
Fundamentos del *big data*: arquitectura del sistema

PID_00247346

José López Vicario

Tiempo mínimo de dedicación recomendado: 2 horas



Universitat
Oberta
de Catalunya

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea éste eléctrico, químico, mecánico, óptico, grabación, fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del copyright.

Índice

Introducción	5
1. Motivación del <i>big data</i> en entorno industrial	6
2. Arquitectura de un sistema de <i>big data</i>	8
3. Fuentes y tipos de datos en la empresa industrial	9
4. Extracción y homogeneización de datos	11
4.1. Importancia de la calidad de los datos	11
4.2. Datos brutos frente a datos procesados	11
4.3. Métodos de extracción de datos	13
5. Computación y almacenamiento en la nube	15
6. Herramientas para el análisis de datos	16
7. Ejemplos de aplicación	17
Resumen	18
Bibliografía	19

Introducción

El objetivo de este material es presentar la arquitectura de un sistema *big data* y las consideraciones que deben tenerse en cuenta cuando se aplica a un entorno industrial.

En primer lugar, se motiva el uso de tecnología *big data* en industria, y se presentan los principales elementos que suelen formar parte de una arquitectura *big data*, desde la adquisición de datos hasta el almacenamiento de los mismos y su posterior análisis.

A continuación, se hará una revisión de la naturaleza de las principales fuentes de datos que se pueden encontrar en el entorno industrial, y se presentarán las principales consideraciones que hay que tener en cuenta a la hora de recolectar estos datos. En concreto, se pondrá énfasis en la importancia que tiene tratar datos relevantes, qué se entiende por trabajar con datos de calidad y los métodos existentes para extraerlos.

Finalmente, se abordará el concepto de almacenamiento en la nube (*cloud*) de los datos recopilados y la posterior computación de los mismos para extraer los patrones de interés. Se expondrán las principales herramientas disponibles, tanto libres como propietarias, para llevar a cabo análisis de datos y, a modo de ejemplos de aplicación, se presentarán un conjunto de aplicaciones totalmente alineadas con el entorno industrial.

1. Motivación del *big data* en entorno industrial

Actualmente, la industria está viviendo una cuarta revolución industrial, conocida como industria 4.0 (véase figura 1 para un resumen de las cuatro revoluciones industriales). Esta nueva revolución se basa principalmente en la conectividad ofrecida por sistemas ciberfísicos (gracias al auge de las redes de sensores y actuadores inalámbricos, paradigma IoT, *internet of things*), lo cual permite tomar un mayor conjunto de medidas en los procesos industriales y, a su vez, proporcionar sistemas de actuación más inteligentes y en tiempo real. Para que esto sea posible, otras tecnologías, también actualmente en auge, como *cloud computing* y *big data*, deben ser explotadas. La primera permite almacenar datos de forma más eficiente, con menor coste y con una mayor accesibilidad desde cualquier lugar. Las tecnologías *big data*, por su parte, permiten analizar conjuntos masivos de datos. El objetivo de este material es mostrar cómo se plantea esta última tecnología en el entorno industrial.

Figura 1. Evolución industrial



Cuando se hace referencia a *big data*, se habla de las 4 V en cuanto a gran *volumen*, *velocidad* y *variedad* que pueden tener los datos que se están tratando, y el problema existente con la *veracidad* de los mismos. Aplicado al entorno industrial, *big data* hace referencia a la gran cantidad de datos generados por un conjunto de fuentes muy distintas, relacionadas con este entorno. Estas fuentes de datos pueden ser sensores de la maquinaria, consumo energético, precio de las materias primas, planificación de procesos y tiempos de ejecución, utilización de recursos, etc.

Al igual que en otros entornos, la motivación de la utilización de tecnología *big data* viene dada por los beneficios, en un ámbito económico, que pueden obtenerse al analizar los datos, ya sea reduciendo costes o incrementando ventas. En el entorno industrial, estos beneficios pueden obtenerse mediante la gestión eficiente de los re-

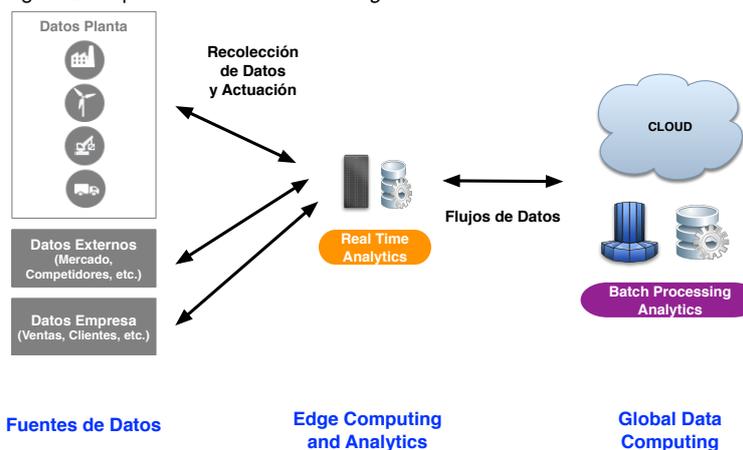
cursos industriales, la extracción de patrones para conocer/mejorar los procesos y la predicción de comportamientos para adaptarse a demandas de producción variantes o para prevenir anomalías (es decir, para mejorar la **gestión de las operaciones**). Se presentarán algunos ejemplos con más detalle en el apartado 7, pero algunas aplicaciones de *big data* en una planta industrial pueden ser: a) optimización del proceso de producción aumentando su capacidad mediante la detección y optimización de cuellos de botella; b) mantenimiento predictivo basado en el uso del histórico de datos de sensores para predecir cuándo es necesario reparar o sustituir maquinaria; o c) producción eficiente energéticamente a partir del análisis de datos de consumo y su relación con los tipos de productos fabricados.

En un ámbito más global, es decir, considerando una compañía con diferentes plantas repartidas en un ámbito geográfico, el análisis de los datos puede utilizarse para mejorar el rendimiento de las plantas de producción y la interacción entre ellas (es decir, para mejorar la **estrategia en operaciones**). Por ejemplo, comparando los resultados entre plantas, es posible detectar buenas prácticas aplicables a las plantas más ineficientes, mejorar la flexibilidad en un ámbito de producción global y reducir los costes totales de producción (analizando costes laborales, de transporte o de adquisición de materias primas y/o energía en las diferentes regiones).

2. Arquitectura de un sistema de *big data*

En la figura 2, se presenta el esquema normalmente utilizado para implementar un sistema *big data* en el entorno industrial.

Figura 2. Arquitectura de un sistema *big data* en entorno industrial



Tal y como puede observarse, este esquema está formado por:

- **Fuentes de datos:** elementos, dispositivos o generadores de datos utilizados en el análisis llevado a cabo por el sistema *big data*. Como se detalla en el apartado siguiente, el entorno industrial presenta un conjunto muy heterogéneo en cuanto a tipos de fuentes.
- **Edge data computing and actuation:** hace referencia al conjunto de mecanismos que llevan a cabo tareas de computación y actuación en un ámbito de planta industrial. Mecanismos encargados de preprocesar/limpiar/preparar los datos para poder almacenarlos en servidores de una forma más organizada, o encargados de ejecutar análisis preliminares para aquellas acciones que requieren actuación en tiempo real. En el material «Gestión de los datos y su uso: *data analytics*» de este mismo PLA, se profundizará en este tipo de análisis, que se denomina *real time analytics*.
- **Global data computing:** en esta etapa, los datos almacenados en servidores (propiarios o servicios *cloud*) son analizados con técnicas de análisis de datos que obtienen mejores prestaciones, al analizarse los datos en todo su conjunto y sin restricciones temporales. Este concepto también se profundizará en el material «Gestión de los datos y su uso: *data analytics*» de este mismo PLA, y se denominará *batch processing analytics*.

3. Fuentes y tipos de datos en la empresa industrial

En un entorno industrial, existe un conjunto muy amplio de datos. A continuación, se presentan las principales fuentes de datos:

- **Datos de mercado:** se refiere a datos relacionados con el mercado o mercados que está abordando la empresa (por ejemplo, mercado de automóviles en Estados Unidos). En este contexto, datos relevantes serían tamaño del mercado (normalmente medido en euros/dólares o la unidad monetaria pertinente); *market share* (porcentaje del mercado capturado por empresa); datos sobre competidores; tendencia del tamaño del mercado, etc.
- **Datos de proveedores:** en este caso, los datos se refieren a los suministradores de materias primas utilizadas por la empresa para generar sus productos o servicios. Es decir, datos relacionados con los proveedores, su naturaleza (tamaño, poder de negociación), número de los mismos, precios ofrecidos, tiempos de entrega y calidad de servicio.
- **Datos de logística:** se refiere a los datos relacionados con el transporte, almacenamiento y distribución de las mercancías a los clientes. Según el tipo de sector al cual pertenece la empresa, estos datos pueden ser de vital importancia para mejorar la competitividad. Esto es así en sectores en los que se maneja un tamaño de inventario elevado, ya que el inventario y su transporte/almacenamiento deben financiarse hasta que se lleva a cabo su venta.
- **Datos de planta:** aquí se puede encontrar un conjunto más rico de fuentes de datos, teniendo en cuenta el contexto de industria 4.0 y el uso de IoT para poder tomar un conjunto mayor de medidas a bajo coste. Estas medidas abarcarían datos de sensores equipados en la maquinaria de la planta, datos sobre consumo energético, datos sobre la capacidad/*throughput*/retardos en los procesos de producción o datos sobre planificación de producción.
- **Datos de ventas:** datos relacionados con las ventas efectuadas por la empresa, tendencia de las mismas, ventas por tipo de producto o servicio, mercado o cliente.
- **Datos de clientes:** información relevante sobre los clientes de la empresa, su tamaño, sector, poder de negociación, ventas anuales, precios y productos ofrecidos, localización geográfica, perfil de cliente y segmento al cual pertenece.

- **Datos relevantes del entorno:** datos relativos al contexto que rodea la actividad empresarial, como países donde se opera, legislación y normativa aplicable, costumbres y características de la región.

4. Extracción y homogeneización de datos

4.1. Importancia de la calidad de los datos

Tal y como se ha observado en el apartado anterior, los datos utilizados en el entorno industrial provienen de fuentes heterogéneas y de naturaleza muy diversa. Esto hace que disponer de un conjunto de datos organizados y bien estructurados para su posterior análisis resulte una tarea muy difícil. Por este motivo, en este entorno es vital la utilización de técnicas de preprocesado para organizar todos los datos y garantizar que sean completos, continuos y sincronizados.

Por otro lado, en el entorno industrial, los atributos analizados tienen una clara interpretación física, y garantizar la integridad de los datos que los caracterizan es también de gran importancia para no llegar a análisis o tomas de decisiones erróneas. Disponer de datos de baja calidad (alterados por ruido) o mal almacenados (por errores de sensores, mala conectividad o fallos de almacenamiento) puede llevar a caracterizar de forma errónea las relaciones causa-efecto, y provocar de manera catastrófica la toma de decisiones basada en el análisis de los datos.

Como conclusión, en este apartado cabe mencionar que el conocimiento sobre el entorno industrial considerado es fundamental para poder desarrollar proyectos *big data*. Este conocimiento, así como el hecho de contar con un equipo apropiado, permiten que sea posible diseñar técnicas de preprocesado adecuadas y detectar cuándo los resultados obtenidos no son fiables, debido a que los datos han sido de baja calidad.

4.2. Datos brutos frente a datos procesados

Cuando se hace referencia a **datos brutos**, nos estamos refiriendo a los datos recibidos directamente de sus fuentes. Estos datos están caracterizados por las siguientes propiedades:

- Inexistencia de formato unificado.
- Posibilidad de valores no asignados (*missing values*).
- Posibilidad de medidas corruptas por ruido en medidas o fallos en comunicación/almacenamiento de los datos.
- Dificultad de ser tratados por algoritmos de análisis de datos directamente.

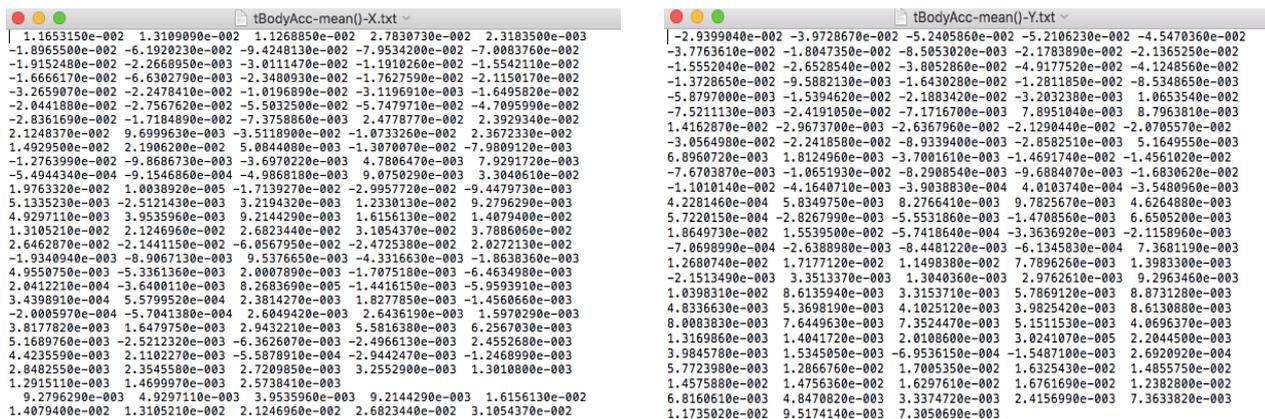
En cuanto a los **datos procesados**, son los datos a los que se ha aplicado algún tipo de tratamiento (preprocesado/limpieza/preparación de datos, mencionado en el apartado 2). El objetivo de este tratamiento es que los datos puedan ser almacenados en unas estructuras de datos, normalmente referidas como *data sets*, para su posterior análisis. Es decir, los datos pueden expresarse en forma de tabla: cada variable/atributo está en una columna, cada valor de cada variable está en una fila y cada columna tiene el nombre de la variable a la que se refiere la misma.

En las figuras 3 y 4, se presenta de forma ilustrativa cómo se desea que estén los datos organizados para poder llevar a cabo un análisis de los mismos frente a cómo se suelen encontrar normalmente en la práctica. Observad cómo en el caso de datos procesados, la primera figura, los datos obtenidos de un conjunto de acelerómetros (por ejemplo, equipados en un robot industrial), se presentan bien organizados, homogeneizados y preparados para poder ejecutar mecanismos de análisis de datos. En el segundo caso, en cambio, se muestran los datos brutos obtenidos de diferentes sensores. En este caso, los datos presentan un aspecto más caótico, ya que se obtienen consultando directamente la fuente que los genera. Por tanto, se observa la necesidad de manipularlos y procesarlos para poder generar el *data set* deseado.

Figura 3. Datos procesados de medidas de acelerómetros equipados en un robot industrial

	subject	activity	tBodyAcc-mean()-X	tBodyAcc-mean()-Y	tBodyAcc-mean()-Z	tGravityAcc-mean()-X	tGravityAcc-mean()-Y	tGravityAcc-mean()-Z	tBodyAccJerk-mean()-X
1	1	andar	0.2773308	-0.017383819	-0.11114810	0.9352232	-0.282165021	-0.068102864	0.07404163
31	1	bajarEsc	0.2891883	-0.009918505	-0.10756619	0.9318744	-0.266610339	-0.062119959	0.05415532
61	1	depie	0.2789176	-0.016137590	-0.11060182	0.9429520	-0.272983832	0.013490582	0.07537665
91	1	sentado	0.2612376	-0.001308288	-0.10454418	0.8315099	0.204411593	0.332043703	0.07748252
121	1	subirEsc	0.2554617	-0.023953149	-0.09730200	0.8933511	-0.362153364	-0.075402940	0.10137273
151	1	tumbado	0.2215982	-0.040513953	-0.11320355	-0.2488818	0.705549773	0.445817720	0.08108653
2	2	andar	0.2764266	-0.018594920	-0.10550036	0.9130173	-0.346607090	0.084727087	0.06180807
32	2	bajarEsc	0.2776153	-0.022661416	-0.11681294	0.8618313	-0.325780101	-0.043889016	0.11004062
62	2	depie	0.2779115	-0.018420827	-0.10590854	0.8969286	-0.370062697	0.129747161	0.07475886
92	2	sentado	0.2770874	-0.015687994	-0.10921827	0.9404773	-0.105630024	0.198726769	0.07225644
122	2	subirEsc	0.2471648	-0.021412113	-0.15251390	0.7907174	-0.416214890	-0.195888239	0.07445078
152	2	tumbado	0.2813734	-0.018158740	-0.10724561	-0.5097542	0.752536639	0.646834880	0.08259725
3	3	andar	0.2755675	-0.017176784	-0.11267486	0.9365067	-0.261986358	-0.138107866	0.08147459
33	3	bajarEsc	0.2924235	-0.019355408	-0.11613984	0.9390578	-0.228829214	-0.102352758	0.07256893
63	3	depie	0.2800465	-0.014337656	-0.10162172	0.9350308	-0.301735130	0.024763107	0.07509006
93	3	sentado	0.2571976	-0.003502998	-0.09835792	0.9010990	0.127303379	0.139020582	0.07260984
123	3	subirEsc	0.2608199	-0.032410941	-0.11006486	0.8835334	-0.382851154	-0.162943955	0.04268810
153	3	tumbado	0.2755169	-0.018955679	-0.10130048	-0.2417585	0.837032100	0.488703185	0.07698111
4	4	andar	0.2785820	-0.014839948	-0.11140306	0.9639997	-0.085854027	0.127764113	0.07835291

Figura 4. Datos brutos de diferentes acelerómetros



Finalmente, en la siguiente tabla se presenta un resumen de la comparación entre datos brutos y datos procesados.

Tabla 1

Datos brutos	Datos procesados
Datos extraídos directamente de la fuente original	Datos organizados y listos para ser analizados
Difícilmente pueden utilizarse para llevar a cabo un proceso de análisis.	Pueden existir unas normas más o menos estrictas para su análisis (legales, tecnológicas, estadísticas).
El análisis de los datos requiere un procesado previo.	El análisis puede requerir algunas operaciones previas (aglutinación, fusión, segmentación).
Puede requerir una o más transformaciones.	Todos los pasos del análisis deben documentarse.

4.3. Métodos de extracción de datos

Tal y como se ha presentado en el apartado 3, en el entorno industrial existe un conjunto de datos muy heterogéneo. Por este motivo, se requiere el uso de diferentes métodos de extracción de datos según la fuente de datos. Las herramientas de software utilizadas para el análisis de datos (véase el apartado 6) cuentan con multitud de librerías/funcionalidades para poder extraer datos. El desarrollador debería, simplemente, consultar cada caso específico según su entorno de trabajo. Algunos casos usuales serían los siguientes:

- **Acceso a bases de datos:** para extraer datos sobre clientes, proveedores, información de la empresa, ventas, precios, etc.
- **Lectura de ficheros locales:** para extraer información similar a la del punto anterior, pero que ha sido almacenada en ficheros Excel, por ejemplo.

- **Descarga de ficheros de la web:** para extraer datos disponibles en páginas web con información sobre el entorno empresarial.
- **Lectura de formatos web:** para extraer información contenida en formatos comúnmente utilizados en la web, como XML o JSON.
- **Web scrapping:** hace referencia a la extracción directa de los datos contenidos en una página web. La aplicación industrial sería similar a los dos casos anteriores.
- **Lectura de datos recolectados por sensores:** la cual normalmente se lleva a cabo, en un entorno industrial, mediante plataformas basadas en SCADA o el uso de nodos recolectores (llamados *gateways*), encargados de agregar las medidas de diferentes sensores.

5. Computación y almacenamiento en la nube

Como se ha mencionado en el apartado 1, una de las tecnologías explotadas en el paradigma de industria 4.0 es la tecnología *cloud* ('nube'). Esta tecnología se basa en un nuevo concepto en el que proveedores de servicio ofrecen servicios de almacenamiento y/o computación a través de internet. Es decir, los clientes de estos servicios utilizan las infraestructuras de servidores de almacenamiento y computación de los proveedores para almacenar sus datos o llevar a cabo tareas de procesamiento de los mismos.

En el ámbito de *big data*, esta tecnología facilita la posibilidad de operar con un gran volumen de datos. Esto se debe a que la principal ventaja que ofrecen soluciones de almacenamiento basadas en la nube es que el coste de las mismas suele basarse en un pago por uso de recursos de almacenamiento (por ejemplo, pago basado en euros/GB). Por tanto, la empresa que opta por esta opción no debe incurrir en una gran inversión en infraestructura (hardware/servidores) y su posterior mantenimiento (del orden del 30 % en algunos casos) necesario para almacenar *big data* [Minelli y otros (2012) Minelli, Chambers y Dhiraj]. Estamos, por tanto, ante un modelo basado en OPEX (gastos operativos, es decir, gastos mensuales/anuales por un servicio) frente al tradicional, basado en CAPEX (inversiones en bienes de capitales, las cuales requieren de una cierta capacidad financiera). El modelo basado en OPEX ha ido ganando mucha popularidad en los últimos años, debido a la crisis financiera vivida. También son modelos muy adecuados para empresas que no tienen un gran tamaño. Otra ventaja es que la capacidad de almacenamiento ofrecida por el proveedor puede escalar según las necesidades del cliente, y las actualizaciones de la plataforma son automáticamente llevadas a cabo por parte del proveedor. En un modelo CAPEX, en cambio, se debería ir reinvertiendo en nueva infraestructura si el volumen de los datos fuera creciendo.

No obstante, muchas empresas no han migrado aún a la nube debido a la desconfianza que puede generar tener los datos almacenados en instalaciones de terceros. Ante este punto, cabe mencionar que los principales proveedores, como Amazon (AWS, Amazon Web Services), ofrecen unos niveles de servicio garantizados mediante SLA (*service level agreement*). Por otro lado, los niveles de inversión y especialización de estos proveedores permiten ofrecer una seguridad mayor que la que muchas empresas podrían tener en sus propias instalaciones.

Aparte de servicios de almacenamiento, muchos proveedores *cloud* ofrecen servicios de computación o de conectividad de dispositivos. Siguiendo con el ejemplo de Amazon, este proveedor ofrece servicios como Amazon EC2 (capacidad de cómputo variable) o AWS IoT (plataforma que permite la conectividad de dispositivos en la nube). De nuevo, son servicios basados en pago por consumo (bajo demanda o reservado), y resulta interesante para aquellas empresas que no dispongan de capacidad financiera para abordar una gran inversión.

6. Herramientas para el análisis de datos

Actualmente, existe un conjunto de herramientas para el análisis de datos realmente amplio. En un ámbito de herramientas propietarias, encontramos soluciones de principales proveedores como, por ejemplo, General Electric (<https://www.ge.com/digital/>, con productos como Asset Performance Management o Predix); SAS (https://www.sas.com/es_es/home.html, con productos/soluciones como Supply Chain Intelligence o Customer Intelligence); NTT Data (<https://us.nttdata.com/en/Industries>) o IBM (<https://www.ibm.com/internet-of-things/iot-solutions/iot-manufacturing/>).

No obstante, el uso de herramientas libres se ha hecho muy popular en los últimos años, debido a su gran soporte por parte de comunidades amplias y por las excelentes prestaciones ofrecidas. Dentro del ecosistema amplio de herramientas *big data* libres, en este apartado destacamos un conjunto de las mismas (algunas se analizan con más detalle en los materiales «Gestión de los datos y su uso: *data analytics*» y «Tecnologías de *big data*» de este PLA):

- **Recolección de datos:** herramientas encargadas de recolectar datos emitidos por fuentes, y su posterior organización para asegurar la no pérdida de los mismos. Soluciones como Apache Flume, Apache Sqoop o Apache Kafa.
- **Almacenamiento de datos:** herramientas encargadas de almacenar datos masivos para su posterior análisis. Soluciones como HDFS o HBase.
- **Análisis de datos basados en *batch processing*:** herramientas encargadas de analizar datos almacenados en bases de datos distribuidas en servidores propios o en la nube. Soluciones como Hadoop o Apache Spark.
- **Análisis de datos en *stream processing*:** herramientas encargadas de analizar datos en tiempo real. Soluciones como Apache Spark Streaming o Apache Storm.
- **Análisis estadístico:** herramientas encargadas de llevar a cabo análisis estadísticos de los datos o de aplicar algoritmos *machine learning*. Suelen estar basadas en lenguajes de programación que pueden utilizarse para desarrollar sistemas ejecutados en los dos puntos anteriores (*batch* o *stream processing*). Soluciones como R o Python.

7. Ejemplos de aplicación

En la siguiente tabla, se presentan una serie de ejemplos de aplicación de *big data* en el sector industrial.

Tabla 2

Ámbito	Aplicación
Excelencia operativa	<ul style="list-style-type: none"> • Automatización de toma de decisiones paramétrica • Optimización de procesos de negocio • Optimización de flujos de producción • Optimización y monitorización en tiempo real • Optimización eficiencia energética • Control y aseguramiento de calidad • <i>Data center management</i>
Eficiencia de proceso y producto	<ul style="list-style-type: none"> • Mejora en el diseño de procesos y productos con datos de uso • Análisis y predicción de la demanda • Detección de patrones y <i>root cause analysis</i> • Inteligencia de mercado • Mantenimiento predictivo y preventivo de equipamiento, instalaciones y estructuras • Precios dinámicos • <i>Product lifecycle</i>
Gestión de riesgos	<ul style="list-style-type: none"> • Seguros • Fraude • Seguridad (videoanálisis en tiempo real)
Productos y servicios inteligentes	<ul style="list-style-type: none"> • Dato como producto y dato como servicio (<i>data as a service</i>) • Monitorización remota • Edificios inteligentes • Fábrica conectada • Inteligencia ambiental • Gestión del conocimiento técnico (distintas fuentes de información)
Conocimiento del cliente	<ul style="list-style-type: none"> • <i>Customer support</i> • Análisis de sentimiento • <i>Targeting</i> • Recomendación • Prevención del abandono • <i>Customer lifetime value</i> • Análisis del <i>product mix</i> • <i>Personal analytics</i>

Resumen

Este material tiene como objetivo presentar los fundamentos del *big data* aplicado a un entorno industrial. En primer lugar, se ha motivado el uso de esta tecnología en la industria para poder satisfacer las necesidades de la cuarta revolución que se está viviendo actualmente y, hecho esto, se ha mostrado la arquitectura de un sistema *big data* típico. Teniendo en cuenta el gran conjunto de datos que pueden encontrarse en un entorno industrial, se han presentado las fuentes de datos existentes, la necesidad de homogeneizarlos para su posterior análisis y los métodos de extracción de datos más utilizados.

Finalmente, se describirá el concepto de *cloud computing*, uno de los pilares para facilitar la penetración de *big data* en la industria actual, para pasar luego a presentar las herramientas normalmente utilizadas en el análisis de datos, y algunos ejemplos de aplicación industrial.

Referencia bibliográfica

Trovati, M.; Hill, R.; Anjum, A.; Ying Zhu, S.; Liu, L.
(2016). *Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications*. Springer.

Bibliografía

Indústria (2016). *La transformación digital de la Industria española*. Informe del Ministerio de Industria, Energía y Turismo, disponible en «<http://www.industriaconectada40.gob.es>».

Minelli, M.; Chambers, M.; Dhiraj, A. (2012). *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. John Wiley and Sons.

Trovati, M.; Hill, R.; Anjum, A.; Ying Zhu, S.; Liu, L. (2016). *Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications*. Springer.