated software systems to collect, define and visualize textile and clothing trends through innovative image analysis from open data [49]. Herramienta para analítica de datos y visualización de grandes conjuntos de datos no estructurados, relacionados con el uso y las preferencias de los productos de moda por parte de los consumidores, apoyando la rápida reacción de las empresas a la dinámica del mercado y una mejor adaptación del diseño a la demanda real de los consumidores. SOMATCH proporcionará a los diseñadores estimaciones de tendencias y pronósticos de aceptación del usuario. Además integrará los sistemas con los nuevos dispositivos SoA mobile y wearable (por ejemplo, Google Glass) para recopilar información y visualizar la interpretación de tendencias.

2.6 AERONÁUTICA

El volumen de datos que pueden generarse en un avión a lo largo de todo un recorrido, como la temperatura registrada en sensores repartidos por toda la aeronave, nivel de combustible, humedad, altitud, velocidad, posición, imágenes de cabina, condiciones climáticas externas, etc. es muy elevado. Esto hace que **este sector sea proclive a la utilización de estas tecnologías**, desde aspectos relacionados directamente con los vuelos como la monitorización de rutas, o la seguridad de datos de cajas negras, etc. hasta los procesos de fabricación, como los mencionados para otros sectores:

- Eficiencia del mantenimiento de las aeronaves. Integración de fuentes de datos dispares, como registros electrónicos de mantenimiento, datos paramétricos de aeronaves y datos operacionales para crear un conjunto de Big Data, sobre el que nuevas tecnologías de análisis y optimización de decisiones podrán ser aplicadas
- Planificación y asignación óptima de recursos de producción
- Predicción de fallos en cadenas de producción
- Optimización logística de suministro y distribución
- Pronóstico de la demanda
- Eficiencia energética en producción
- Despliegue óptimo de bienes y cadenas de producción (lay-outs)
- Seguridad en la planta
- Análisis de riesgos y predicción de fallos
- Reducción del tiempo de pruebas de una aeronave [50]
- Producción de series donde cada componente aeronáutico se personaliza siguiendo especificaciones definidas potencialmente por cada cliente individual.

Entre algunos de los casos de uso se encuentran:

La **Universidad de Michigan** colaborando con **IBM** [51] para desarrollar sistemas de supercomputación "centrados en datos" en campos tan diversos como diseño de motores y aviones, tratamiento de enfermedades cardiovasculares, física de materiales, modelado climático y cosmología. Los sistemas de IBM usan un enfoque acelerado basado en datos de GPU, integrando datasets masivos. ConFlux, el nuevo sistema, permitirá que los clústeres de computación de alto rendimiento se comuniquen directamente y a velocidades interactivas con operaciones de uso intensivo de datos. El proyecto establece un ecosistema de

hardware y software para permitir el modelado a gran escala basado en datos de problemas físicos complejos, como el rendimiento de un motor de avión.

La compañía **Lokad** ha desarrollado un software de optimización cuantitativa para la cadena de suministro aeronáutica (mantenimiento, reparación, reacondicionamiento y fabricantes de piezas originales). Las aeronaves requieren una gran variedad de piezas, desde las más caras hasta las más baratas pero con un alto grado de rotación. El hecho de que no se disponga de una pieza en un determinado momento se traduce en altos costes, que pueden afectar tanto a la empresa fabricante como a la aerolínea. El software desarrollado ofrece una solución estadística que proporciona una optimización exhaustiva del inventario a través del pronóstico de la demanda de las aerolíneas.

En el caso de la evolución de **las cajas negras**, Big Data y Cloud Computing y la obtención de los datos en tiempo real, tienen una gran relevancia. Las investigaciones actuales tratan de posibilitar el acceso a las cajas negras antes de que el avión tome tierra, momento en el cual los datos son disponibles. De esta forma, entraríamos en una nueva dimensión en la que el análisis de los datos de vuelo pasaría de ser de correctivo a preventivo.

BRITISH AIRWAYS: el programa "Know Me" combina la información de fidelidad ya existente con los datos recopilados de los clientes en función de su comportamiento en línea. Con la combinación de estas dos fuentes de información, British Airways puede hacer ofertas más específicas [52].

DELTA: aerolínea que permite a los clientes rastrear sus maletas desde dispositivos móviles [52].

2.6.1 Proyectos de I+D

DART - Data-driven Aircraft Trajectory prediction research. [53]. DART brindará comprensión sobre la idoneidad de aplicar técnicas de big data para predecir las trayectorias de aeronaves.

2.7 TIC

La industria de las tecnologías de la información y comunicación es propicia para la aplicación de técnicas de Análisis de Datos y/o Big Data. La aplicación de esta tecnología redundará en un incremento de la eficiencia operacional, apoyo a la toma de decisiones en tiempo real, aumento de eficiencia en las campañas de marketing, mejora de la experiencia con clientes, o la creación de modelos de negocio innovadores, etc.

El tratamiento de los datos y la obtención de información **aportan valor añadido** a las operaciones propias del modelo de negocio de las empresas TIC. El sector TIC es uno de los primeros interesados en ofrecer soluciones de captura y análisis de datos a sus clientes de forma que puedan obtener ventajas competitivas.

El sector se encuentra delante de una **multitud de oportunidades de desarrollo** de software relacionado con [54]:

- Gestión avanzada de call centers: identificación de problemas en tiempo crítico, maximización de retorno de los clientes y eficiencia en la conservación de clientes
- Analítica para redes: identificación de cuellos de botella en base a logs, predicción de capacidad y demanda para dimensionamiento óptimo de redes, descarga de tráfico celular a redes oportunistas
- Despliegue óptimo de infraestructura de redes con criterios económicos, de capacidad, impacto visual ...

- Análisis predictivo de niveles de ocupación de recursos de red (espectral)
- Servicios de valor añadido basados en información de localización estimada por redes de telecomunicación (útil en dominio de transporte y comercio).
- Análisis de contenido de la red y perfil de usuario (Deep packed analyses)

Entre los casos de uso podemos destacar:

IBM ha desarrollado un paquete de soluciones llamado IBM InfoSphere BigInsights, que permite optimizar un entorno analítico aplicando tecnologías para el tratamiento masivo de datos.

HPE proporciona también otro paquete de software de similares características llamado Vertica, con características orientadas a la obtención de valor a través de grandes cantidades de datos, además de ayudar a ser más eficientes y adaptativos a cambios de entorno.

DENODO. La plataforma Denodo se integra con plataformas Big Data como Hewlett Packard Enterprise (HPE), Vertica y Hortonworks Data Platform (HDP). La plataforma Denodo permite a los clientes construir fácilmente data warehouses lógicos.

En relación a Cloud Computing alguno de los ejemplos actuales son:

Servicios de espacio en la nube como **Dropbox o Google Drive**; de mensajería como Slack; o paquetes de ofimática online como Microsoft Office 365 son algunos de los ejemplos más conocidos de **SAAS**.

Google App Engine o Microsoft Azure. Plataforma como Servicio (**PAAS**) donde se proporciona un entorno de desarrollo, facilitando la creación de aplicaciones y servicios web a través de internet.

Amazon Web Service, vCloud o Rackspace Cloud. “Infraestructura como Servicio” (**IAAS**)

2.7.1 Proyectos I+D

RePhrase - REfactoring Parallel Heterogeneous Resource-Aware Applications - a Software Engineering Approach [55]. Desarrollo de nueva metodología de ingeniería de software para el desarrollo de aplicaciones complejas, a gran escala, de uso intensivo de datos en paralelo, respaldada por un modelo de programación de muy alto nivel. La generalidad del enfoque se garantiza mediante la orientación C \\\ y los modelos de programación paralelos de bajo nivel más populares, como los estándares C \\\ 11/14/17, pthreads, OpenMP, Intel TBB, OpenCL y CUDA.

COCOA CLOUD - Collaborative CO-creation of web Applications on the CLOUD. Es una innovadora herramienta de desarrollo de aplicaciones que cambiará la forma en que se crean las aplicaciones web complejas y las interfaces de usuario diseñadas.

2.8 ENERGÍAS RENOVABLES

Todas las instalaciones generadoras de energía renovable (solar fotovoltaica, térmica, eólica, hidráulica, geotérmica, mareomotriz, undimotriz, etc.) trabajan con gran cantidad de datos. Los parques eólicos o las granjas solares, por ejemplo, tienen la capacidad de recoger cada vez más información, las agencias meteorológicas son capaces de predecir más variables de forma más precisa y el operador del sistema tiene la posibilidad de recabar cada vez más datos en un mundo cada vez más conectado.

Para poder gestionar y extraer información y conocimiento de los datos disponibles es necesario el empleo de técnicas de **Big Data Analytics y Cloud Computing**. Con estas herramientas podemos conseguir, por ejemplo:

- Analizar en tiempo real las variables de funcionamiento de los equipos
- Analizar en tiempo real los datos meteorológicos
- Extraer patrones de comportamiento de la instalación. Mantenimiento predictivo. Reducción de tiempos de parada, etc.
- Realizar predicciones que inciden directamente en la eficiencia y en los costes de la instalación
- Extraer de patrones de consumo
- Etc.

En el caso de la **energía eólica**, la realización de predicciones eólicas a corto y medio plazo tiene una incidencia directa en las labores de operación y mantenimiento de los parques. Igualmente, para acudir al mercado eléctrico es necesario disponer de predicciones horarias de producción con un día de antelación (en el mercado diario).

La realización de predicciones en el sector eólico depende de multitud de factores, uno de ellos es el viento y su carácter variable, lo que hace necesario el control de multitud de datos. Para extraer el conocimiento y la información necesaria de todo este gran volumen de datos que inciden en la producción, gestión, distribución, etc. se hace necesario la aplicación de técnicas de Big Data, como por ejemplo el Machine-Learning o aprendizaje máquina.

Los modelos que actualmente están siendo más estudiados en relación a las energías renovables son, por un lado los modelos de **Bosques Aleatorios o de Gradient Boosting** [30], y por otro lado, se está demostrando la **eficacia de las redes profundas** [31].

Como ejemplos de la utilización de todas estas tecnologías, podemos destacar:

A.U.R.A. GAMESA, S.A. Mantenimiento predictivo de maquinaria eólica. La plataforma de monitorización A.U.R.A. de NEM Solutions realiza diagnósticos expertos de la maquinaria eólica permitiendo un ajuste más fino del mantenimiento predictivo

EA2 [32]: sistema de predicción de producción de energía eólica, desarrollado por IIC (UAM), capaz de llevar a cabo la predicción horaria de parques individuales, pequeñas agrupaciones o áreas más amplias, que puede complementarse con Argestes Planner, una herramienta de visualización que permite, en tiempo real, analizar y comparar las predicciones realizadas.

EA2 está orientado operadores del sistema eléctrico, operadores de distribución, generadores de energía, comercializadoras o, en general, empresas relacionadas con la eficiencia energética que requieran técnicas de modelado y predicción de energía

El sistema se ofrece en modalidad Software as a Service (SaaS) que emite predicciones de producción eólica para parques en cualquier parte del mundo, adaptándose a las características de cualquier mercado. Gracias a su versatilidad se puede aplicar a un parque eólico o sobre un conjunto de parques como los de una granja, agrupación o clúster, o incluso sobre una gran área como la Península Ibérica. Para la elaboración de las predicciones EA2 utiliza técnicas de analítica predictiva y métodos de machine learning, como SVM redes neuronales.

Esta herramienta ha sido aplicada al parque eólico experimental Sotavento, en Galicia.



ILUSTRACIÓN 29. FUENTE: EA2. [HTTP://WWW.IIC.UAM.ES/SOLUCIONES/ENERGIA/EA2/](http://www.iic.uam.es/soluciones/energia/ea2/)

Aristoles de Kaiserwetter ha desarrollado un sistema que combina el Internet de las cosas, el despliegue de sensores, técnicas de análisis de Big Data y una infraestructura digital centralizada en la nube.

The Hybrid Renewable Energy Forecasting Solution (HyRef) de IBM, utiliza datos de equipos de monitoreo como cámaras que siguen el movimiento de las nubes, datos meteorológicos, sensores en aerogeneradores para monitorear velocidad, dirección y temperatura del viento, y realizar una predicción de condiciones hasta un mes en adelante. El análisis de turbulencia y la tecnología de imágenes en la nube se usan para predecir la generación de energía solar y eólica con precisión.



ILUSTRACIÓN 30. THE HYBRID RENEWABLE ENERGY FORECASTING SOLUTION (HYREF) DE IBM. FUENTE: IBM'S HYREF SEEKS TO SOLVE WIND'S INTERMITTENCY PROBLEM [HTTP://WWW.RENEWABLEENERGYWORLD.COM/ARTICLES/2013/08/IBMS-HYREF-SEEKS-TO-SOLVE-WINDS-INTERMITTENCY-PROBLEM.HTML](http://www.renewableenergyworld.com/articles/2013/08/ibms-hyref-seeks-to-solve-winds-intermittency-problem.html)

Intelligent Renewable Energy Performance Deep Analytics & Optimization iPAO de IBM proporciona evaluación y mejora de la eficiencia de trabajo, evaluación de amenazas y mantenimiento predictivo, recomendación de tipo de activos, optimización de piezas de repuesto a gran escala y optimización de planes O & M. Además, iPAO ayuda a las plantas de energía renovable a proporcionar toda la gestión de la operación del ciclo de vida, aumentar la eficiencia de los activos, ampliar la vida útil de los activos y mejorar su nivel de gestión.

Vi-POC (Virtual Power Operating Center) recopila variables de instalaciones fotovoltaicas, eólicas, cogeneración, biomasa, geotermia y de predicción del clima. El módulo de Big Data utiliza:

- Mondrian as OLAP Server
- Hive as Query Executor on Hadoop MapReduce
- HBase as NoSQL Data Storage

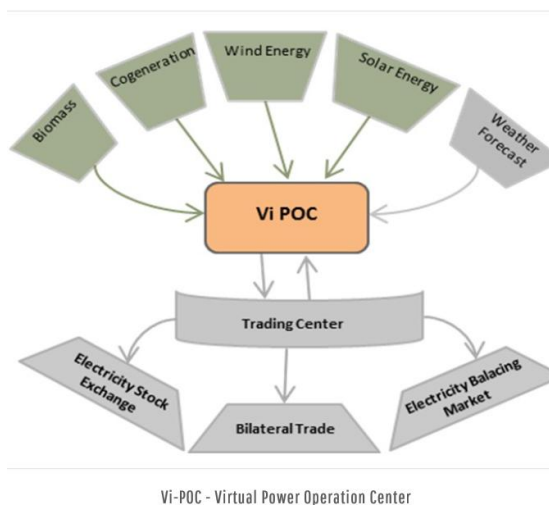


ILUSTRACIÓN 31: VI-POC (VIRTUAL POWER OPERATING CENTER) FUENTE: [HTTP://WWW.SMAU.IT/MILANO15/PARTNER_PRODUCTS/33555/](http://www.smau.it/milano15/partner_products/33555/)

Enervalis [57] ha desarrollado una plataforma de TI enfocada en optimizar el uso de la energía verde. Enervalis desarrolla software que proporciona soluciones de energía sostenible para vehículos eléctricos, edificios y microrredes. La plataforma monitorea las fuentes de energía y los usuarios disponibles, y puede predecir la demanda y el suministro de energía en el futuro a través de la predicción meteorológica, los aportes de los usuarios, la Inteligencia Artificial y Big Data.

2.9 PIEDRA NATURAL

Big Data como tecnología habilitadora y transversal de la industria 4.0 puede ser de aplicación a todos los sectores fabriles, entre ellos la piedra natural. El uso de los datos para la predicción, control de calidad, operación y toma de decisiones puede ser de utilidad en procesos como la **extracción del material, el proceso constructivo, e incluso la entrada de nuevos materiales inteligentes** que podrían entrar en competencia con la piedra natural.

Distinguimos como posibles aplicaciones:

- Extracción de patrones de comportamiento de la instalación. Mantenimiento predictivo. Reducción de tiempos de parada, etc.
- Planificación y asignación óptima de recursos de producción
- Predicción de fallos en cadenas de producción
- Optimización logística de suministro y distribución
- Pronóstico de la demanda
- Eficiencia energética en producción
- Despliegue óptimo de bienes y cadenas de producción (lay-outs)
- Seguridad en la planta
- Análisis de riesgos y predicción de fallos
- Optimización del stock
- Customización del producto: nuevos acabados y formas más ajustados a las necesidades del cliente
- Clasificación de materiales.

Como ejemplos de la utilización de esta tecnología, podemos destacar:

Antolini Luigi & C. S.p.A [59]: compañía especializada en la industria de piedra natural, ha desplegado la tecnología RFID, **StoneID**, para cuantificar su material almacenado y dar seguimiento a cerca de 10 mil bloques de diferentes materiales provenientes de todas partes del mundo y la localización exacta de alrededor de 900 mil losas producidas cada año. El proyecto abarcaba todo el proceso productivo y logístico de la empresa.

El bloque es etiquetado en su llegada a fábrica, a lo largo del proceso de fabricación y en el stock, de forma que en todo momento la empresa dispone de los datos necesarios para la identificación y localización del material (stock en tiempo real), y de los datos necesarios para llevar a cabo el procesamiento a través del ajuste automático de la maquinaria necesaria, evitando tiempos de espera. El tratamiento posterior de los datos permite una planificación optimizada de la oferta, la predicción del mantenimiento de los equipos y la eficiencia del proceso de fabricación.

Clasificación de Materiales: Metodología de Multiclasificación para la pizarra. [60] Nueva metodología de clasificación basada en clasificadores binarios, construidos utilizando máquinas de vectores de soporte y aplicando un enfoque de uno contra todos respaldado por el uso de gráficos acíclicos dirigidos.

Datos en Obra: la posibilidad de obtener datos en obra de forma que la información esté actualizada y sea precisa, es fundamental para la gestión de materiales en constante movimiento a lo largo de la vida de la obra. El establecimiento de Big Data en Obra tendrá una incidencia directa en:

- Reparación de equipamiento: Mitigación de fallos y diagnóstico remoto.
- Consumo de energía: soluciones que evitan gastos innecesarios en iluminación o que monitoricen la temperatura pueden aportar grandes ahorros.
- Gestión de personas y suministros: Control de stocks de materiales dispersos y de personal en obra.

Nuevos materiales: la introducción en materiales constructivos de sensores para la captación de datos físicos o del usuario, será un factor relevante para Big Data en el sector, en la medida en la que podremos acercarnos cada vez más a la experiencia del consumidor y a sus necesidades.

2.10 METALMECANICO

El sector metalmecánico puede beneficiarse de la mayor parte de las aplicaciones mencionadas para el resto de sectores analizados pues, en la mayoría de los casos, éste forma parte de la propia cadena de valor de sectores como los de automoción, aeronáutico, renovable o naval. Como por ejemplo:

- Extracción de patrones de comportamiento de la instalación. Mantenimiento predictivo. Reducción de tiempos de parada, etc.
- Planificación y asignación óptima de recursos de producción
- Predicción de fallos en cadenas de producción
- Optimización logística de suministro y distribución
- Pronóstico de la demanda
- Eficiencia energética en producción
- Despliegue óptimo de bienes y cadenas de producción (lay-outs)
- Seguridad en la planta
- Análisis de riesgos y predicción de fallos
- Optimización del stock
- Customización del producto: nuevos acabados y formas más ajustados a las necesidades del cliente

Como ejemplos de la utilización de esta tecnología, podemos destacar:

GESTAMP [61]. La plataforma Siemens de Big Data monitoriza las necesidades de consumo energético de Gestamp y conecta sus infraestructuras a una solución cloud. Este sistema permite definir algoritmos basados en los patrones de consumo para identificar y advertir sobre posibles fallos de los equipos. Los datos del consumo energético pueden ser procesados a través de técnicas de análisis de datos para definir de forma predictiva el mantenimiento, así como gestionar los procesos de producción o las previsiones de consumo energético en base a las necesidades de producción futuras.

2.10.1 Proyectos I+D

MC-SUITE - ICT Powered Machining Software Suite[62]. MC-SUITE quiere aumentar la productividad de la industria manufacturera, mejorando el desempeño de la simulación y el mecanizado, al correlacionar el modelado de procesos utilizando computación de alto rendimiento, y la monitorización de las máquinas.

3. CONCLUSIONES / IMPACTO EN LA INDUSTRIA

Tal y como hemos desarrollado en anteriores apartados, el beneficio último de Big Data / Big Data Analytics es la obtención de valor a partir de los datos. Big Data Analytics es la capacidad de extraer valor de Big Data, sin tener la certeza de que ese valor exista. Esto implica estar expuestos a la volatilidad y el ruido que pueden tener estos datos durante su procesamiento. Partiendo de este hecho los principales retos de Big Data y de Big Data Analytics están relacionados íntimamente con la Captura, Almacenamiento, Transmisión, Procesamiento, Análisis, Visualización, Seguridad, Escalabilidad, Desempeño y Consistencia de los datos.

3.1 RETOS

A continuación, se describen brevemente algunos de los **desafíos más relevantes**, en los cuales se centra gran parte de las investigaciones académicas:

- **Captura y almacenamiento de datos:** La distribución (alta concurrencia y el manejo de un alto volumen de operaciones por cada servidor), replicación, migración, desempeño, confiabilidad y escalabilidad.
- **Transmisión de los datos:** La transmisión de Big Data en las tecnologías orientadas a la nube y a sistemas distribuidos tiene diferentes retos a abordar, entre los que se destacan: el ancho de banda, la seguridad y la integridad de los datos.
- **Procesamiento de los datos:** desde la identificación de los datos hasta su recuperación, aseguramiento de calidad, adición de valor, reutilización y preservación en el tiempo. El inconveniente con las herramientas tradicionales (principalmente los modelos que trabajan con datos estructurados) es que éstas no tienen la capacidad de manejar estos procesos con Big Data de forma eficiente. Por lo tanto, el análisis de Big Data debe incluir técnicas innovadoras que van desde la captura, representación y almacenamiento de los datos hasta su visualización después de los procesos de análisis, teniendo en cuenta que dichas técnicas deben realizarse a bajo coste.
- **La seguridad, privacidad y propiedad de los datos.** Los entornos en los que se mueve Big Data son entornos en la nube o manejan arquitecturas de datos compartidos. Este hecho hace que las organizaciones tiendan a desconfiar de dichos entornos y ralenticen los procesos de adaptación de estas tecnologías en las empresas. Los retos son numerosos y tienen que ver por un lado con aspectos legales, reputación, seguridad nacional, competitividad, secretos de industria, etc. y por otro con estructuras de información que puedan facilitar datos públicos, como los historiales clínicos o los perfiles académicos.
- **Análisis de datos:**
 - **Obtener una población correcta de los datos,** obtención de datos limpios (libres de ruido) y veraces (que puedan ser verificables), con el fin de evitar hechos falsos que puedan alterar la percepción de la realidad.
 - **Interpretar los datos,** realizar proyecciones y tendencias a partir de los datos recolectados. Este reto tiene que ver en gran parte con el anterior, ya que sin los datos adecuados, las interpretaciones posiblemente serán incorrectas.
 - **Definir y detectar anomalías.**
- **Arquitectura de los datos:** Se plantean una serie de inquietudes:
 - Determinar cómo la integración de servicios en la nube permite el manejo de Big Data.
 - Facilitar el escalamiento de sistemas a través de Cloud Computing.

Desarrollo de Frameworks genéricos de gestión de workflows de Big Data, con el objetivo de implementar sistemas flexibles y reutilizables en las diferentes organizaciones, en forma de servicios (XaaS, donde X pueden ser estos Frameworks).

- **Visualización de los datos:** El principal objetivo de la visualización de los datos es la representación del conocimiento para el entendimiento del ser humano. En el campo de Big Data, el reto principal frente a la visualización se encuentra en el gran tamaño y diversidad de los datos. Actualmente, la mayoría de las herramientas de visualización deben enfrentarse a inconvenientes de desempeño y escalabilidad. Esto hace que las técnicas y tecnologías que se vienen utilizando por años deban re-pensarse para abordar estos retos de una forma más efectiva.

A pesar de las múltiples oportunidades que se vislumbran, **las empresas aún están lejos de aprovecharla al 100% de las capacidades que el Big Data es capaz de ofrecer**. El verdadero reto no está tanto en crear las arquitecturas que permitan implementar Big Data (y su posterior análisis) en las organizaciones, si no en ser capaces de separar, en el menor tiempo de respuesta posible, lo que es relevante de lo que no aporta valor en todo el volumen de datos que se genere. La mayoría de los autores plantean que las organizaciones deben determinar los objetivos de análisis para luego decidir qué datos almacenar y cómo serán devueltos en forma de información valiosa. Las organizaciones que no se planteen objetivos claros desde el principio, simplemente almacenarán datos que no sabrán cómo aprovechar en un futuro, dando lugar a frustraciones y al abandono de esta tecnología. El informe de IDC “Big Data. Retos y Oportunidades” para Europa, establece como principales inhibidores de la implantación de estas metodologías en las empresas los reflejados en la siguiente ilustración.



Fuente: IDC BDA Pulse Survey

ILUSTRACIÓN 23: INHIBIDORES PARA LA ADOCIÓN DE BIG DATA. FUENTE: IDC BDA PULSE SURVEY

En relación a las herramientas, propiamente dichas de Big Data, todavía existen numerosos retos a los que enfrentarse:

- Medir el rendimiento a través de los distintos nodos operativos.
- Determinar, entre tanto componente co-dependiente, si el rendimiento es óptimo o no, y por qué.
- Medir el rendimiento del hardware y las múltiples conexiones necesarias para el funcionamiento de estos sistemas.
- Monitorizar las infraestructuras de Big Data y TI en una misma arquitectura.
- Integrar datos clave en la infraestructura de Hadoop desde servidores propios.

Un número creciente de empresas está utilizando la tecnología para almacenar y analizar ingentes cantidades de datos, incluyendo registros web, datos de flujo de clics, y contenido de redes sociales o incluso una combinación de diferentes fuentes. Esta cantidad de información es utilizada por las empresas para conseguir un **mayor conocimiento de sus clientes y de sus negocios**. Como resultado a esta nueva visión, la clasificación de la información (niveles de privacidad) se vuelve aún más crítica. A continuación se describirán brevemente algunos de los **principales problemas de privacidad y seguridad** que se generan en los entornos Big Data.

La información recogida en los entornos Big Data proviene principalmente de las redes sociales, entornos bancarios y los registros médicos. Estos entornos contienen información altamente privada y por tanto susceptible de ser tratada con mayor seguridad.

Además, en estos entornos puede haber más tipos de actores que sólo proveedores y consumidores, principalmente propietarios de datos, como los usuarios móviles y los usuarios de redes sociales. Los actores no son meros usuarios que reciben información, también pueden ser dispositivos que recolectan más información de diferentes fuentes para otros consumidores de datos diferentes.



ILUSTRACIÓN 24: CATEGORIZACIÓN DE FUENTES DE DATOS. FUENTE: [HTTP://WWW.SOFOSCORP.COM/BIG-DATA/](http://www.sofoscorp.com/big-data/)

El gran volumen de datos que maneja Big Data requiere su almacenamiento en diferentes medios, algunos de los cuales pueden almacenar datos agregados. La agregación y el movimiento de dicha información entre aplicaciones, pueden provocar la **pérdida total o parcial de la misma**, abriendo de esta forma otra puerta a violaciones de seguridad y privacidad.

Por tanto, la seguridad y privacidad son importantes tanto para la calidad de los datos como para la protección del contenido. La información en entornos Big Data con frecuencia se mueve de límites individuales a colectivos, hacia una comunidad de interés, estado, fronteras nacionales e internacionales. La procedencia, aborda el problema de la comprensión de la fuente original de los datos y de lo que se ha hecho con ellos e incluye el aseguramiento de la información de los métodos a través de los cuales se recogió la información. Por ejemplo, cuando se recibe información de sensores, es necesario rastrear la calibración, la versión, el muestreo y la configuración del dispositivo.

La **propiedad del dato** como característica universal debe abordarse en el contexto de la seguridad y la privacidad de Big Data. La propiedad es una característica (que puede o no ser visible para los usuarios) que vincula a los datos con una o más entidades que poseen o pueden influir en lo que se puede hacer con los datos (por ejemplo, las entidades bancarias pueden influir pero no pueden cambiar el historial de crédito). En las bases de datos, la propiedad confiere los privilegios para crear, leer, actualizar y eliminar datos. La transparencia de la propiedad permite la confianza y el control de los propietarios de los datos, así como la apertura y la utilidad para las empresas y la sociedad. El mantenimiento de la procedencia de los datos permite la trazabilidad a lo largo del ciclo de vida de los datos y controla la propiedad y el cambio de los mismos.

Los **frameworks de programación** distribuidos fueron desarrollados teniendo en cuenta el volumen y la velocidad de acceso, pero no se diseñaron para tener en cuenta la seguridad. Por ejemplo, aquellos nodos que no funcionan correctamente pueden llegar a perder datos confidenciales. También, ataques a parte de la infraestructura podrían comprometer una gran parte del sistema debido a los altos niveles de conectividad. Si el sistema diseñado no gestiona la autenticación entre los nodos distribuidos, es posible la intromisión de nuevos nodos no autenticados en la infraestructura que pueden extraer datos.

La búsqueda y selección de datos también pueden generar nuevos problemas relacionados con la privacidad o la política de seguridad como la **pérdida de datos en el proceso de búsqueda y selección**. Es probable que se necesite una combinación de las competencias del usuario y las protecciones del sistema, incluida la exclusión de las bases de datos que permiten la re-identificación.

Debido a que puede haber procesos de procesamiento dispares entre el propietario de los datos, el proveedor y el consumidor de datos, **debe garantizarse la integridad de los mismos**. Las prácticas de aseguramiento de información de extremo a extremo para Big Data -por ejemplo, para verificabilidad- no difieren de otros sistemas, sino que deben diseñarse a mayor escala.

Redefinir la seguridad de las bases de datos relacionales tradicionales hacia las **bases de datos no relacionales** supone un gran reto, ya que estos sistemas no han sido diseñados teniendo en cuenta la seguridad, delegando este problema a la creación de un middleware.

El movimiento y la agregación de datos entre aplicaciones provocan el análisis sistemático de posibles amenazas y con ello la investigación y desarrollo continuo de nuevas técnicas para ofrecer sistemas más seguros. Las amenazas que sufren los sistemas distribuidos incluyen los siguientes escenarios principales: **confidencialidad e integridad, procedencia, disponibilidad, consistencia, colusión, ataques de retroceso y disputas de registros**.

Otro de los principales problemas que surgen en términos de seguridad y privacidad es el elemento humano. Al igual que el resto de problemas comentados con anterioridad, el **elemento humano** en los sistemas Big Data también generará nuevos problemas. A medida que se disponga de más datos a través

de los motores de análisis, habrá más "analistas" que puedan acceder a dichos datos y por tanto generar más problemas para preservar la seguridad y la privacidad. De manera similar, es probable que los analistas tengan acceso a datos cuya procedencia desconozcan.

Por otro lado, las medidas de seguridad y privacidad en Big Data deben escalar de forma no lineal. Deberán surgir nuevas regulaciones para abordar los riesgos detectados en entornos reales y percibidos a medida que los usuarios y los reguladores tomen conciencia de las capacidades de Big Data. El aseguramiento de la información, generará diferentes especializaciones dentro de la informática.

3.2 PERSPECTIVAS A MEDIO Y LARGO PLAZO

En los próximos años, continuará el crecimiento de los sistemas que admiten grandes volúmenes de datos, tanto estructurados como no estructurados. El mercado exigirá plataformas que faciliten a los responsables de los datos las tareas de administración y seguridad de los Big Data. Además, estas plataformas deberán permitir a los usuarios finales poder analizar dichos datos. Los sistemas deberán madurar para funcionar correctamente en el marco de los sistemas y estándares empresariales de las Tecnologías de la Información.

Los sistemas Big Data se volverán más rápidos y flexibles. Poniendo a disposición de los usuarios herramientas de aprendizaje automático / aprendizaje máquina (Machine Learning), de forma que a medida que las máquinas aprendan y los sistemas se vuelvan inteligentes, las tendencias se centrarán en los proveedores de software de autoservicio.

Las plataformas y servicios distribuidos, Hadoop, Spark, Graph, análisis de datos en flujo continuo, aprendizaje de máquinas/machine learning, transformación de datos, procesamiento de lenguaje natural, visualización de datos y otras funciones y procedimientos, se adoptarán ampliamente como bloques de ensamblaje para aplicaciones analíticas personalizables. La naturaleza distribuida de estos servicios también permitirá la analítica en dispositivos y la gestión de la información en casos de uso como Internet de Cosas (IoT) y robótica

De una forma más concreta, podemos identificar las siguientes tendencias por tecnología o herramienta Big Data.

Apache Hadoop

Una tendencia en aumento es la transformación de Hadoop en una parte fundamental del entorno de TI empresarial.

La principal tendencia que se puede observar con respecto a la implementación de Hadoop para análisis de Big Data es la transformación de este framework y sus diferentes componentes relacionados, en arquitecturas para el análisis de datos de cualquier tipo de industria (o al menos las más representativas). Estas arquitecturas se encuentran apoyadas por la computación en la nube, haciendo posible la habilitación de estas plataformas de análisis de datos como servicios.

NoSQL y Sistemas Híbridos

NoSQL, continuará siendo tendencia en la medida que se generen datos no estructurados a grandes velocidades. Hoy en día, las tendencias apuntan a sistemas híbridos entre SQL y NoSQL para tomar lo mejor de cada sistema manejador, como es el caso de HadoopDB [25]. HadoopDB combina Hadoop con

PostgreSQL [26] teniendo a PostgreSQL como capa de datos, Hadoop como capa de comunicación y almacenamiento, y Hive como capa de traducción de SQL a MapReduce.

Existen otras arquitecturas implementadas generalmente como sistemas RDBMS paralelos capaces de conectarse con Hadoop para la carga de datos y la ejecución de tareas MapReduce. La mayoría de estas soluciones ofrecen una especie de semántica MapReduce-SQL nativo. Las tres representaciones más destacadas de este estilo arquitectónico son Pivotal Greenplum [27], Aster Data – Teradata y HP Vertica [28].

Data Analytics as a Service (DAaaS)

El desarrollo de estas plataformas continuará siendo imparable en los próximos años.

DAaaS son plataformas de analítica extensible, a través de un modelo basado en la nube, donde se encuentran disponibles varias herramientas configurables para análisis de datos. La idea con este tipo de plataformas es que las organizaciones cliente las alimenten con datos empresariales y obtengan información concreta y útil para la toma de decisiones. Esta información es generada por aplicaciones analíticas, las cuales generan flujos de trabajo específicos para análisis de datos, utilizando colecciones de servicios que implementan algoritmos analíticos. Una característica de estas plataformas es que son extensibles, lo que permite manejar diferentes casos de uso o áreas de aplicación (venta al por menor, telecomunicaciones, salud, administración pública). El principal beneficio que se obtiene de DAaaS es la reducción de la barrera de entrada relacionada con las capacidades analíticas avanzadas por parte de organizaciones, que apenas se encuentran introduciéndose en el mundo de Big Data Analytics. Esto permite a estas organizaciones concentrarse más en sus KPIs (el qué) que en la forma como se obtendrán (el cómo).

Compresión de Datos

A pesar de que los costos de almacenamiento de datos se han venido reduciendo, el enorme crecimiento en el volumen de los datos hace que el almacenamiento sea uno de los elementos más costosos. Las tecnologías actuales de compresión de datos utilizan una combinación de métodos orientados a filas y a columnas que permiten almacenar datos para ahorrar espacio y mejorar el desempeño.

In-database Analytics

Este término hace referencia a las técnicas de análisis de datos que son aplicadas directamente en los DBMS. Esto permite eliminar la necesidad de mover datos entre servidores, optimizando el data warehousing y reduciendo costos de implementación. El hecho de no tener que mover los datos hacia otras fuentes de almacenamiento para su análisis, permite a los analistas obtener información valiosa en mejores tiempos a costes más bajos. Adicionalmente, esto permite apuntar hacia atributos de calidad como la seguridad, escalabilidad y desempeño. Las principales compañías que suministran soluciones de data warehousing, en la actualidad, incluyen análisis “In-database” como una de sus alternativas. Como por ejemplo: Teradata, Oracle, IBM Netezza, Pivotal Greenplum, Sybase, ParAccel (Actian), SAS y Exasol.

Arquitecturas caracterizadas por su temporalidad

Las tendencias se centrarán en el desarrollo de arquitecturas en función de la temporalidad del análisis requerido, como las plataformas para análisis Batch, plataformas para análisis Interactivo y plataformas para análisis de datos de Streaming.

Arquitecturas caracterizadas por el almacenamiento y representación de los datos

Las principales tendencias en cuanto a estas arquitecturas son:

- **Bases de Datos distribuidas.** La distribución de las bases de datos es una tendencia que se seguirá observando junto con otras que tiene que ver con la forma cómo se comparte el almacenamiento y con las condiciones del teorema de CAP.
- **Bases de Datos “In-memory”:** Actualmente son utilizadas tanto para sistemas transaccionales como para sistemas analíticos interactivos y de streaming, donde la latencia y el tiempo de respuesta son críticos. Estas bases de datos se implementan frecuentemente como modelos relacionales sin logging o como modelos clave-valor en tablas o mapas hash.
- **Linked Data Oriented (LOA):** Esta forma de almacenamiento y organización de los datos será bastante utilizada en los próximos años por la facilidad con la que se accede al conocimiento y como se representa.

Arquitecturas caracterizadas por la plataforma para el cómputo de los datos

- **Computación Granular (Granular Computing):** Se encuentra basada en el manejo eficiente de Big Data a través de la utilización de gránulos tales como clases, clústeres, subsets, grupos e intervalos para separar los datos, etc., con el fin de reducir el volumen de los mismos en diferentes grados de granularidad, permitiendo aplicar distintos algoritmos de minería de datos.
- **Computación en la Nube (Cloud Computing):** Es una de las tendencias más comunes en Big Data. El beneficio más importante que ha brindado esta forma de computación es la reducción de costos, permitiendo a las empresas mayores capacidades de procesamiento sin ampliaciones relevantes en los presupuestos de infraestructura.
- **Computación Bio-inspirada (Bio-inspired Computing).** Consolida técnicas y modelos que se han venido estudiando en los últimos años, sobre la forma en la que el cerebro humano y, en general, la naturaleza, almacena, organiza y procesa los datos. Estos modelos son más apropiados para Big Data ya que tienen mecanismos más eficientes para organizar, acceder y procesar datos que los modelos tradicionales. Pueden ser utilizados tanto para el diseño de software como de hardware y comienzan a ser una tendencia, aunque por ahora experimental y con costos muy elevados.
- **Computación Cuántica (Quantum Computing):** Se basa en la utilización de computadores y componentes cuánticos, capaces de incrementar exponencialmente la capacidad de memoria y procesamiento de los computadores tradicionales. Esto resultaría bastante útil para el análisis de Big Data por el desempeño la gran reducción de tiempos que se podría obtener en el procesado y análisis de los datos. Sin embargo, tiene la desventaja de ser extremadamente costoso por ahora.

Machine Learning y Deep Learning

A medida que Machine Learning y Deep Learning alcancen la madurez, las máquinas irán adquiriendo habilidades de pensamiento, de resolución de problemas y de entendimiento del lenguaje.

Los sistemas capacitados para Machine o Deep Learning facilitarán mejores servicios y experiencia al cliente, gestionarán la logística, analizarán registros médicos, etc. La mayor parte del valor potencial de estas herramientas se encuentra aún por descubrir. Estas nuevas tecnologías incrementarán los **ratios de productividad y la calidad de vida**. Según una investigación de MGI [29] “A future that works: automation, employment, and productivity”, machine learning puede ser el habilitador de la automatización del 80% de las actividades empresariales. Los descubrimientos en procesamiento de lenguaje natural pueden hacer que este efecto sea aún mayor.

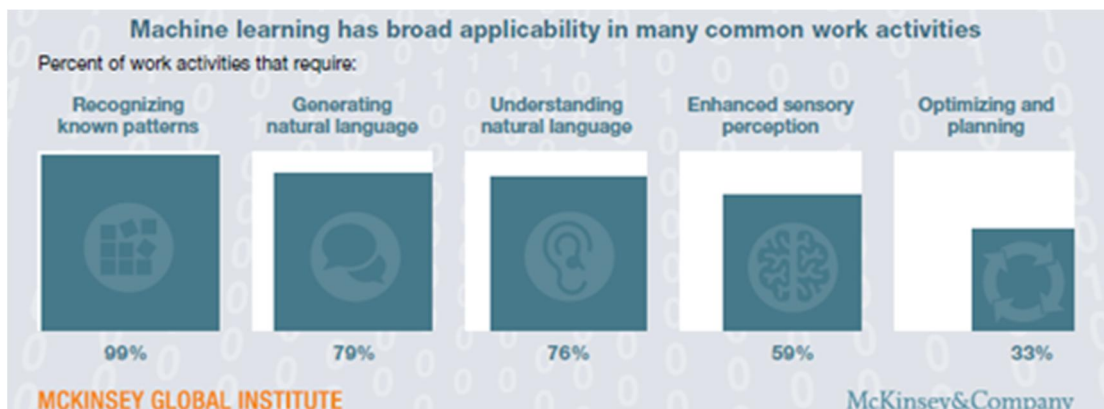


ILUSTRACIÓN 31: APLICACIÓN DE MACHINE LEARNING EN TRABAJOS HABITUALES. FUENTE: THE AGE OF ANALYTICS. COMPETING IN A DATA-DRIVEN WORLD. MCKINSEY GLOBAL INSTITUTE

El uso de **Machine Learning** junto con otras técnicas tendrá enormes rangos de uso, tal y como se refleja en la siguiente ilustración, tomada del MGI [29].

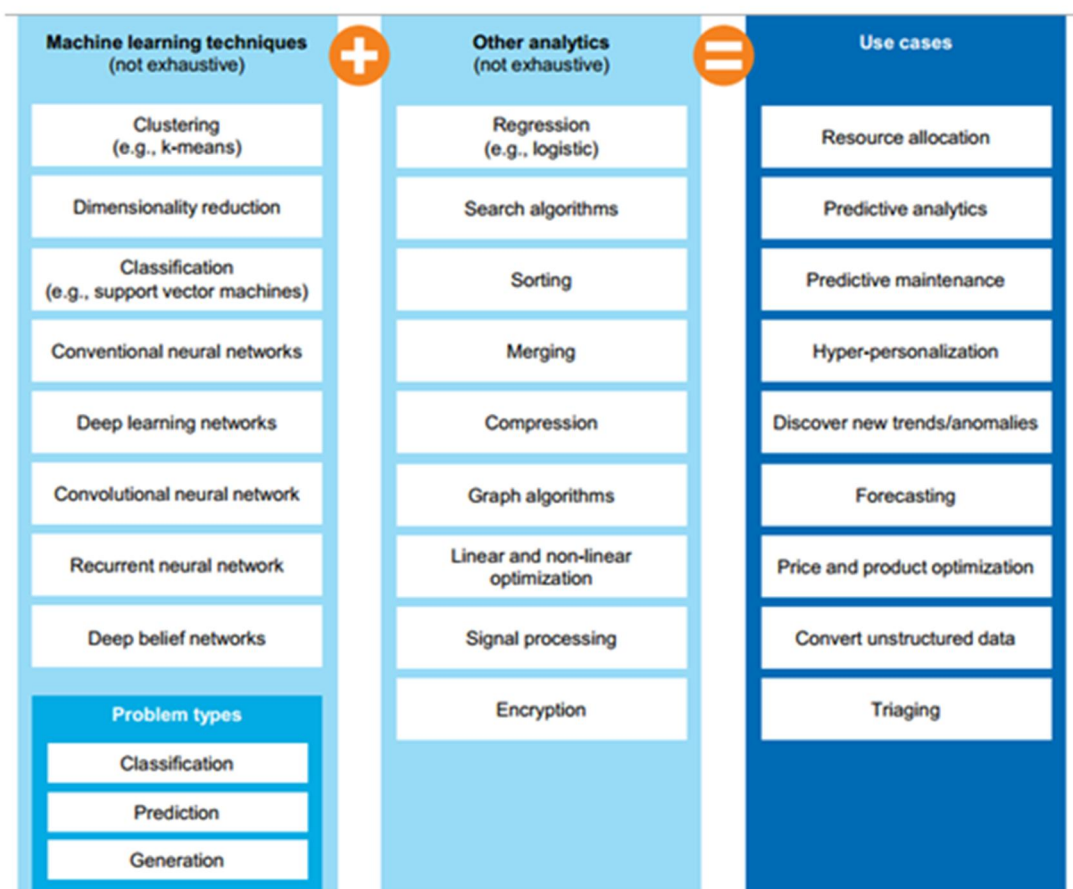


ILUSTRACIÓN 32: IMPACTO DE MACHINE LEARNING COMBINADO CON OTRAS TECNOLOGÍAS. FUENTE: MCKINSEY GLOBAL INSTITUTE ANALYSIS

Deep Learning, está todavía en la frontera del conocimiento, utilizando las redes neuronales para incrementar las capacidades de las máquinas. Los últimos avances científicos se centran en el uso de deep learning para reconocimiento de objetos y caras, así como en la generación de lenguaje.

3.3 CONCLUSIONES

3.3.1 Políticas De Apoyo

A nivel europeo, nacional y regional, se están llevando a cabo numerosas iniciativas con el fin de acelerar el proceso de adaptación de Europa al nuevo paradigma tecnológico. De esta forma, se han iniciado importantes proyectos como **The Digital Single Market y Digital Agenda for Europe 2020**, con los que la Unión Europea pretende crear las condiciones necesarias para el futuro crecimiento económico. Como parte de este esfuerzo, la Comisión Europea también ha lanzado un partenariado público-privado (PPP Big Data), con el fin incrementar la posición europea en la economía digital.

A nivel nacional España ha elaborado la **“Agenda Digital”** y a nivel regional **“A Axenda Galega”**.

Digital Agenda for Europe 2020, aprobada en 2010, sienta las bases para impulsar la economía europea a través de los beneficios económicos y sociales del mercado único digital.

La Agenda Digital para España, aprobada en 2013, establece la estrategia del Gobierno para desarrollar la economía y la sociedad digital en nuestro país. Esta estrategia se configura como el paraguas de todas las acciones del Gobierno en materia de Telecomunicaciones y de Sociedad de la Información.

La Agenda Digital de Galicia 2020, se desarrolla para crear una estrategia coordinada y alineada con las estrategias de ámbito nacional y europeo, introduciendo elementos que permitan maximizar el impacto de las políticas tecnológicas.

3.3.2 Impacto Económico

Según los informes de IDC “Q4 2015: Worldwide Server and Storage in Big Data Forecast, 2015–2019” y “Big Data. Retos y Oportunidades”, el avance del impacto de estas tecnologías en la economía se prevé de la siguiente forma:

- la **cuota de mercado** en EMEA, en servidores para Big Data crecerá de un 5,9% en 2015 a un 15,8% en 2019,
- el **valor económico** del mercado de servidores pasará de 1000 millones de dólares en 2015 a 2700 millones en 2019,
- la **cuota en capacidades** almacenamiento alcanzará el 19,5% (19,8 exabytes) en 2019, con un valor de 2700 millones de dólares,
- en 2019, se prevé que el **gasto** en soluciones de Big Data y Analítica sobre Cloud crecerá 4,5 veces más rápido que las que están alojadas en entornos bajo premisa. Y se observa un claro crecimiento de la utilización de la Cloud Pública para soluciones de Analítica de Datos en EMEA.
- se ha previsto un **crecimiento del volumen de los datos** no estructurados de un 80% en 2016 respecto al año anterior en EMEA.

Resumiendo datos recogidos del informe IDC FutureScape: Worldwide Big Data and Analytics 2016 Predictions:

- El gasto en la tecnología Big Data Analytic basada en la nube crecerá **4,5 veces más rápido** que el gasto en soluciones locales. La tecnología de código abierto representará el núcleo de esta nueva arquitectura.
- El **50% del software** de análisis de negocios incorporará análisis predictivos basados en computación cognitiva.
- El **gasto en herramientas de preparación de datos** y visualización de autoservicios crecerá 2,5 veces más rápido que las herramientas tradicionales.
- Los **esfuerzos de monetización** de datos darán lugar a que las empresas sigan las iniciativas de transformación digital aumentando el número de sus propios datos en 100 veces o más.
- Las organizaciones capaces de analizar todos los datos relevantes y obtener información de valor, lograrán 430\$ mil millones más en **beneficios de productividad** sobre sus competidores menos orientados analíticamente.

Las nuevas tecnologías relacionadas con el Big Data y el análisis de datos han originado la aparición de **oportunidades de negocio y nuevos perfiles de trabajadores**.

IDC, en su informe “European Data Market Smart 2013/0063. D8 — Second Interim Report. Junio 2016., analiza el impacto de Big Data basándose en la tendencia de indicadores relacionados con Data Trabajadores, Data Compañías y Data Usuarios. Para el análisis de los indicadores, IDC realiza comparativas

en 3 escenarios denominados “Baseline (Línea Base-Escenario Conservador), “Challenge (Desafío - Escenario Medio)”, “High Growth (Alto crecimiento – Escenario más favorable).

Data trabajadores

Se entiende por data trabajadores la mano de obra dedicada a recolectar, almacenar, gestionar y analizar datos como la principal actividad de su trabajo.

La tendencia actual muestra una tendencia creciente en el número de “Data Trabajadores”. Con una previsión a 2020, según el mencionado estudio de ECI, de 6.6 millones en 2020 en el escenario Challenge, 7.3 millones en el escenario Baseline, y 9.3 millones en el escenario High Growth. La tendencia general, muestra un sólido y constante crecimiento de los Data trabajadores, en cualquiera de los tres escenarios.

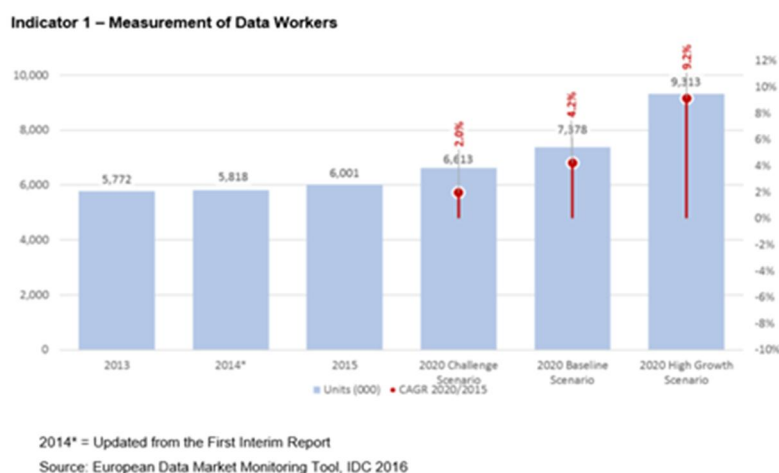


ILUSTRACIÓN 16: CRECIMIENTO EN LOS PRÓXIMOS AÑOS DEL NÚMERO DE DATA TRABAJADORES.

FUENTE: EUROPEAN DATA MARKET MONITORING TOOL. IDC 2016

Además, no solo serán necesarios los Científicos de los Datos, si no gestores que sepan interpretar esos datos en función del know-how empresarial, los llamados “data-literate managers”.

Data Compañías

Se entiende por data Compañías las organizaciones cuya principal actividad es la producción y entrega de productos, servicios y tecnologías digitales.

El crecimiento potencial en EU de las Data Compañías y compañías relacionadas, es muy alto, pudiendo llegar a alcanzar 360,000 unidades (escenario – high) en 2020, en el territorio europeo. Este alto crecimiento es debido al impulso de las inversiones en Investigación y Desarrollo, y la continua innovación en tecnologías de datos.

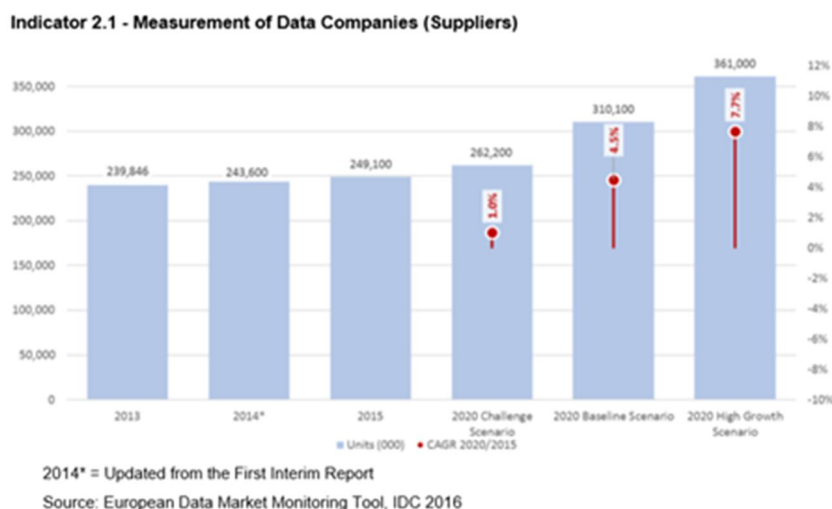


ILUSTRACIÓN 17: CRECIMIENTO EN LOS PRÓXIMOS AÑOS DEL NÚMERO DE DATA COMPAÑÍAS. FUENTE: EUROPEAN DATA MARKET MONITORING TOOL. IDC 2016

Data Usuarios

Se entiende por Data Usuarios a las organizaciones que generan, explotan, capturan y analizan datos para mejorar su negocio. Se prevé un sólido incremento de las compañías usuarias en los próximos años, tal y como puede verse en la siguiente figura.

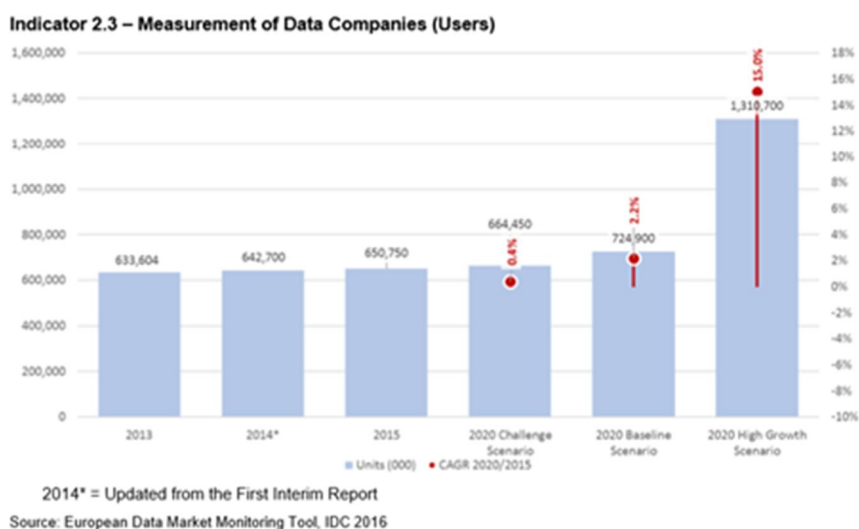


ILUSTRACIÓN 18: CRECIMIENTO EN LOS PRÓXIMOS AÑOS DEL NÚMERO DE DATA USUARIOS. FUENTE: EUROPEAN DATA MARKET MONITORING TOOL. IDC 2016

3.3.3 Impacto Industrial

La convergencia de distintas tecnologías y herramientas está acelerando el proceso de explosión de Big Data Analytics y Cloud Computing en las empresas. Big Data Analytics ha modificado la dinámica de trabajo en

muchas organizaciones. Sin embargo, a día de hoy existen multitud de oportunidades sin explorar y sin explotar.

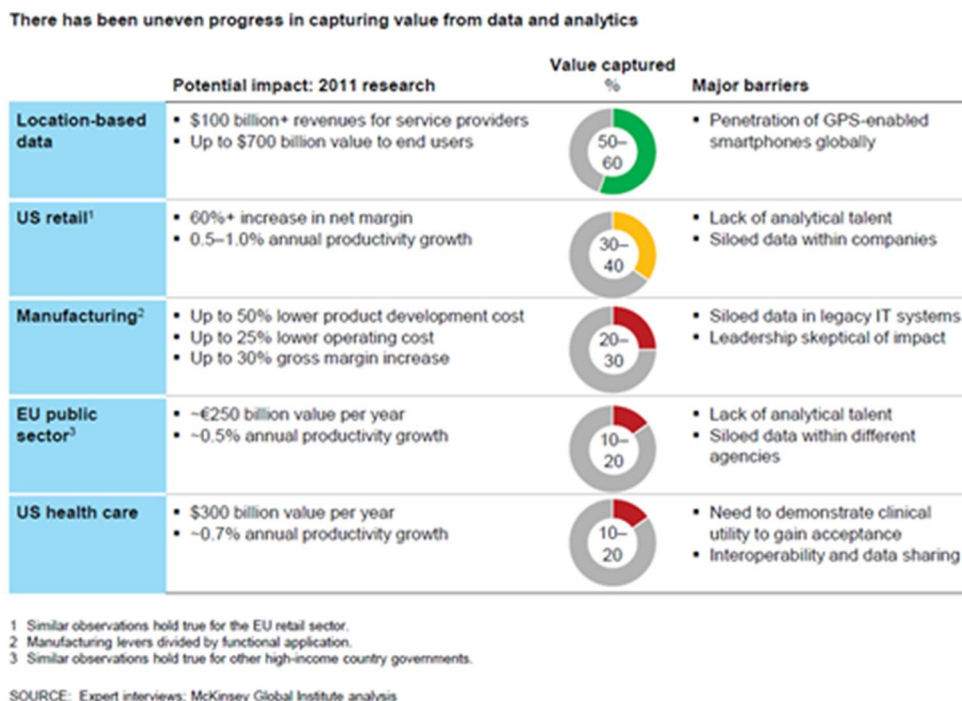


ILUSTRACIÓN 19: IMPACTO POTENCIAL DE LA IMPLANTACIÓN DE BIG DATA ANALYTICS EN VARIOS SECTORES Y BARRERAS QUE LA RETRASAN. FUENTE: MCKINSEY GLOBAL INSTITUTE

Así mismo, y siguiendo el mismo estudio de MGI, no todo el impacto potencial establecido por MGI en 2011 ha sido conseguido, **siendo la industria uno de los sectores que va más rezagado** en comparación con otros como: Sanidad, Sector Financiero o Venta al por menor.

En este estudio se establece que, por ejemplo:

- En **Investigación y Desarrollo**, el uso de Big Data en ingeniería concurrente o gestión del ciclo de vida del producto, ha conseguido reducciones en el coste de un 10-30%, frente al 20-50% esperado por MGI en 2011.
- En la **Producción**, el uso de “factorías Digitales”, sensores y analítica aplicada ha supuesto un descenso en los costes de operación de 10-15% frente al 10-25% previsto por MGI en 2011.

En la próxima ilustración puede verse, de forma sectorial, el potencial que esta nueva era plantea para todos los sectores:

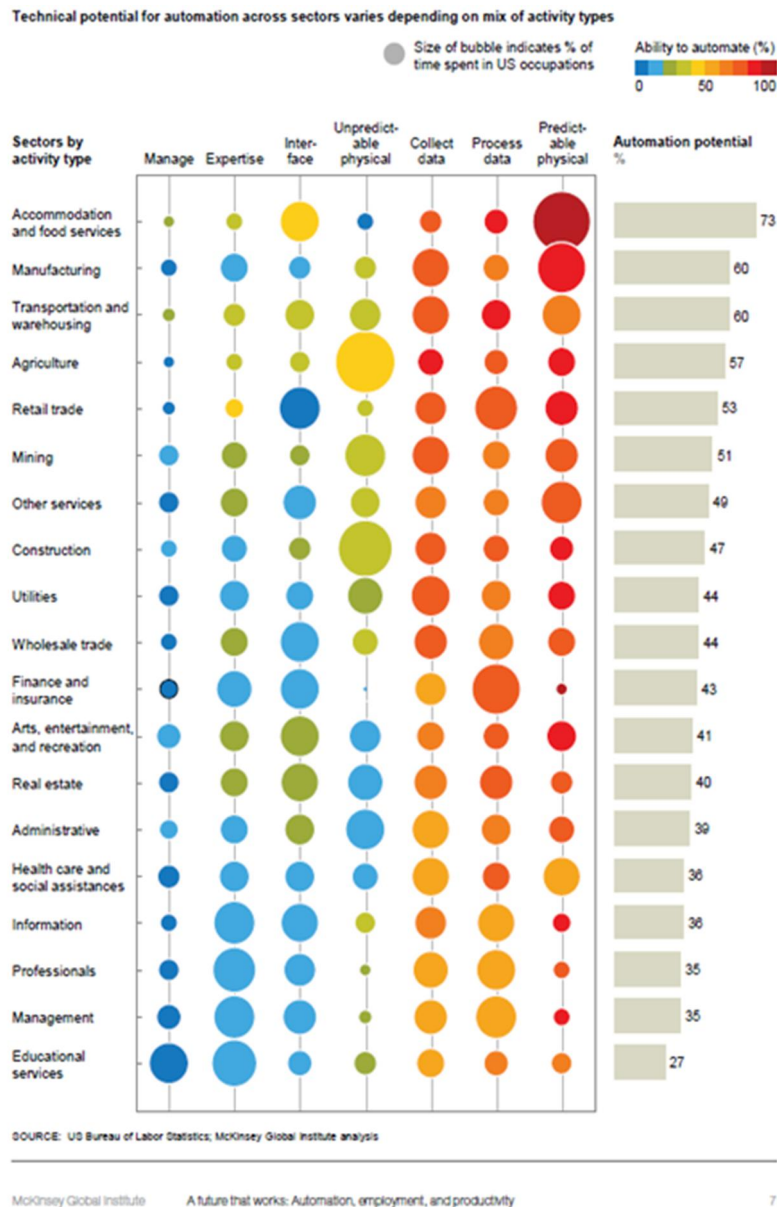
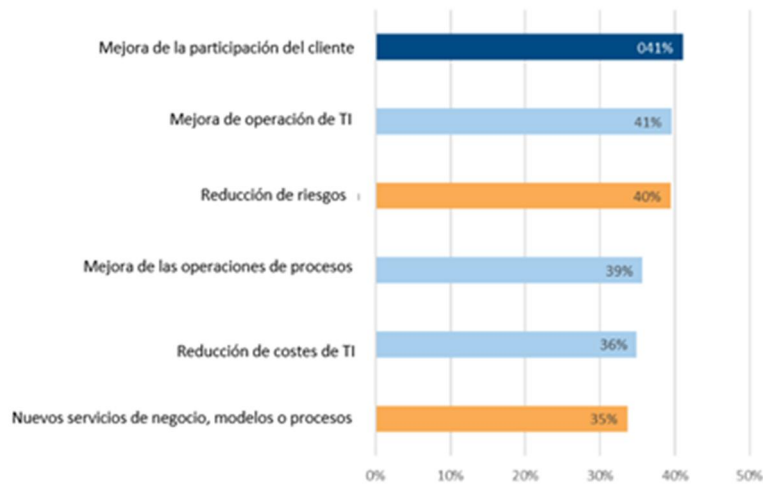


ILUSTRACIÓN 20: POTENCIAL TÉCNICO DE AUTOMATIZACIÓN DE LOS SECTORES PRODUCTIVOS. FUENTE: MCKINSEY GLOBAL INSTITUTE

Así mismo, serán las **soluciones End to End**, desde la captura de la información hasta la toma de decisión, las que tomen fuerza, teniendo en cuenta que uno de sus elementos claves deberá ser la necesidad de reducir la latencia.

En relación a los casos de uso en Europa, el informe de IDC “Big Data. Retos y Oportunidades”, establece que, “el caso de uso relativo a la **mejora de la participación del cliente**, es el más implantado, mientras que la mejora de las operaciones de procesos, se encuentra en un cuarto lugar”.

P. ¿Cuáles son los objetivos actuales y previstos por su organización alrededor de Big Data?



Fuente: IDC Market Analysis Perspective: European Big Data and Analytics Software, 2016

ILUSTRACIÓN 21: PERSPECTIVA ANÁLISIS DE MERCADO EN BIG DATA ANALYTICS. FUENTE: IDC MARKET ANALYSIS PERSPECTIVE: EUROPEAN BIG DATA AND ANALYTICS SOFTWARE, 2016

MGI prevé, además, una nueva ola de impacto en la economía derivada de la maduración de las tecnologías relacionadas con **Machine Learning y Deep Learning**. Tras el estudio realizado con 12 grupos de industria, identificándose 300 casos de uso sobre los que valorar la utilización de Machine Learning, MGI concluyó que el potencial de Machine Learning podría resumirse en la siguiente figura:

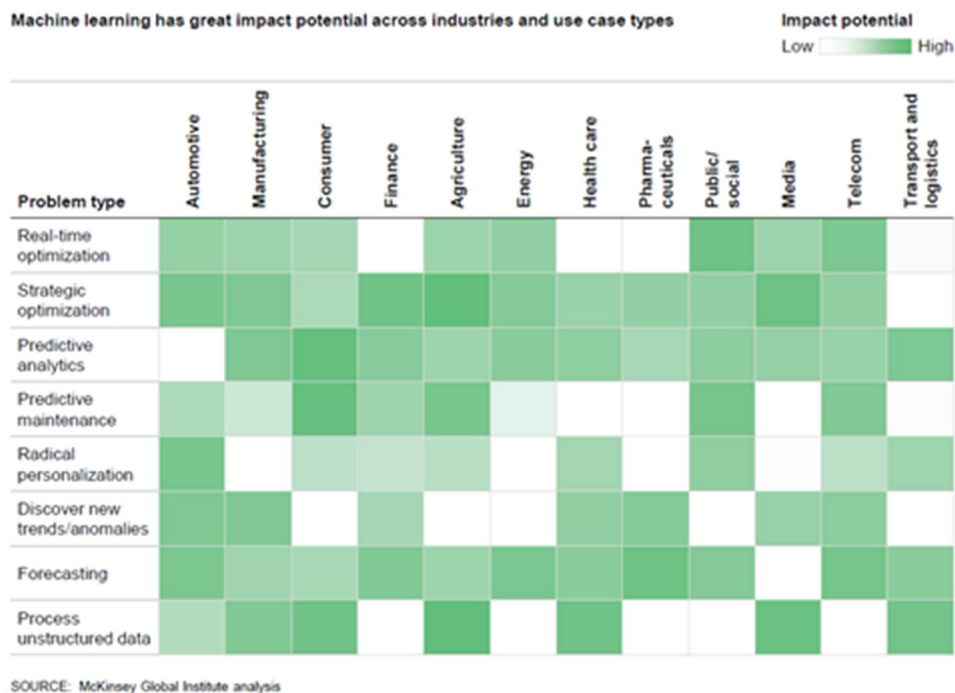


ILUSTRACIÓN 22: IMPACTO DE TÉCNICAS DE MACHINE LEARNING EN DISTINTOS TIPOS DE INDUSTRIA. FUENTE: MCKINSEY GLOBAL INSTITUTE ANALYSIS

En los apartados anteriores se han identificado casos de uso para las tecnologías por separado: Big Data, Data Analytics y Cloud Computing. A continuación se resaltan las tendencias futuras a nivel sectorial implicando las tres tecnologías mencionadas:

Educación

Los datos como habilitadores de innovación tendrán el potencial de transformar la educación, no solo a través de herramientas de formación on-line, sino integrando la analítica de datos con el software educativo, de forma que se consiga adaptar los materiales de estudio a las debilidades o fortalezas de los alumnos. Además, estas tecnologías nos ayudarán a conocer las competencias y habilidades del estudiante, diseñando así materiales escolares individualizados. Las escuelas podrán obtener datos del comportamiento de los estudiantes, identificar tendencias e intervenir en problemas como el absentismo.

Energía

Los contadores inteligentes podrán recoger y transmitir datos, formando parte clave de la red, realizando previsiones de demanda, optimizando la producción de energía para un vecindario o ciudad. Así mismo se podrán establecer políticas dinámicas de precio para reducir las puntas de consumo energético. Incluso los electrodomésticos inteligentes podrán hacer un uso inteligente de los horarios más favorables para reducir los costes de consumo. Globalmente, se espera que las reducciones de las emisiones de CO2 se reduzcan en más de 20 billones de toneladas en 2020.

En Finlandia, Italia y Suecia se han instalado cerca de 45 millones de contadores inteligentes para electricidad, y se espera que Europa tenga instalados 200 millones en 2020, cubriendo el 70% de los consumidores europeos; y 45 millones de contadores inteligentes de gas, cubriendo el 40% de los consumidores.

Gestión Ambiental

Uno de los avances más interesantes a este respecto es el desarrollo de sensores medioambientales instalados en satélites, con la capacidad de transferir enormes cantidades de datos a la tierra. Los científicos podrán monitorizar y realizar previsiones de calidad del aire, cambio climático, emisiones a la atmósfera, prevenir catástrofes, tormentas, terremotos, incendios, creando las alertas necesarias para mitigar las pérdidas de un posible desastre natural. El programa Copérnico, financiado por la UE, provee a los observatorios terrestres de datos de satélites y sensores que monitorizan agua, tierra y aire.

Las políticas marítimo-pesqueras pueden ser, ahora, evaluadas en función de los datos obtenidos vía satélite para medición de temperatura de océanos o color, etc.

Salud

Casi cualquier aspecto relacionado con la salud, desde el descubrimiento de nuevos fármacos hasta la prevención de enfermedades, se van a ver beneficiados por las tecnologías de análisis de datos. Una vez que las historias clínicas estén totalmente digitalizadas, se abre un enorme abanico de oportunidades

relacionadas con la salud: reducción de costes, tratamientos personalizados apoyados en secuencias genéticas y analítica de datos, etc.

Open Data

Se espera que el mercado de bienes y servicios basados en Open Data, en la UE, alcance los 75,7 billones de euros. Por ejemplo, se podrán usar datos de población demográfica, infraestructura, tráfico, etc. con el fin de determinar la ubicación de negocio; y las aseguradoras podrán usar datos de salud pública, condiciones ambientales, estadísticas criminológicas, para el establecimiento dinámico de sus precios.

Smart Cities

El Internet de las Cosas está reconfigurando las ciudades, de forma que los datos capturados faciliten la realización de mejores servicios públicos y se mejore nuestra calidad de vida. La inclusión de sensores y conectividad a las redes de una ciudad, desde la red semafórica hasta la red de saneamiento, permitirá tomar decisiones con el fin de mejorar los servicios públicos y la vida de los ciudadanos.

Smart Manufacturing

Incrementar el ahorro, fortalecer la eficiencia operacional, mejorar la gestión de la calidad de los productos y servicios, disminuir los tiempos de lanzamiento al mercado de nuevos productos, prevenir fallos de equipos, y mejorar la gestión de las materias primas, son algunas de las ventajas que se pueden derivar de la obtención de datos en la industria. Los datos obtenidos tanto en fábrica como fuera de fábrica, redundarán en prestación de servicios y productos más económicos para la industria y más ajustados a las necesidades de los usuarios.

Sector Financiero

El sector financiero es uno de los sectores donde la aplicación de Big Data es más inmediata. Según el último informe de IDC “La Información: Valor diferencial en el Sector Financiero”, la evolución del gasto de las instituciones financieras españolas en Hardware, Software, Comunicaciones y Servicios alrededor de Big Data, para el periodo 2014-2018 prevé una Tasa de Crecimiento Anual Compuesta (en inglés, CAGR) de más de un 18%.

Más del 70% de las entidades financieras españolas ya están utilizando de forma intensiva Big Data, o lo esperan hacer en un plazo inferior a 12 meses. Los usos más comunes son: prevención del riesgo y fraude ligado a las tarjetas de crédito, o reducción de riesgos crediticios utilizando Big Data, para analizar las ventas de los Terminales Punto de Venta (TPV), de forma que la entidad financiera conozca si mejora o empeora la facturación de un determinado negocio, sector o zona geográfica, etc.

4. BIBLIOGRAFÍA

REFERENCIAS

- [1] H. Kagermann and W. Wahlster, “Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative INDUSTRIE 4.0, “Working group, Acatech –National Academy of Science and Engineering, Germany 2013. Final report of the Industrie 4.0.
- [2] Big Data for Modern Industry: Challenges and trends. Shen Yin Okyay Kaynak. 2015 IEEE.
- [3] Gantz, John, and David Reinsel. "Extracting value from chaos." IDC iview 1142 (2011): 9-10.
- [4] An architecture for a business and information system (Devlin & Murphy, IBM Systems Journal 1988).
- [5] Big Data y la historia del almacenamiento de la información. Winshuttle. <http://www.winshuttle.es/big-data-historia-cronologica/>.
- [6] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H. Byers, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011.
- [7] Cukier K Data, data everywhere: a special report on managing information. Economist Newspaper, 2010
- [8]. Big data and analytics—an IDC four pillar research area, IDC, Tech. Rep. 2013. [Online]. Available: <http://www.idc.com/prodserv/FourPillars/bigData/index.jsp>.
- [9] K. Davis, D. Patterson, Ethics of Big Data: Balancing Risk and Innovation, O’Reilly Media, 2012.
- [10] I. O’Reilly Media, Big Data Now: 2014 Edition, O’Reilly Media, 2014.
- [11] Forrester Research, TechRadar: Big Data, Q1 March 10, 2016.
- [12] IDC Research “The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things”, March 2014.
- [13] A Very Short History of Data Science. Gil Press. Revista Forbes May 28 2013.
- [14] Big Data Analytics. Towi Best Practices report. SAS 2011.
- [15]Silipo R. Adae I. Hart A. Berthold M. Seven Techniques for Dimensionality Reduction. Open for Innovation KNIME, 2014.
- [16]Gauss, C. F. (1823). Theoria combinationis observationum erroribus minimis obnoxiae.- Gottingae, Henricus Dieterich 1823.
- [17] Donald, E. K. (1999). The art of computer programming.
- [18] W.S. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5:115–133, 1943.
- [19] Konishi, S. Kitagawa G. Information Criteria and Statistical Modeling, Spring 2008.
- [20] D. R. Cox Principles of Statistical Inference. Cambridge University Press 2006.
- [21] Wattenberg, Martin. "Baby names, visualization, and social data analysis." Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on. IEEE, 2005.
- [22] Gartner. 2015 “Big Data Industry Insight”.
- [23] Business opportunities: Big Data Jun 2013. IDC.
- [24] RackSpace, “Understanding the Cloud Computing Stack: SaaS, PaaS, IaaS” . [Online]. Available: http://www.rackspace.com/knowledge_center/sites/default/files/whitepaper_pdf/Understanding-the-Cloud-Computing-Stack.pdf

- [25] A. Abouzeid and K. Bajda-pawlikowski, “HadoopDB : An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads.”
- [26] “PostgreSQL: The world’s most advanced open source database.” [Online]. Available: <http://www.postgresql.org/>. [Accessed: 21Aug-2014].
- [27] “Home | Pivotal.” [Online]. Available: <http://www.gopivotal.com/>. [Accessed: 16-Jul-2014].
- [28] “Real-Time Analytics Platform | Big Data Analytics | MPP Data Warehouse.” [Online]. Available: <http://www.vertica.com/>. [Accessed: 21-Aug-2014].
- [29] McKinsey Global Institute. A future that works: automation, employment, and productivity - january 2017. MGI.
- [30] Alonso, A. Torres, A., & Dorronsoro, J.R. (2015). Random Forests and Gradient Boosting for Wind Energy Prediction. LNCS, Lecture Notes in Computer Science. 9121, 26-37. Springer. http://link.springer.com/chapter/10.1007%2F978-3-319-19644-2_3 (2016)).
- [31] (Bengio, Y. (2009). Learning Deep Architectures for AI. Foundations and Trends in Machine Learning. 2 (1), 1–127. Y. Bengio, Canada. <http://dx.doi.org/10.1561/2200000006> (2016)
- [32] <http://bynse.com/bynse-precision-agriculture>. BYNSE
- [33] http://cordis.europa.eu/project/rcn/206584_es.html. DataBio
- [34] Regional crop monitoring and assessment with quantitative remote sensing and data assimilation. <http://gtr.rcuk.ac.uk/projects?ref=ST%2FN006798%2F1>
- [35] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1636891. PLANNING:MIDWEST
- [36] <http://gtr.rcuk.ac.uk/projects?ref=BB%2FM011860%2F1>. 14TSB_ATC_IR
- [37] Connected car report 2016. Opportunities, risk, and turmoil on the road to autonomous vehicles. 2016. <https://www.strategyand.pwc.com/media/file/Connected-car-report-2016.pdf>
- [38] Daimler AG. Predictive Analytics to Increase Productivity in Powertrain Production (Cylinder Head) Increase in productivity for the production of cylinder heads: Systematically evaluated empirical knowledge supports faster and more targeted control. <http://www.plattform-i40.de/I40/Redaktion/EN/Use-Cases/116-predictive-analytics-to-increase-productivity-in-powertrain-production/article-predictive-analytics-to-increase-productivity-in-powertrain-production.html>
- [39] Cloud-LSVA. Cloud Large Scale Video Analysis http://cordis.europa.eu/project/rcn/199579_en.html
- [40] Improving access to big data in agriculture and forestry using semantic technologies. Lokers R., Wageningen University | Van Randen Y., Wageningen University | Knapen R., Wageningen University | Gaubitzer S., AIT Austrian Institute of Technology | And 2 more authors. Communications in Computer and Information Science | Year: 2015
- [41] Fuente: La predicción de energías renovables: oportunidades Big Data para la energía eólica. Autor: Ángela Fernández Pascual Data Scientist del IIC en Health & Energy Predictive Analytics. <http://www.iic.uam.es/>
- [42] Integración de datos de inventario y modelos de hábitat para predecir la regeneración de especies leñosas mediterráneas en repoblaciones forestales. Rafael M^a Navarro Cerrillo, Inmaculada Clavero Rumbaó, Astrid Lorenzo Vidaña, José Luis Quero Pérez, Joaquín Duque-Lazo
- [43]. Ship block manufacturing process performance evaluation. Jaehun Park, Dongha Lee, Joe Zhu, Int. J, Production Economics. ELSEVIER, 2014.

- [44] Exponav El astillero 4.0 para las fragatas F 110
https://es.slideshare.net/presentaciones_exponav/el-astillero-40-para-las-fragatas-f110
- [45] Wärtsilä Technical Journal: Big data: a Genius engine for efficiency. PAUL CONNOLLY.
<http://www.wartsila.com/twentyfour7/in-detail/big-data-a-genius-engine-for-efficiency>.
- [46] AkzoNobel and Tessella <http://www.prweb.com/releases/2016/11/prweb13877179.htm>
- [47] EMMA Advisory. Predicting ship behavior navigating through heavily trafficked fairways by analyzing AIS data on apache HBase. Wijaya W.M.,Japan National Defense Academy | Nakamura Y.,Japan National Defense Academy Proceedings - 2013 1st International Symposium on Computing and Networking, CANDAR 2013 | Year: 2013] <http://www.plattform-i40.de/I40/Redaktion/EN/Use-Cases/177-emma-advisory-suite-en/article-emma-advisory-suite-en.html>
- [48] Design of Textile Manufacturing Execution System Based on Big Data. SHAO Jingfeng, HE Xingshi, WANG Jinfu, BAI Xiaobo, LEI Xia2 LIU Congying. .School of Information Engineering, Chang'an University; School of Management, Xi'an Polytechnic University (2015)
- [49] SOMATCH - Support IT solution for creative fashion designers by integrated software systems to collect, define and visualize textile and clothing trends through innovative image analysis from open data [49] http://cordis.europa.eu/project/rcn/196627_en.html].
- [50] Spacecraft electrical signal classification method based on improved artificial neural network. Li K.,Beihang University | Wang Q.,Beihang University | Song S.,China Academy of Space Technology | Sun Y.,China Academy of Space Technology | Wang J.,Beihang University Beijing Hangkong Hangtian Daxue Xuebao/Journal of Beijing University of Aeronautics and Astronautics | Year: 2016
- [51] ConFlux. http://www.greencarcongress.com/2016/04/20160407-umibm.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+greencarcongress%2FTrBK+%28Green+Car+Congress%29
- [52] Big Data case studies. Rob Petersen.
- [53] DART - Data-driven AiRcraft Trajectory prediction research. <http://dart-research.eu/>
- [54] Análisis de la estrategia Big Data en España. Felipe Mirón, Clara Pezuela, Nuria de Lama, Juan Carlos Trujillo, Juan Luis Sobreira, Miguel Angel Mayer, Patricia Miralles, Amelia Martín, Fernando Martín, María Belén García, Jesús Poveda. PLANETIC, España.
- [55] RePhrase - REfactoring Parallel Heterogeneous Resource-Aware Applications - a Software Engineering Approach. http://cordis.europa.eu/project/rcn/194174_en.html
- [56] COCOA CLOUD - Collaborative CO-creation of web Applications on the CLOUD [xx http://cordis.europa.eu/project/rcn/199197_en.html
- [57] Enervalis. <http://tech.eu/brief/enervalis-raises-funds/>
- [58] Zenotech Ltd. Z-DATA (Remote Big Data Rendering for Simulation)
<http://gtr.rcuk.ac.uk/projects?ref=131489>].
- [59] <http://www.rfidworld.ca/manufacturing-process-and-logistics-management-through-the-rfid-technology-in-the-natural-stone-industry/792>
- [60] Solving the slate tile classification problem using a DAGSVM multiclassification algorithm based on SVM binary classifiers with a one-versus-all approach. [xx J. Martinez, C. Iglesias, J.: Matías, J. Taboada, M. Araújo. Applied Mathematics and Computation 230 (2014)].

- [61] La tecnología de Siemens permite a Gestamp reducir un 15% el consumo energético de sus plantas. <http://www.gestamp.com/prensa/comunicados-de-prensa?NewID=2528>
- [62] MC-SUITE - ICT Powered Machining Software Suite http://cordis.europa.eu/project/rcn/198764_en.html].
- [63] Volvo single view of vehicle: Building a big data service from scratch in the automotive industry. Paweł Woźniak, Robert Valton, Morten Fjeld. Proceeding:CHI EA '15 Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. Pages 671-678. Seoul, Republic of Korea — April 18 - 23, 2015

PÁGINAS WEB DE REFERENCIA

- IDC: <http://idcspain.com/>
- Mckinsey Global Institute: <http://www.mckinsey.com/mgi/overview>
- Agenda Digital para España: www.agendadigital.gob.es/
- Data. Unión Europea: <https://ec.europa.eu/digital-single-market/en/big-data>
- Center for data innovation: <http://www.datainnovation.org/>
- The Boston Consulting Group: <http://www.thebostonconsultinggroup.es/>
- Gartner Research: <http://www.gartner.com/technology/research.jsp>
- Computer Sciences Corporation: <http://www.csc.com/>
- EFFRA European Factories of the Future Research Association: <http://www.effra.eu/>
- National Academy of Science and Engineering. ACATECH <http://www.acatech.de/>
- Germany Trade & Invest: www.gtai.com
- PPP Big Data Value. <http://www.bdva.eu/>
- KDD: <http://www.kdd.org/explorations>
- Tic Beat: <http://www.datainnovation.org/>
- Center for data innovation: <http://www.datainnovation.org/>
- AT SCALE: <http://info.atscale.com/atscale-business-intelligence-on-hadoop-benchmark>
- Skytree: <http://www.skytree.net/company/pr/skytree-releases-2013-big-data-analytics-report/>
- IBM Institute for Business Value: <https://www-935.ibm.com/services/us/gbs/thoughtleadership/>
- Leading edge forum: <https://www.leadingedgeforum.com/>
- Roland Berger GmbH: <https://www.rolandberger.com>
- World Economic Forum: <https://www.weforum.org/>
- PricewaterhouseCoopers International Limited www.pwc.com/industry40
- Excelacom: <http://www.excelacom.com/>
- SAS Institute Inc. ("SAS"): https://www.sas.com/es_es/home.html
- Kaggle Inc.: <https://www.kaggle.com/>
- Hadoop: <http://hadoop.apache.org/>
- Apache Spark: <http://spark.apache.org/>
- Apache Avro: <https://avro.apache.org/>
- Apache Flume: <https://flume.apache.org/>
- Apache Kafka: <https://kafka.apache.org/>
- Apache Oozie: <http://oozie.apache.org/>
- Apache Pig: <https://pig.apache.org/>

- Apache Zookeeper: <https://zookeeper.apache.org/>
- Apache Mahout: <http://mahout.apache.org/>
- Apache Lucene: <https://lucene.apache.org/core/>
- ElasticSearch: <https://www.elastic.co/>
- Apache Zeppelin: <https://zeppelin.apache.org/>
- Apache Flink: <https://flink.apache.org/>
- Apache Storm: <http://storm.apache.org/>
- Apache Samza: <http://samza.apache.org/>
- MongoDB: <https://www.mongodb.com/es>
- Apache CouchDB: <http://couchdb.apache.org/>
- Apache Cassandra: <http://cassandra.apache.org/>
- Apache Hbase: <https://hbase.apache.org/>
- Apache Singa: <https://singa.incubator.apache.org/en/index.html>
- Amazon Machine Learning: <https://aws.amazon.com/es/machine-learning/>
- Amazon Web Services, <http://aws.amazon.com>
- XenSource Inc, Xen: <http://www.xensource.com>
- Amazon DynamoDB: <https://aws.amazon.com/es/dynamodb/>
- Azure Machine Learning Studio: <https://azure.microsoft.com/es-es/services/machine-learning/>
- Caffe: <http://caffe.berkeleyvision.org/>
- Massive Online Analysis (MOA): <http://moa.cms.waikato.ac.nz/>
- Weka: <http://www.cs.waikato.ac.nz/ml/weka/>
- MLlib: <http://spark.apache.org/ml/lib/>
- NumPy: <http://www.numpy.org/>
- Mlpack: <http://mlpack.org/>
- Pattern: <http://www.clips.ua.ac.be/pattern>
- Scikit-Learn: <http://scikit-learn.org/stable/>
- Shogun: <http://www.shogun-toolbox.org/>
- TensorFlow: <https://www.tensorflow.org/>
- Theano: <http://deeplearning.net/software/theano/>
- Torch: <http://torch.ch/>
- Veles: <https://velesnet.ml/>
- Keras: <https://keras.io/>
- Google App Engine: <https://appengine.google.com>
- Big Table: <https://cloud.google.com/bigtable/>
- Google Analytics: <https://www.google.com/intl/es/analytics/>
- Woopra, <https://www.woopra.com/>
- Piwik: <https://piwik.org/>
- WebTrends: <https://www.webtrends.com/>
- Google AdWords: <https://www.google.es/adwords/>
- CrazyEgg: <https://www.crazyegg.com/>
- Nvidia: <http://www.nvidia.es/object/cuda-parallel-computing-es.html>
- Redis: <https://redis.io/>

- InfiniteGraph: <http://www.objectivity.com/products/infinitegraph/>
- Neo4j: <https://neo4j.com/>
- Towi Research: <https://tdwi.org/Home.aspx>
- The Internet of Food & Farm 2020 <https://iof2020.eu/iof/iof2020>
- Hispatec: <http://www.hispatec.es/proyectos/hortisys-ininterconecta-feder/>
- Mobileye: <http://www.mobileye.com/en-us/technology/features/>
- Nvidia: <http://.www.nvidia.com/object/drive-px.html>
- IIC. <http://www.iic.uam.es/soluciones/energia/ea2/>
- HyRef de IBM. <http://www.renewableenergyworld.com/articles/2013/08/ibms-hyref-seeks-to-solve-winds-intermittency-problem.html>
- Vi-POC (Virtual Power Operating Center: http://www.smau.it/milano15/partner_products/33555/