UNIVERSIDAD D SEVILLA

MÁSTER UNIVERSITARIO EN ESTUDIOS AVANZADOS EN DIRECCIÓN DE EMPRESAS

BIG DATA Y BUSINESS INTELLIGENCE PARA LA GENERACIÓN DE CONOCIMIENTO

[BIG DATA AND BUSINESS INTELLIGENCE TO GENERATE KNOWLEDGE]

TRABAJO FIN DE MÁSTER

Inés Ameijeiras Lois, septiembre 2019

Departamento de Economía Financiera y Dirección de Operaciones

BIG DATA Y BUSINESS INTELLIGENCE PARA LA GENERACIÓN DE CONOCIMIENTO

[BIG DATA AND BUSINESS INTELLIGENCE TO GENERATE KNOWLEDGE]

Trabajo Fin de Máster presentado para optar al Título de Máster Universitario de Estudios Avanzados en Dirección de Empresas por Inés Ameijeiras Lois, siendo el tutor del mismo el Doctor José Carlos Ruiz del Castillo.

Vº. Bº. del Tutor/a:

RUIZ DEL CASTILLO JOSE CARLOS - 30493998T

Digitally signed by RUIZ DEL CASTILLO JOSE CARLOS - 30493998T
Date: 2019.08.19 09:45:08 +02'00'

D. José Carlos Ruiz del Castillo

Alumno/a:

Dª. Inés Ameijeiras Lois

Sevilla, 2 de septiembre de 2019

MÁSTER UNIVERSITARIO DE ESTUDIOS AVANZADOS EN DIRECCIÓN DE EMPRESAS

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

TRABAJO FIN DE MÁSTER

CURSO ACADÉMICO [2018-2019]

TÍTULO:

BIG DATA Y BUSINESS INTELLIGENCE PARA GENERAR CONOCIMIENTO

[BIG DATA AND BUSINESS INTELLIGENCE TO GENERATE KNOWLEDGE]

AUTOR/A:

INÉS AMEIJEIRAS LOIS

TUTOR/A:

JOSÉ CARLOS RUIZ DEL CASTILLO

LÍNEA DE TRABAJO:

SISTEMAS DE INFORMACIÓN. BUSINESS INTELLIGENCE Y BIG DATA

RESUMEN:

En este trabajo se definen las disciplinas de Business Intelligence y Big Data, se aporta un contexto teórico, se exponen las fases del proceso de utilización de herramientas de estos ámbitos y se explican y utilizan herramientas útiles y accesibles para las PYMES para el tratamiento de sus datos y la obtención de ventajas competitivas.

PALABRAS CLAVE:

Business Intelligence; Inteligencia de negocio; Big Data; Generación de conocimiento; Visualización de datos; Web scraping

ÍNDICE

Relación de Figuras

Relación de Tablas

Relación de abreviaturas

AWS	Amazon Web Service. Servicio Web de Amazon
BD	Big Data o macrodatos.
BI	Business Intelligence o Inteligencia de Negocio
CRM	Customer Relationship Management
CSV	Comma-separated values. Valores separados por comas
ERP	Enterprise Resource Planning
ETL	Extract, Transform and Load. Extracción, transformación y carga.
EDI	Intercambio Electrónico de Datos
INE	Instituto Nacional de Estadística
KPI	Key Performance Indicator
KRI	Key Results Indicator
MUEADE	Máster en Estudios Avanzados en Dirección de Empresas
ONTSI	Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información
p.	Página
p.ej.	Por ejemplo
PI	Performance indicator
PYMES	Pequeñas Y Medianas Empresas
RI	Result indicator
SABI	Sistema de Análisis de Balances Ibéricos
SQL	Structured Query Language. Lenguaje de consulta estructurada.
US	Universidad de Sevilla
TPS	Programas de Procesamiento de Transacciones

CAPÍTULO 1. Introducción

1.1. Justificación

El Big Data y el Business Intelligence están de plena actualidad en las empresas. Y es que cada vez más negocios deciden aplicar estas soluciones en el área de los Sistemas de la Información para buscar ventajas competitivas. Además, en el contexto actual, en el que nos encaminamos a la era de la información -si es que no estamos ya en ella-, es importante para la supervivencia y crecimiento de muchas empresas que puedan tener información en tiempo real, por ejemplo, saber lo que piensan y publican en la red sus clientes.

Al igual que muchas tendencias en el mundo de los negocios, muchas pequeñas y medianas empresas (PYMES) consideran que esto del Big Data y el Business Intelligence no les es aplicable. Y es que los directivos de estas empresas o bien no conocen estas herramientas o desconocen las ventajas de su implantación y piensan que son muy costosas.

1.2. Objetivos

El objetivo primario de este trabajo es estudiar la implantación y utilización de herramientas de Big Data y Business Intelligence en las PYMES con el fin de que obtengan ventajas competitivas.

Además, se determinan otros objetivos secundarios que consisten en:

- Proponer soluciones gratuitas o de bajo coste para que las PYMES también gestionen sus datos e información con estas herramientas.
- Exponer la utilidad y las ventajas que suponen la aplicación de este tipo de soluciones.

1.3. Metodología

Para alcanzar los objetivos recogidos en el punto anterior se combinaron tres metodologías: revisión bibliográfica, utilización de las herramientas en ejercicios prácticos grabados en vídeo y creación de un blog para incluir los anexos.

En primer lugar, se realizó una revisión bibliográfica sobre el Big Data y el Business Intelligence y su utilización en la empresa con las siguientes características:

- Fuentes de datos: las búsquedas se realizaron en el catálogo de la biblioteca de la Universidad de Sevilla, Google Scholar y Dialnet.
- Las palabras clave empleadas fueron: Big Data, Business Intelligence, big data implementation process, ETL, web scraping, datos e información en la empresa, database, datwarehouse, cuadro de mando, etc.
- Filtrado de información: entre los resultados se seleccionó aquella información acorde con los objetivos, de fuentes fiables, de redacción adecuada para un trabajo formal e incluibles en la estructura del trabajo, obteniendo como resultado 45 fuentes gestionadas con Mendeley y que incluyen revistas, libros y publicaciones en Internet. Entre las revistas consultadas se encuentra la MIT Sloan Management Review y IEEE Access. Y entre las publicaciones en Internet destacan los artículos de empresas como Oracle, Aukera, Gartner o Amazon.

- Período temporal: se estableció un filtro inicial de publicaciones del 2011 o posteriores. Se permitió una excepción de 2004 para incluir las definiciones de tipos de bases de datos de Sánchez Sánchez et al.

En segundo lugar, se realizaron una serie de pequeños ejercicios prácticos recogidos en vídeos de demostración en los que se emplean las herramientas citadas en la revisión bibliográfica. Para grabarlos se empleó el software FastStone. Se incluyen en los apartados dedicados a las herramientas y en el anexo los enlaces correspondientes, así como un código QR para la versión en papel.

Por último, se detectó que había definiciones que se podrían anexar al trabajo, así como, ejemplos de los tipos de ficheros citados y se decidió adjuntarlos empleando un blog. Las entradas del blog se referencian por hipervínculos en las palabras clave que los titulan. Para la versión en papel se incluye un código QR en el anexo.

1.4. Estructura

Además de este capítulo introductorio y del de conclusión, el trabajo consta de otros ocho capítulos. En el segundo se establece el contexto del Big Data y el Business Intelligence y de la empresa en general en la era histórica actual: la de la información.

En el tercer capítulo se tratan los tipos de datos, fuentes y bases de datos.

En el cuarto, se introducen el Big Data y el Business Intelligence y se comparan.

En el quinto, se establece un proceso de implantación válido para ambas disciplinas, cuyas actividades se detallan en los tres capítulos siguientes.

CAPÍTULO 2. La era de la información

2.1. Introducción

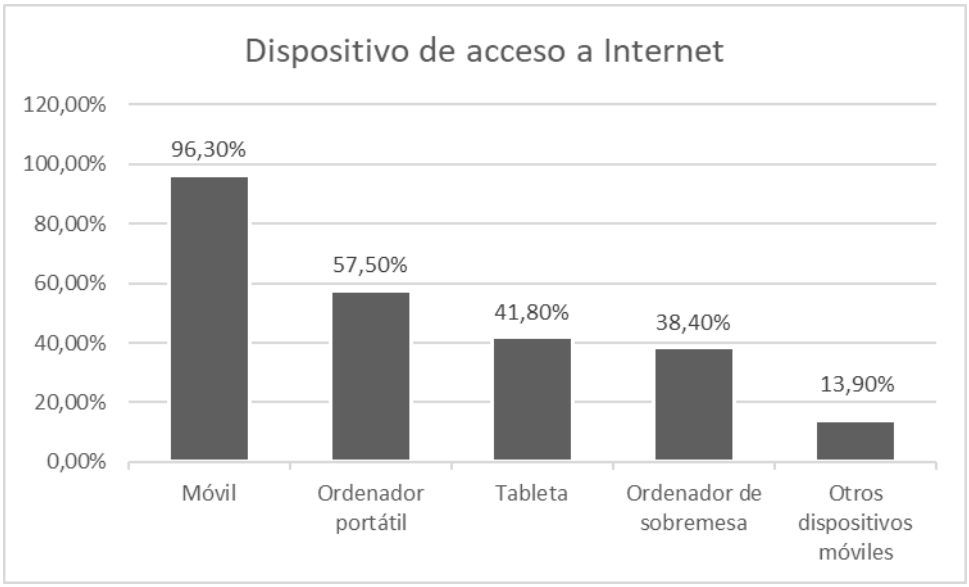
En este capítulo se va a explicar el contexto en el que se emplean las herramientas de Business Intelligence y Big Data, así como, reflejar su relevancia en la situación mundial actual e introducir conceptos necesarios para la introducción del resto del trabajo.

2.2. Contexto

En el siglo XXI se está viviendo una transformación social comparable a la del Neolítico o a la Revolución Industrial, como reflexionó el Doctor en Ciencias Económicas y Empresariales, Francisco José Martínez López, en su ponencia como parte de la asignatura ERP en el plan de estudios de MUEADE (Máster de Estudios Avanzados en Dirección de Empresas), de la Universidad de Sevilla. En el neolítico el ser humano se preocupó por cubrir la necesidad de comer, sin necesidad de ir de caza o recolección todos los días, pues inventó la agricultura (Martínez López, 2019). Y en la Revolución Industrial, con la comida ya cubierta, los productos prioritarios en los que pasó a centrarse la economía fueron ciertas comodidades como las casas, las camas o los electrodomésticos. Actualmente, como bien remarcaba el citado doctor, no se ha llegado a acuerdo sobre si ya ha empezado la tercera gran era de la historia: la de la información. Al menos los cambios que nos llevan a ella, ya se pueden observar en la sociedad.

Esta tercera era se caracteriza por la terciarización, la automatización, la globalización, la complejidad, la interactividad y la información como factor de producción (De Pablos Heredero, López Hermoso Agius, Martín-Romo Romero, & Medina Salgado, 2012, p.56). Lo característico de la información es que se está generando a gran velocidad y desde diversas fuentes. Las personas generamos datos continuamente: con nuestros accesos a Internet y también con muchas otras acciones cotidianas como hacer la compra o entrar en un parking de pago. Y las máquinas también generan continuamente datos, materia objeto de estudio en el Internet de las cosas (Murillo González, 2016, p.9). En la actualidad, prácticamente cualquier fenómeno puede ser observado y almacenado, a esto se le denomina dataficación (Maheshwari, 2015, p.8). Esta generación de datos la facilitan aplicaciones de mensajería, el correo electrónico, las redes sociales, las transacciones bancarias electrónicas, etc. La mayoría de ellas se caracterizan porque las usamos gratis, las queremos actualizadas en tiempo real y las llevamos siempre con nosotros en nuestro teléfono móvil. De hecho, este es el dispositivo más empleado para acceder a Internet, de acuerdo con los datos recogidos por la ONTSI en los últimos tres meses de 2018 (Figura 1).

Figura 1 Dispositivos utilizados para conectarse a Internet en los últimos tres meses de 2018



Fuente: Ontsi red.es, 2018

Es un hecho que se generan más datos, a mayor velocidad y con menor coste y menores necesidades de hardware (Mochón Morcillo & Gonzálvez Cabañas, 2016, p.18). Y es que ya no es necesario que las empresas tengan sus propios servidores en los que almacenar todos sus datos, pueden subcontratar estos servidores y acceder a ello a través de Internet. Ejemplo de ello son los ERP en la nube, donde la empresa usuaria tiene un enlace desde el que puede acceder al programa y los servidores son propiedad del proveedor y pueden estar en cualquier lugar del mundo. Por ello, los servidores, que antes se vendían como un producto, han pasado a ser un servicio, ya que, se alquila su capacidad de almacenamiento. Lo que puede ser una gran oportunidad para las empresas, que pueden aprovechar toda esta información y crear ventajas competitivas. Pero para ello deben diseñar un proceso de captación y procesamiento, además de aprender a gestionar y utilizar sistemas de información con estas herramientas.

2.3. Datos, información y conocimiento

En la era de la información, la materia prima característica son los datos. Se trata de la unidad mínima en la jerarquía de la información (Rojas Pescio, 2016, p.70). Los datos son una medición objetiva y se caracteriza por la facilidad de almacenamiento y manipulación. Además, no tienen significado de forma aislada (Espinoza & Secaira, 2016, p.664). Por ejemplo, 22, 1’80, 60, rubia y azules son datos.

La información surge del procesamiento de un conjunto de datos mediante la contextualización, categorización, cálculo, corrección, condensación o relación de los datos, entre otras operaciones que entregan valor a los datos (Rojas Pescio, 2016, p.69). Por lo que, se puede definir como un conjunto de datos relacionados e interpretados en un contexto específico (Espinoza & Secaira, 2016, p.664) (Figura 2). Volviendo al ejemplo, los datos citados se convierten en información al contextualizarlos explicando que 22 son los años que tiene una persona, 1’80 cm los que mide esa persona, 60 los kilos que pesa, rubia se refiere a la cerveza que se ha bebido y no al color de su pelo y azules son los pantalones que lleva puestos hoy. De forma similar se pueden realizar las otras cinco acciones enumeradas por Rojas Pescio para pasar de los datos a la información.

Sin embargo, Farradane asemejaba la información a datos ordenados (Suárez Sánchez, 2017, p.6), olvidando los numerosos y más significativos procesos que sí

considera Rojas Pescio. Además, el simple hecho de ordenar datos, por ejemplo, de menor a mayor o por orden alfabético, no les dota de significado, por lo que, siguen siendo datos.

Figura 2 Datos e información

Fuente: elaboración propia

A partir de la información, cada individuo con su experiencia en la materia, genera conocimiento (Figura 3). Para conocer algo tenemos que entender la información sobre ello y elaborar una representación mental. En esa representación mental toma partido el juicio personal (Briones Delgado, 2014, p.75). El autor David Loshin habla de un profundo y preciso entendimiento de la información basado en patrones (Loshin, 2012, p.8). Esos patrones habrían sido aprendidos por esa persona en el pasado por su experiencia vital. Pues es posible que alguien nunca haya comido un pistacho, pero sí conozca las pipas o los cacahuetes y, por lo tanto, un patrón de conducta humana de extraerle las cáscaras. De esta forma, cuando le ofrezcan un pistacho, reconocerá la cáscara y no la comerá. Esos patrones forman parte del conocimiento. A un niño se le puede informar de que esos tres frutos secos tienen cáscara y hay que quitársela, pero solo adquirirá el conocimiento necesario para evitar un atragantamiento por cáscara cuando haya visto, pelado y comido estos frutos secos. Así, podrá crear su representación mental y aplicarla cuando le presenten un fruto seco similar. Actualmente, existe un debate sobre si las máquinas pueden generar conocimiento. Es cierto que el Machine Learning, o aprendizaje de las máquinas, está avanzando rápidamente. Ejemplo de ello son los avances en algoritmos predictivos que mejorar sus resultados practicando con series históricas o el coche autónomo de Amazon que aprende conduciendo.

Farradane destaca que la información es algo físico, mientras que el conocimiento es abstracto (Suárez Sánchez, 2017, p.6). Debido a estas características, la información se puede transmitir y el conocimiento no. Quién posee el conocimiento puede elaborar y compartir información para facilitar que otro aprenda, que genere su propio conocimiento o juicio personal. Esa información que genera tomar forma de escritos, documentos de texto, mensajes de audio o simplemente una conversación hablada; en definitiva, algo captable con los sentidos. Pero, debatiendo a Farradane, no necesariamente físico, pues se entiende físico como algo material y el sonido no lo es. Que alguien sabe o no, que conoce un tema, es algo que solo se puede intuir a partir de la información que transmite.

Figura 3 Información y conocimiento

Fuente: elaboración propia

En resumen, los datos son la materia prima de la era de la información. Se trata de unidades básicas sin significado y fáciles de manipular. Si interpretamos y/o relacionamos los datos mediante su contextualización, categorización, cálculo, corrección o condensación, generamos información. La información sí contiene significado y es algo que podemos percibir por los sentidos y transmitir con facilidad a otros. Si esta información la hacemos parte de nuestra experiencia vital, tratándola, trabajándola y/o estudiándola, generamos conocimiento. El conocimiento es abstracto, no se puede captar por los sentidos, y es intrínseco de cada persona, por lo que, tiene un componente de juicio y se transfiere con mayor dificultad que la información.

CAPÍTULO 3. Tipos de datos, fuentes y bases de datos

3.1. Introducción

En este capítulo se introducen conceptos característicos de la era de la información que son importantes de cara a comprender el proceso de generación de conocimiento con la ayuda del Business Intelligence y el Big Data.

Concretamente, se definirán las bases de datos y las fuentes de datos, se comentará su variedad y se caracterizarán los tipos de fuentes de datos según su estructura.

3.2. Datos, bases de datos y fuentes de datos

Como se han definido en el apartado anterior, los datos son la materia prima de la era de la información y carecen de significado. Estos son proporcionados por fuentes de datos y son almacenados en bases de datos.

Una base de datos es un sistema informático de registro y almacenamiento de datos que permite su consulta (Foster & Godbole, 2014, p.3). Maheshwari (2015, p.9) las define como series da datos accesibles de diversas formas.

Las fuentes de datos suministran datos a otras herramientas del sistema de información. Las fuentes de datos pueden ser bases de datos o tener otros formatos o características. Por ejemplo, se pueden emplear páginas web como fuentes de datos. Puede tratarse del espacio donde se han generado los datos o no. Pueden emitir los datos un número contado de veces o de forma regular. El ejemplo lo encontramos en el ERP, pues cuando se hace una implantación de este sistema en una empresa se migran datos del sistema antiguo, como los activos, el asiento de apertura o las facturas no saldadas. En este caso, el sistema original es la fuente de datos y se realizan los envíos una o varias veces determinadas. Sin embargo, por circunstancias de la empresa, se puede requerir que los usuarios sigan registrando las facturas en el programa antiguo, mientras que, todas las demás transacciones se registrarán en el ERP. Por ello, se establece un proceso automático en el que cada cierto tiempo, las facturas creadas en el antiguo sistema se envíen al ERP. En este caso sería una fuente de datos de forma regular.

3.3. La variedad de los datos

Como ya se ha comentado, una de las características de los datos en la era de la información es la variedad. Los datos que se emplean en las empresas tienen diferentes orígenes y formatos. Por ello, una tarea importante es la homogenización y adaptación de los datos a sus necesidades. Las trasformaciones a realizar dependen de las características de los datos de origen, por ello es importante caracterizarlos.

Los datos pueden provenir, según la estructura, de fuentes de datos estructuradas, semiestructuradas o no estructuradas. Esta clasificación se detalla en los apartados siguientes.

Las fuentes de datos operacionales tradicionales son los programas de procesamiento de transacciones (TPS), los ERP (Enterprise Resource Planning) y las aplicaciones de gestión de las relaciones con los usuarios (CRM) (Minelli, Chambers, & Dhiraj, 2013, p.10). Hoy en día el número de fuentes de datos que maneja una empresa se ha incrementado y diversificado de forma que incluye:

- Datos de internet: p.ej. redes sociales y analítica web.

- Investigación primaria: llevaba a cabo por la propia empresa, p.ej. encuestas.
- Investigación secundaria, llevada a cabo por terceros, p.ej. estudios de mercado, de la industria o del consumo.
- Datos de localización.
- Datos de imagen, como la de los satélites.
- Datos de la cadena de suministros, como el intercambio electrónico de datos (EDI).
- Datos de dispositivos, como la radio fecuencia.

Además, pueden tener distintos formatos: texto, vídeo, audio, imagen, etc.

3.4. Fuentes de datos estructuradas

Los datos proporcionados por fuentes estructuradas siguen un esquema que permite almacenarlos en tablas con campos (columnas) y registros (filas). Estos datos generalmente tienen una longitud y un formato definido (Kaufman, 2013, p.72), por lo que son fáciles de definir, almacenar y analizar (Minelli et al., 2013, p.11) Los números, las fechas y las cadenas de estos caracteres, como los nombres de clientes o las direcciones, son datos estructurados (Kaufman, 2013, p.72). Se almacenan en hojas de cálculo, tablas y bases de datos relacionales, en ellas están definidas las propiedades y relaciones entre ellas (Phillips-Wren, Iyer, Kulkarni, & Ariyachandra, 2015, p.23).

Se trata generalmente de bases de datos relacionales (Sánchez Sánchez, Pan Bermúdez, & Viña Castiñeiras, 2004, p.4). Estas se caracterizan por:

- La persistencia de datos consiste en que la información siga disponible aunque cerremos la pantalla o aunque haya terminado el periodo temporal al que se refiere (Kaufman, 2013, p.75).
- La base de datos tiene una estructura que define las tablas, las etiquetas de las columnas de las tablas (campos) y las relaciones entre ambos (Kaufman, 2013, p.76).
- Los datos se estructuran en filas (registros).
- Cada tabla tiene, al menos, un campo clave sin duplicados.
- Las tablas de datos relacionales pueden ser consultadas con el lenguaje de consulta SQL.

Aunque pueda parecer que utilizar distintas tablas aumenta la información a almacenar, ocurre todo lo contrario. Pues si separamos, por ejemplo, la tabla de clientes de la de facturas, evitaremos repetir en cada registro de una factura el nombre del cliente, sus datos bancarios, su dirección, etc. Simplemente a cada factura se le asociará un número identificador del cliente y en la tabla de clientes estará ese número con todos los datos del mismo, almacenados una sola vez. Lo más importante de este sistema es que se eliminan duplicidades.

Los paquetes de office, de Microsoft o de software libre incluyen un gestor de bases de datos relacionales. En el caso de Microsoft, se denomina Access y permite cargar tablas, relacionarlas y consultarlas.

3.5. Fuentes de datos semiestructuradas

Las fuentes de datos semiestructuradas no tienen un orden tan claramente definido como las estructuradas, pero sí puede intuirse un cierto esquema, aunque flexible (Sánchez Sánchez et al., 2004, p.4).

Contienen datos semiestructurados: los ficheros XML, logs (informes que generan procesos informáticos con información sobre su ejecución, especialmente útiles cuando los procesos terminan en error porque describen el mismo), algunas páginas web (Sánchez Sánchez et al., 2004, p.4), EDI (Intercambio Electrónico de datos), SWIFT (Society for Worldwide Interbank Financial Telecommunication, el formato característico de los archivos financieros como los bancarios) (Kaufman, 2013, p.8).

3.6. Fuentes de datos no estructuradas

Las fuentes de datos no estructuradas no presentan un orden en los datos comparable a los de las tablas, sino que su formato es libre, sin que siga ninguna norma específica en su estructura.

Las colecciones de datos no estructurados no siguen un formato específico (Kaufman, 2013, p.78), por ello, no suelen ser fáciles de definir, necesitan mucha capacidad del almacenamiento y habitualmente son más difíciles de analizar (Minelli et al., 2013, p.11). Este tipo de datos no encajan en las bases de datos relacionales p.11.

Son datos no estructurados algunos documentos de texto, las páginas web estáticas (Sánchez Sánchez et al., 2004, p.3), algunos datos científicos, las imágenes, los vídeos, mucha información interna de las empresas como los correos electrónicos, los datos generados en las redes sociales como comentarios, datos generados con el teléfono móvil, como audios (Kaufman, 2013, p.80), etc.

CAPÍTULO 4. Big Data y Business Intelligence

4.1. Introducción

En ese apartado se definirán por separado el Big Data y el Business Intelligence a partir de una revisión bibliográfica. Luego se identificarán semejanzas y diferencias entre estas dos disciplinas. Y, por último, se analizará si se pueden implantar realizando la misma serie de tareas.

4.2. Definición de Big Data

El Big Data, macrodatos o datos masivos se refiere, de acuerdo con McKinsey Global Institute (2011, p.1), a conjuntos de datos que por su gran tamaño superan las capacidades de las herramientas tradicionales para capturar, almacenar, gestionar y analizar datos. McKinsey se centra en el volumen de los datos, que es a lo que también hace referencia en el propio nombre de Big Data. En el mismo informe en el que lo define, el Instituto de la importante consultora americana comenta que las herramientas de gestión de bases de datos han evolucionado considerablemente en los últimos años aumentando su capacidad. Incluso esta capacidad depende del sector del que se esté hablando (McKinsey Global Institute, 2011, p.1). A pesar de todo ello, actualmente lo más importante del Big Data no es el tamaño, por lo que, se debe buscar otra definición.

Gartner, la importante consultora en el área tecnológica, define el Big Data como la información de gran volumen, alta velocidad y gran variedad que demanda un coste eficiente y formas innovadoras de procesar información que permitan alcanzar ventajas competitivas, tomar decisiones y automatizar procesos (Gartner, s. f.). De esta definición cabe comentar dos puntos: la inclusión de otras características del Big Data a mayores del volumen y la utilidad del Big Data.

En primer lugar, Gartner emplea para definir el Big Data tres de sus características, conocidas como las 3V's y que son: volumen, velocidad y variedad. A estas, diversos estudios han añadido la veracidad y el valor, alcanzando las 5V's (Figura 4):

- Volumen. A pesar de su nombre, el Big Data se centra cada vez menos en el tamaño de los datos y es que, los avances tecnológicos nos permiten procesar datos de la talla del pentabyte en nuestros ordenadores personales (Murillo González, 2016, p.8). Además, el almacenamiento es cada vez más barato e incluso, se puede contratar como servicio en la nube.
- Velocidad. Se refiere a la velocidad con la que se generan los datos y también a que se espera que sea procesados en tiempo real (Anuradha J; Ishwarappa, 2015, p.320-321).
- Variedad. Los datos se pueden presentar estructurados o no y también en diversos formatos. Los datos pueden ser textos, números, imágenes, vídeos, audios…
- Veracidad. Todos podemos generar datos, por ejemplo, publicando en nuestro muro de Facebook, y para ello, no necesitamos pasar ningún control ni tener fuentes contrastada, como sí le ocurrirá a un periodista. Por ello, surgen noticias falsas o fake news. Esto hace que en la era de la información saber distinguir las fuentes fiables sea una habilidad remarcable.
- Valor. Se trata quizás del aspecto más importante, de que los datos permitan a las empresas crear valor (Anuradha J; Ishwarappa, 2015, p.321). Para ello, la empresa debe transformarlos en información y gestionar el conocimiento.

Figura 4 Características del Big Data

Fuente: elaboración propia

En conclusión, y a pesar de llamarse Big Data o macrodatos, de sus cinco características la del volumen sería la menos importante, dados los avances tecnológicos en almacenamiento y velocidad de procesamiento.

En segundo lugar, de la definición de Gartner, destaca la proposición sobre lo que permite el Big Data. Este punto está muy relacionado con la quinta característica del Big Data: el valor. Y es que, para las empresas, y también para las personas, es muy importante que la información que consuma les reporte algún beneficio. La empresa tecnológica destaca que con el Big Data se generan ventajas competitivas, se mejora la toma de decisiones y se automatizan procesos. En este punto es importante ver el Big Data como parte de los Sistemas de Información de la empresa. Una parte importante porque aporta al sistema sus cinco características y también, porque permite automatizar tareas repetitivas que antes debían hacer los empleados. Por ejemplo, tradicionalmente, cada vez que una empresa iba al banco a pedir un crédito con un balance, un empleado del banco debía calcular diversas ratios de acidez y solvencia. Con las nuevas tecnologías y el Big Data, se pueden generar procesos que traten automáticamente los datos obteniendo los cálculos que el banquero necesita. Esto puede hacerse incluso sin tener que abrir el fichero.

Según otros autores, lo importante del Big Data es generar valor empleando la gran cantidad de datos disponibles. Además, se deben buscar métodos fiables y eficientes (Minelli et al., 2013, p.1).

Por todo ello, las herramientas de Big Data son aquellas que procesan datos muy variados (de diversas fuentes y formatos), a gran velocidad y en paquetes de gran tamaño (en relación con lo que se procesaba tradicionalmente) donde su gestión se centra en garantizar su fiabilidad y en conseguir obtener de ellos la mayor información posible para tomar las mejores decisiones informadas, obtener ventajas competitivas y, de esta forma, generar valor para la empresa.

4.3. Definición de Business Intelligence

De acuerdo con Mochón y Gonzálvez, Business Intelligence son los procesos y técnicas para sacar partido a los datos, convirtiéndolos en información y luego en conocimiento (Mochón Morcillo & Gonzálvez Cabañas, 2016, p.17).

Cuatro empresarios con una patente sobre datos desestructurados, definen el Business Intelligence como "technologies, applications and practices for collection, integration, analysis, and presentation of content such as business information" (Guha, Wrabetz, Wu, & Madireddi, 2012, background). La extracción, la integración, el análisis y la presentación del contenido facilitan convertir los datos en información y esta en conocimiento. Concretamente, la extracción y la integración suponen pasar de datos a información. El análisis de los datos contribuye a generar información y el de los datos a generar conocimiento. Y la presentación de la información ayuda a difundirla y a que los individuos la empleen para generar conocimiento. Por lo que, coincide con la definición de Mochón y Gonzálvez en este punto y añade las actividades a realizar para llegar del dato al conocimiento.

Maheshwari define esta disciplina como el conjunto de una serie de aplicaciones de tecnologías de la información que se emplean para analizar datos y transmitirles información a los usuarios (Maheshwari, 2015, p.21). Por lo tanto, este autor limita la utilidad del Business Intelligence a la información, cuando contribuye en gran medida a la generación de conocimiento, por ejemplo, transmitiendo y visualizando la información. Aunque considera que el paso de información a conocimiento se da en la mente del individuo y sin él esto no es posible. El BI ayuda al individuo a comprender esta información e incluirla en sus patrones metales, pero no puede asimilarla por él.

Este autor también define cuatro componentes del BI: "data warehousing, data mining, querying and reporting" (Maheshwari, 2015, p.21). De forma similar, Mochón Morcillo incluye en BI consultas en bases de datos, informes, análisis y herramientas para mostrar información (Mochón Morcillo & Gonzálvez Cabañas, 2016, p. 51-53). Como se verá más adelante, el data mining forma parte de las tareas de análisis, las queries son las búsquedas en bases de datos y el reporting, la generación de informes. El data warehousing que solo menciona Maheshwari se refiere al almacenamiento. Y el cuarto componente de Mochón Mocillo es la visualización.

En cuanto a la finalidad del Business Intelligence, "data-based decisions are more effective than those based on feelings alone" (Maheshwari, 2015, p.21). Las ciencias sociales no son una ciencia exacta, sin embargo, la mayoría de las veces una decisión basada en el conocimiento nos lleva a alcanzar los objetivos, al menos con mayor probabilidad que una basada en sentimientos. Maheshwari dice que se basa la decisión en datos, sin embargo, como se ha expuesto, los datos carecen de significado. Por ello, en el comentario de su afirmación se ha empleado el término conocimiento. Las decisiones basadas en el conocimiento si son más efectivas. También lo son las basadas en información, aunque en menor medida, pues que no llegue a conocimiento significa que no la hemos asimilado ni ligado a nuestra experiencia. Hay que tener en cuenta que todo este asunto está muy ligado y se puede ver influenciado por la veracidad de la información.

Para Davis Loshin, la finalidad del BI está en aplicar esos conocimientos para resolver problemas (Loshin, 2012, p.36). Esta proposición es similar a la de Maheshwari, quizás más limitada. Pues los problemas se pueden resolver con decisiones, pero hay más decisiones que se pueden tomar en la empresa basadas en conocimientos que no tienen por qué ver con problemas, puede tratarse de innovaciones, por ejemplo.

En conclusión, el Business Intelligence es un área de las tecnologías de la información que se centra en la transformación de los datos en conocimiento. Para ello, se extraen los datos, se almacenan, se analizan, se tratan y se prepara la información generada para ser visualizada por el usuario.

4.4. Big Data y Business Intelligence: diferencias y similitudes

En primer lugar, antes de hacer una revisión bibliográfica sobre lo que diversos autores opinan sobre las diferencias y similitudes entre estas dos disciplinas, se comparan las definiciones y características vistas en los dos apartados anteriores y que se resumen en la Tabla 1.

Tabla 1 Big Data y Business Inteligence. Definiciones.

	BD	BI
Son…	Datos, información	Tecnologías, aplicaciones y prácticas.
Permiten…	Procesar gran volumen de datos diversos a gran velocidad.	Transformar datos en conocimiento.
Finalidad	Tomar las mejores decisiones informadas, obtener ventajas competitivas y generar valor para la empresa.	Tomar decisiones basadas en conocimiento, que son más efectivas.
También es importante	La fiabilidad	Las fases del proceso: extracción, almacenamiento, análisis, tratamiento y visualización.

Fuente: elaboración propia

Por lo tanto, centrándonos en la revisión bibliográfica previa, la definición de BD se centra en la materia prima: los grandes volúmenes de datos variados generados a gran velocidad y de los que preocupa la fiabilidad, mientras que, la de BI lo hace en los procesos y herramientas para transformar datos en conocimiento. Se puede decir que hablan de lo mismo, pero con focos distintos: la definición de BD se centra en las características de los datos que se manejan (las 5Vs) y el BI en el proceso para transformar estos en conocimiento. Ambas definiciones coinciden en que el objetivo es tomar decisiones informadas: en la de BD se dijo que para obtener ventajas competitivas y generar valor; y en la de BI para tomar decisiones más efectivas. Y las decisiones son efectivas en las empresas cuando se cumple su fin último: ganar dinero. Lo que se puede traducir en crear valor. En términos financieros, el valor se crea cuando con unos recursos que valen X unidades monetarias se consigue un producto o un servicio que vale Y, siendo Y mayor que X, esa diferencia es el beneficio o valor. Así que, una decisión efectiva es aquella que crea valor. Por ello, se concluye que la finalidad de BI y del BD es la misma. Y de esta comparación, se concluye que se trata de conceptos muy relacionados y poco diferenciados. La delimitación de los dos conceptos se abordará ahora con la revisión bibliográfica sobre los escritos que tratan la relación entre estas disciplinas, y no las definiciones de cada una como se ha hecho hasta ahora.

De acuerdo con Minelli y otros, la diferencia entre Business Intelligence y Big Data es la información que emplean. Afirman que el BI básicamente analiza información interna de la empresa, muy estructurada y de origen transaccional y el BD estudia diferentes fuentes de datos que pueden provenir de fuentes no estructuradas y/o externas (Minelli et al., 2013, p.17). El ejemplo del catálogo de BI de Oracle, incorporado en el ERP, es un ejemplo que respalda esta idea. En el que solo se trabaja con los datos del propio software de información interna de la empresa. Esta herramienta permite realizar informes a partir de Análisis con variables predefinidas o, conociendo el lenguaje de consulta SQL y las tablas de la base de datos, todos los datos del ERP. Aunque, no siempre las empresas diferencian correctamente términos a la hora de vender sus productos (Ruiz del Castillo, 2019) sobre todo cuando se trata de términos que, como se está observando, son muy similares.

Kaufman, en la misma línea que los autores anteriores, asimila el Business Intelligence al análisis tradicional, pensado para datos muy estructurados y fáciles de entender. Además, dice que se aplica a una parte concreta de los datos, no ha todos los disponibles. Mientras que el Big Data puede ser estructurado, semiestructura o no estructurado y complejo. Además, el BD permite analizar datos en tiempo real (Kaufman, 2013, p.252- 253).

Mochón y Gonzálvez están de acuerdo en que las acciones realizadas en ambas disciplinas coinciden (procesar, gestionar, almacenar y analizar datos) y el objetivo (convertirlos en conocimiento y crear valor). También coinciden en que el BI trata datos históricos y estructurados de fuentes internas y el BD tiene capacidad para trabajar con datos en tiempo real y/o no estructurados, cuya fuente puede ser interna o externa. Además, añade que el BI busca detalles en los datos, mientras que el BD busca tendencias globales (Mochón Morcillo & Gonzálvez Cabañas, 2016, p.51-53).

Guha y otros autores (2012, background), confirman, en el texto de su patente, que las herramientas tradicionales de BI no son capaces de procesar datos no estructurados. Además, dicen que las soluciones manuales y de automatización no son viables. Las primeras nos difíciles de aplicar a grandes volúmenes de datos y es más probable que ocurran errores (Guha et al., 2012, background). Las segundas pueden ocultar errores de precisión. Estos autores proponen y patenta su propio método y sistema para el análisis con herramientas BI de datos no estructurados. Dada la información anterior, se ha llegado a la conclusión de que el tradicional no trata datos no estructurados si no han sido tratados previamente, pero una vez que se tratan es posible emplearlos. Por lo que, si no se incluyen las herramientas de transformación dentro del BI, para datos no estructurados requerirá que esta función se haga desde otra disciplina.

En resumen, el Business Intelligence y el Big Data son dos partes muy similares de los Sistemas de Información de las empresas. Se trata de herramientas que permiten transformar los datos en conocimiento para basar en este las decisiones y así, crear valor para la empresa. Estas herramientas realizan un proceso de obtención de los datos, tratamiento, gestión, almacenamiento, análisis y creación de elementos para su visualización. El Business Intelligence se centra en el análisis y la visualización de los datos, aunque también puede importar directamente datos de fuentes estructuradas, para los otros tipos de fuentes necesitará una transformación previa que estas herramientas no pueden realizar. Mientras que el Big Data surgió por el aumento del volumen de datos procesados que requirió herramientas diferentes. Y se ha extendido para tratar otras características de los datos muy importantes en la era del conocimiento como son: la velocidad, la variedad y la veracidad. Por lo que el BD, aunque incluye tareas de análisis y visualización, se centra en la extracción y el tratamiento de los datos, las fases más relacionadas con las características de estos como puede ser la variedad, de hecho, puede tratar datos no estructurados que el BI no. Además, los análisis de BI se centran en los detalles y los de BD en tendencias globales. Por lo tanto, se trata prácticamente de lo mismo, pero dependiendo del término indicado se hace hincapié en una parte u otra de la tarea y existen unas limitaciones. En consecuencia, los procesos empleados para implantar estas herramientas serán muy similares.

CAPÍTULO 5. Proceso de implantación de Big Data y Business Intelligence

5.1. Introducción

En este capítulo se establecerá, revisando dos publicaciones características de los estudios realizados al respeto, el proceso de implantación de las herramientas de los Sistemas de Información que nos ocupan. Se emplea un solo proceso de implantación para Big Data y Business Intelligence porque, como se ha concluido en el apartado anterior, se trata de herramientas que atienden el mismo proceso, pero haciendo énfasis en distintas fases y haciendo diferentes análisis. Este proceso se puede definir también como el de generación de conocimiento, que va un paso más allá que las herramientas de BD/BI, incluyendo la utilización de la información resultante para la toma de decisiones.

5.2. Estudio de Dutta y Bose

Los investigadores del Indian Institute of Management Calcutta (Dutta & Bose, 2015, p.295) establecen un marco de trabajo para aplicar proyectos de Big Data diferenciando tres fases principales: la estrategia, el análisis de datos y la implementación. En la Tabla 2 se muestran estas fases divididas en tareas:

Tabla 2 Marco de trabajo para aplicar proyectos de Big Data

1. Estrategia	2. Análisis de datos	3. Implementación
1.1. Problema del negocio	1.1. Obtener los datos y examinarlos	3.1. Integra las herramientas en el sistema de información de la empresa
1.2. Investigación	1.2. Análisis de datos y modelación	3.2. Formar a los empleados
1.3. Formación de un grupo multidisciplinar	1.3. Visualización de los datos	
1.4. Planificación del proyecto		

Fuente: (Dutta & Bose, 2015)

Las fases de estrategia e implementación son comunes a la mayoría de procesos y actividades a realizar en las empresas, pues en la administración de organizaciones siempre es importante definir los objetivos con los que se hacen las cosas, analizar previamente el reto a resolver, investigar las posibles soluciones, elegir una y planificar la aplicación de la solución. En la era del conocimiento es muy importante la implicación de las personas en el proceso y la formación de grupos multidisciplinares.

Que la estrategia sea algo común en las actividades y procesos de la empresa no le resta importancia en la implantación de soluciones de BI y BD, ya que, son aspectos clave para que estos tengan éxito. En esta fase se determina si realmente son necesarias estas herramientas, cuales en concreto se emplearán, cómo se utilizarán, etc.

En cuanto a la implementación, consiste en emplear en la toma de decisiones los conocimientos adquiridos en el análisis. Dutta y Bose (2015, p.295) destacan en este punto la formación de los empleados. Se trata de una tarea importante ya que serán estos los que deban emplear la información y para ello deben conocer que datos se están empleando, que modificaciones o adaptaciones han hecho y cualquier

característica de los mismos y del proceso que contribuya a mejorar la comprensión de los datos visualizados. En definitiva, todo lo que les pueda ayudar a convertir la información en conocimiento.

Sobre la segunda fase del proceso de Dutta y Bose (2015, p.295), a la que denominan "análisis de los datos", a pesar del título, esta fase incluye en la obtención, el análisis y modelación y la visualización de datos. A juicio de la autora de este trabajo, y de acuerdo con la bibliografía consultada para redactar los apartados anteriores, esta fase incluiría también el tratamiento de los datos y su almacenamiento.

Además, en esa última fase de implementación, se considera importante añadir un espacio para la evaluación del proceso realizado, pararse a pensar si los resultados satisfacen realmente las necesidades de la organización. Esta tarea tiene el objetivo de llevar a cabo una mejora continua del proceso, convirtiéndolo en un ciclo, ya que, con frecuencia, se detectarán puntos de mejora que requerirán volver al inicio del proceso.

5.3. Patente de Guha, Wrabetz, Wun y Madireddi

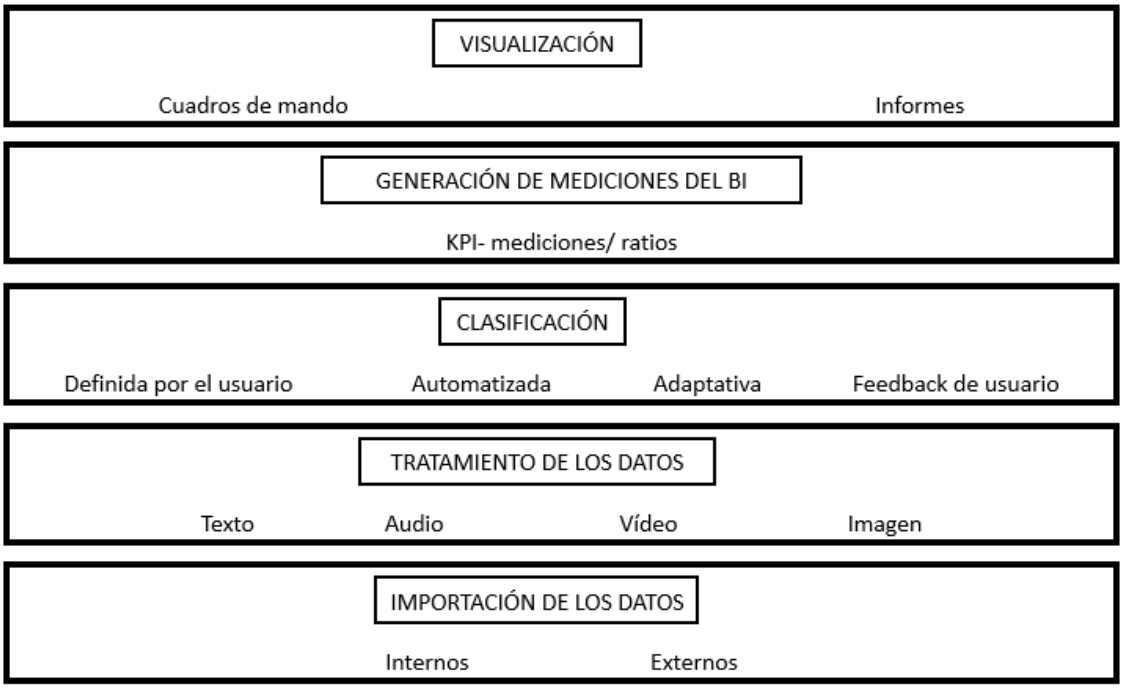
Guha, Wrabetz, Wu y Madireddi (2012, p.295) han definido el proceso de implantación de herramientas de Business Intelligence con la intención de establecer un marco para emplear estas herramientas con datos no estructurados.

Este comienza (de abajo a arriba en la Figura 5) por la importación de los datos, que pueden ser de fuentes internas o externas. Esos datos pueden tener diversos formatos, por ejemplo, los archivos que se pueden crear con las aplicaciones habituales en cualquier ordenador: textos, hojas de cálculo, presentaciones de diapositivas, dibujos, etc.

Una vez tratados se estable una estructura o clasificación, bien porque la definen los usuarios, bien porque lo hace una máquina en base a patrones, bien porque se adapta una estructura existente al fichero o bien por correcciones del usuario a clasificaciones anteriores.

Cuando ya la materia prima está preparada, se calculan una serie de ratios o índices relevantes para la solución del reto. A estos indicadores se les denomina KPIs, por las siglas de Key Performance Indicators. Y, por último, se muestra la información empleando dashboards o cuadros de mando e informes.

Figura 5 Sistema para implantación de BI para fuente no estructuradas



Fuente: Guha et al., 2012

A esta última propuesta se le añade una fase de análisis en paralelo a las mediciones del BI. De esta manera, se hace hincapié en la busca de patrones por parte del BD, ya que, los detalles que busca el BI se entienden cubiertos por estos KPI.

Otro punto a destacar es la obtención de los datos, la cuál tiene que ver con la diferenciación mencionada en la fase de importación por Guha y otros. Aunque la obtención o extración está muy relacionada con el tratamiento ao la extracción, cabe recordar que en esencia son acciones diferentes.

Además, se añadiría la clasificación en el tratamiento de los datos porque se decidirá como clasificarlos en la fase previa de estrategia y la execución de esta decisión no llevará tanto en tiempo como para dedicarle una fase.

5.4. Proceso resultante de la revisión bibliográfica

A los procesos analizados en los dos apartados anteriories, se incluyen, como actividades tranversales, es decir, que no siguen un orden secuencial, sino que están presentes en varios puntos del proceso: la gestión de la base de datos y el almacenamiento.

Se entiende la gestión como la administración de las personas que tienen acceso a la información, las que son responsables de ellos y quién debe hacer que tareas de su transformación, análisis y visualización. En este punto, dado que muchas empresas subcontratan estos servicios de obtención, tratamiento, almacenamiento, análisis y visualización de los datos, es importante gestionar la coordinación entre la empresa propietaria de los datos y la asesoría que se encarga de las herramientas de BI/BD.

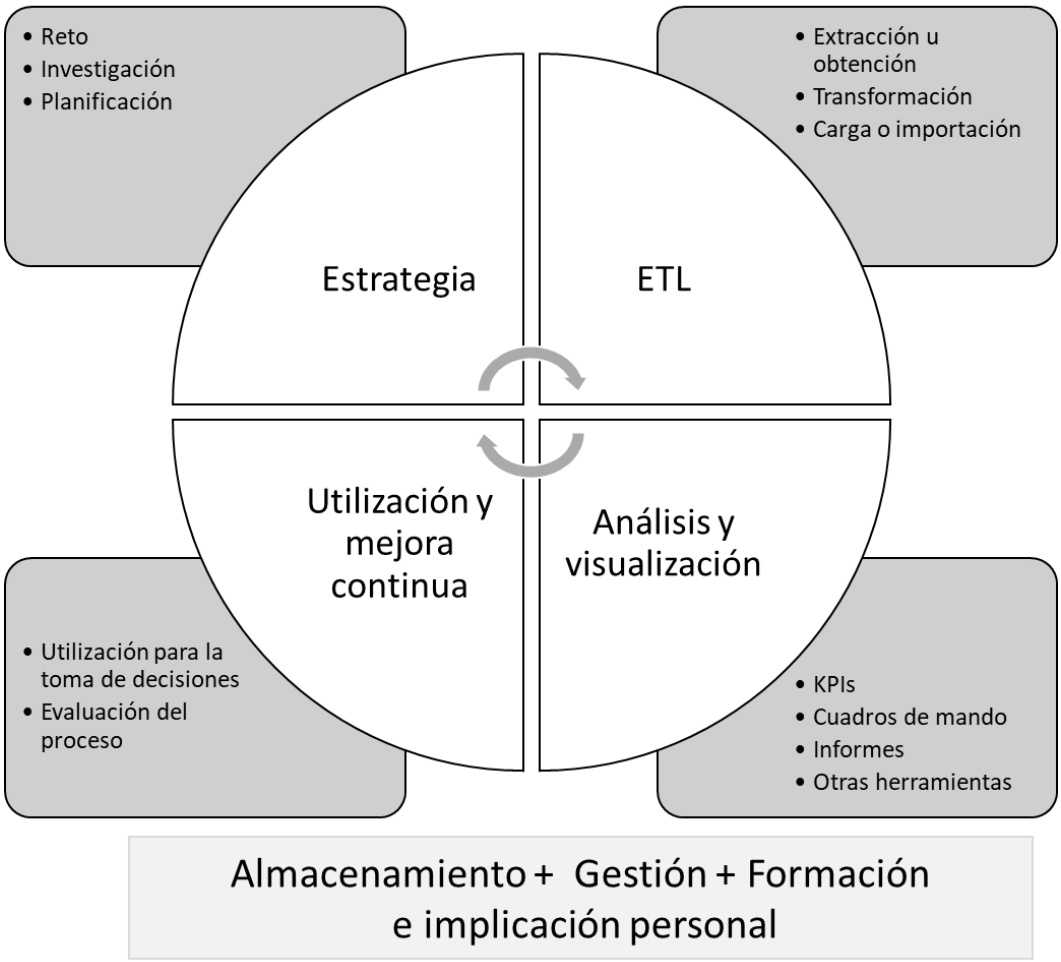
En cuanto al almacenamiento, destaca la elección de una plataforma en la nube, donde otra empresa alquila parte de un servidor a la dueña de los datos; o una solución on premise, en los propios servidores de la empresa. Las decisiones de almacenamiento, tanto de lugar de almacenamiento como de forma, se toman principalmente al inicio en la estrategia y tienen una importante repercusión en el proceso.

Por último, se incluye la formación del personal de la empresa fuera del orden secuencial, ya que, se considera que debe ser continuada y que incluso, se puede empezar antes de que los datos esten disponibles para la visualización, informándolos y teniendo en cuenta sus opiniones en la definición de la estrategia. También se incluye en ese punto no secuencial el grupo multidisciplinar del que hablaban Dutta y Bose en la estrategia de la empresa.

Además, se ha decidio unir las fases de análisis y visualización, dado que, comparten herramientas y sus fronteras son difusas. Y se han agregado la extración u obtención, la transformación y la carga o importación, pues existe un proceso conocido como ETL que los agrupa. Se debe tener en cuenta que el tratamiento se puede realizar tanto antes como después de la importanción.

En la Figura 6 se resume el proceso final resultante de esta revisión.

Figura 6 Proceso de implantación de herramientas de BI y BD



Fuente: elaboración propia

En este trabajo se desarrolla la fase de investigación incluida en el marco de la estrategia. Como base e hilo conductor de la investigación se utilizan el proceso definido en la Figura 6, de forma que, las cuatro grandes actividades corresponden con los próximos capítulos del TFM.

CAPÍTULO 6. Estrategia y consejos para la implantación

6.1. Introducción

Como se definió en el apartado anterior, la primera fase del proceso de BI y BD es la definición de la estrategia. En este apartado se expondrá por qué es necesario llevar a cabo esta fase. También se explicarán las subfases de reto, investigación y planificación.

6.2. Necesidad de una estrategia

Para implantar herramientas de BI y BD en una empresa es necesario definir qué objetivos busca conseguir la empresa con ellos. Y, partiendo de dicho objetivo, se debe determinar la información que necesita y cuáles son las fuentes de las que se debe extraer. Además, una vez obtenidos los datos, se debe confirmar que son útiles para conseguir los objetivos y si no, realizar las correcciones oportunas (Kaufman, 2013, capítulo 15).

Es necesario aclarar, que no siempre es adecuado el uso de estas herramientas de BI y BD. Solo lo es cuando los retos a los que se enfrenta la empresa justifiquen un gran volumen de datos a tratar, gran velocidad de procesamiento o respuesta, o bien, gran variedad de datos. También se debe tener en cuenta la evolución esperada del reto a enfrentar (Mochón Morcillo & Gonzálvez Cabañas, 2016, p.127-144), si es algo puntual o a largo plazo. Por ello, es tan importante definir los objetivos y lo que se espera lograr, no solo para aumentar las probabilidades de éxito de la implantación, sino también, para determinar si realmente es lo que necesita la empresa.

6.2.1. **Reto**

Se trata de identificar el reto al que se enfenta la empresa y que pueden aportar las soluciones de BI y BD. Es el momento ideal de la estrategia para tener en cuenta las expectativas de los grupos de interés (Dutta & Bose, 2015, p.295), como por ejemplo, los empleados que usarán la información o los accionistas que recibirán algunos informes.

Este reto o problema del negocio es la necesidad de información a partir de la cuál y para la cual se realizará todo el proceso de implantación. Solo es razonable implantar estas herramientas si van a generar un valor para la empresa. De manera general, este tipo de herramientas sirven para ayudar en la toma de decisiones, por ejemplo, en la fijación del precio de los productos.

6.2.2. **Investigación**

Bajo este nombre Dutta y Bose explican una base de búsqueda de información dos aspectos: cómo otras empresas han resuelto estos problemas y que herramientas hay disponibles con tales fines (Dutta & Bose, 2015, p.295). Como se puede observar, estos dos aspectos se interrelacionan y podría encontrarse información de ambos en las mismas fuentes.

Observar cómo otras empresas han resuelto estos problemas no es más que llevar a cabo Benchmarking, una herramienta empresarial de comparación con la propia empresa (datos históricos), con empresas competidoras o con los líderes del sector. El Benchmarking permite a los directivos de una empresa estudiar si podrían realizar de manera más eficaz y/o efectiva diversas tareas en diversas áreas, por supuesto, también en Big Data y Business Intelligence.

En cuanto al segundo aspecto, buscar que herramientas hay disponibles permite no solo determinar si la empresa podrá obtener los datos que quiere de la forma que quiere, también estudiar los costes de las distintas aplicaciones y ver, según su presupuesto, que se puede permitir.

6.2.3. **Planificar el proyecto de implantación**

La planificación del proyecto consiste en identificar las actividades a realizar, los plazos y las personas que realizarán las tareas. También se presupuestan los costes. De cara al desarrollo del proyecto lo más importante es que se trate de una planificación flexible, sujeta a cambios según se vaya realizando el proceso y se vayan descubriendo nuevas necesidades y limitaciones. Además, para llevar a cabo un correcto seguimiento del proyecto se deben establecer puntos de control (Dutta & Bose, 2015, p.295).

6.3. Consejos en la implantación de Big Data

A la hora de implantar herramientas de Big Data, es importante (Mochón Morcillo & Gonzálvez Cabañas, 2016):

- Definir objetivos ambiciosos, específicos y medibles.
- Centrarse en los grandes retos y de mayor valor. Es mucho más útil y reduce el riesgo de fracaso centrarse en un reto grande con posibles altos beneficios que centrarse en pequeños con poca repercusión.
- Empezar definiendo a qué preguntas se les quiere dar respuestas o que problema se quiere abordar, pero no que datos se van a emplear.
- Implicar grupos multidisciplinares para que aporten distintos puntos de vista, también a los que van a usar las herramientas y a los directivos y mandos intermedios que deben fomentar su uso.
- Formación del personal.
- La implantación de herramientas más complejas no supone dejar de emplear hojas de cálculo o gráficos simples, ambas herramientas deben complementarse.
- Planificar la implantación dividiéndola en diversas tareas con plazo, para evitar perderse en el análisis.
- Planificar la gestión de la información a largo plazo. Revisar periódicamente si las herramientas siguen siendo lo que necesitan.

Y para tomar decisiones con los datos es importante conocer las fases anteriores por las que han pasado los datos y conocer sus fuentes originales, de forma que el conocimiento generado sea lo más veraz posible.

Para interpretar la información resultante de la aplicación de herramientas de BI y BD se deben tener en cuenta (Mochón Morcillo & Gonzálvez Cabañas, 2016, p.88-96):

- La correlación no es causalidad. Con cálculos muchas cosas son posibles, le toca a la persona determinar si está ante una relación de causalidad o ante una relación espuria, para ello, debe aplicar su experiencia.
- Limitaciones en las redes sociales. Muchos de los datos que se generan hoy en día sobre las empresas provienen de las redes sociales. Es importante analizarlos y que la empresa sepa cuál es su reputación en esto canales. A la hora de hacer este análisis se debe tener en cuenta que:
    - Los usuarios comparten información limitada por las normas de la red social, por ejemplo, los 300 caracteres de Twitter.
    - Un solo usuario puede tener varias cuentas y una cuenta la pueden emplear varios usuarios.
    - Hay cuentas spam o falsas.

- La información de la ubicación es poco fiable, no todo el mundo la comparten y si lo hacen no suele ser correcta.
- Dificultades para analizar si un comentario es positivo, negativo o neutro de forma automática.
- La red social más utilizada, la forma o los fines con los que se emplea cambian continuamente.

CAPÍTULO 7. ETL y almacenamiento

7.1. Introducción

En este apartado se definirá el ETL y se explicarán las fases de obtención, importación y tratamiento del proceso de generación de conocimiento con BI y BD. En las fases se incluirán técnicas y herramientas que se pueden emplear para llevarlas a cabo.

Además, en el último epígrafe del capítulo se caracterizará el almacenamiento, en dónde se importan los datos. Esta fase se ha definido como no secuencial, sino transversal a todo el proceso de generación de conocimiento. Con respecto al almacenamiento se tratan: el almacén de datos o data warehouse, el Hadoop y almacenamiento en la nube o local.

7.2. Definición de ETL

Las fases de extracción, transformación y carga de datos también se conocen por sus siglas en inglés: ETL (Extract, Transfer and Load). Se trata de recopilar datos de varios orígenes, transformarlos de acuerdo con las necesidades del negocio y cargarlos en la base de datos de destino. Estas tres fases pueden ejecutarse en paralelo para ganar tiempo (Tejada, 2018).

En ocasiones, la transformación se realiza en el almacén de datos de destino, por lo que, la carga se produce antes que la transformación. En este caso, estamos hablando de ELT (Extract, Load and Transfer; extracción, carga y transformación).

7.3. Obtención de los datos

Dependiendo de si las fuentes de datos son internas o externas o de si se trata de datos estructurados, semiestructurados o no estructurados, las herramientas necesarias para obtener o extraer los datos pueden ser muy diferentes.

Los datos internos y estructurados, como la contabilidad de la empresa o los datos de los clientes y proveedores, son, por lo general, sencillos de obtener y apenas necesitan edición. Esto se debe a que la empresa ya los conoce, de alguna manera los tiene personalizados y adaptados a su ciclo de negocio, y por ello, no requieren demasiados cambios.

Sin embargo, los datos externos, incluso los que ya están estructurados, pueden requerir diversas tareas para su obtención y tratamiento. La casuística es muy variada según de que fuentes se trate:

- Si se quiere extraer la información de una página web, se empleará el web scraping, explicado en el apartado siguiente.
- Si se quiere migrar datos de fuentes estructuradas se emplean consultas en lenguaje de bases de datos.
- Algunas fuentes como el Instituto Nacional de Estadística (INE), permiten extraer los datos en diversos formatos, generalmente en .xls (el de las hojas de cálculo) o .csv (documentos de texto separado por comas).

7.4. Web scraping

De acuerdo con la consultara en análisis de datos Aukera, el web scraping consiste en “navegar automáticamente a una web y extraer de ella información” (Lafuente, 2019). La empresa referente en esta actividad es Google, que está constantemente scrapeando la web entera como base para el funcionamiento de su buscador (Lafuente, 2019).

El software encargado de extraer los datos se denomina bot, spider o crawler y existen herramientas que permiten utilízalos sin conocer programación, aunque sus funcionalidades son limitadas (Lafuente, 2019).

Los crawlers, además de para extraer información, pueden programarse para rellenar formularios, crear cuentas falsas o realizar cualquier acción en la red de forma automática. Por ello, algunas páginas web están protegidas frente a este tipo de software con widgets como los captcha (Lafuente, 2019).

Extraer información de las páginas web empleando crawlers tiene muchas utilidades, como por ejemplo agregar contenido, estudiar la reputación online de la empresa, seguir las tendencias o modas, obtener datos de precios, monotorizar a la competencia y optimizar el SEO.

Para emplear web scraping hay dos opciones: con o sin saber un lenguaje de programación adecuado para esta función. El lenguaje más utilizado para scrapear es Python. Este requiere de una framework o entorno de trabajo, desde la consultora Aukera recomiendan Scarpy (Lafuente, 2019). Este es un software libre y de código abierto.

Entre los crawlers ya programados para tareas de web scraping, destacan import.io y Octoparse que se comparan en la Tabla 3 y cuyos precios se recogen en las

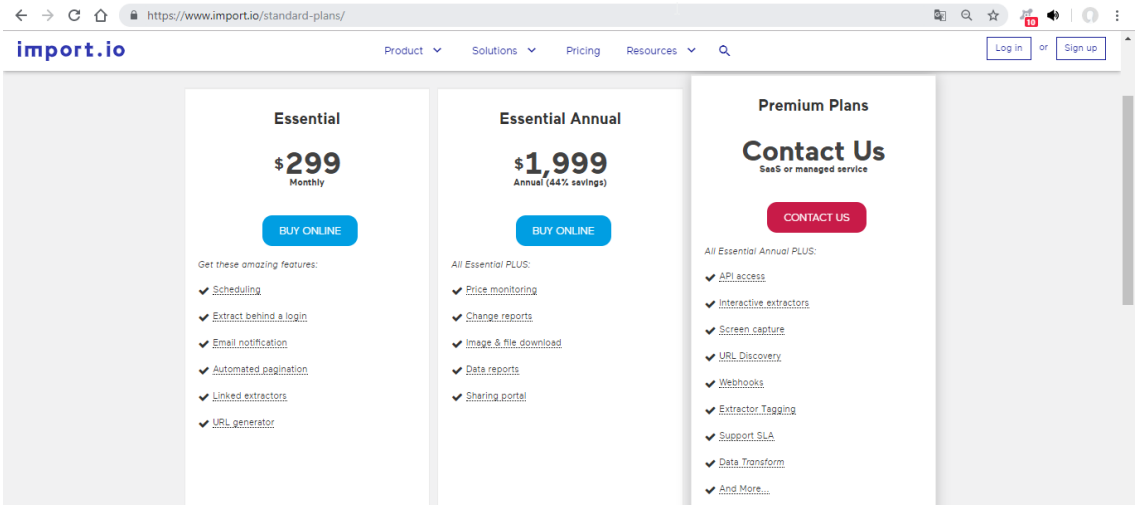
Figura 7 y Figura 8 respectivamente.

Tabla 3 Comparativa de import.io y Octoparse

	Import.io	**Octoparse**
Prueba gratuita	7 días	Versión local
Almacenamiento	Nube	Local y nube
Funciones	Seleccionar y extraer	Seleccinar, extraer, crear variables, loops y condiciones
Posibilidad de programas extracciones en la versión gratuita	Sí	No

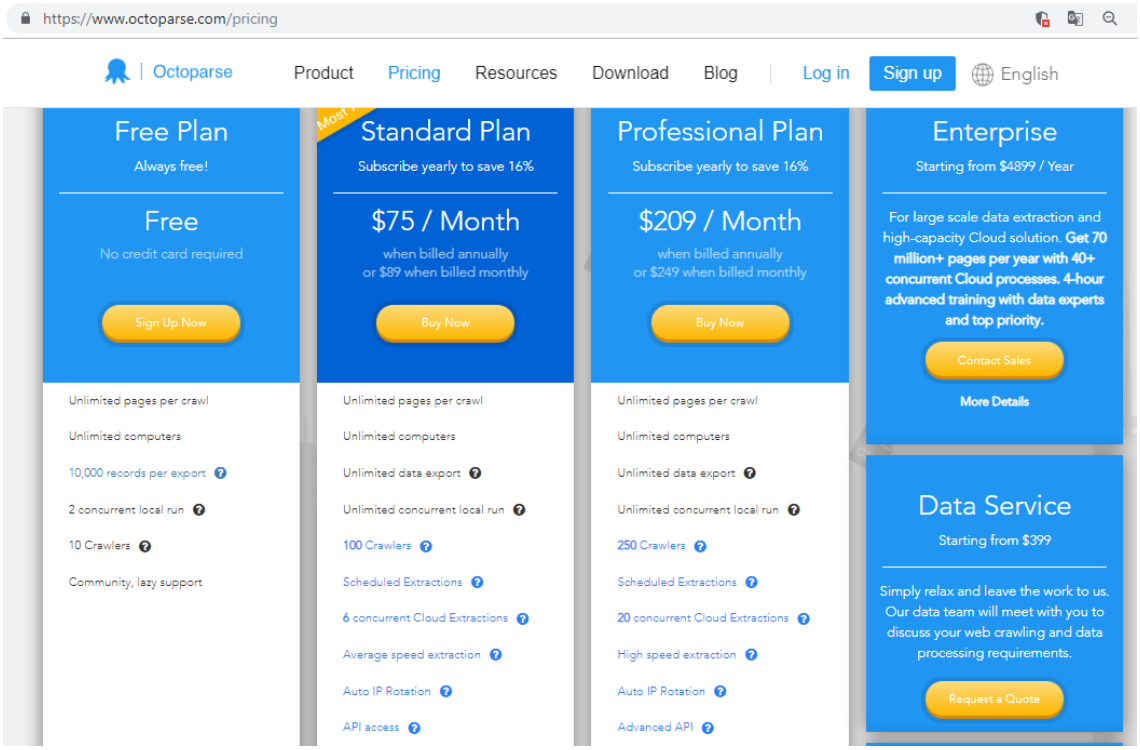
Fuente: Import.oi, 2019; Octopus Data Inc., 2019

Figura 7 Precios import.io



Fuente: Import.oi, 2019

Figura 8 Precios Octoparse



Fuente: Octopus Data Inc., 2019

	Se ha grabado un pequeño vídeo que muestra un sencillo caso de importación de datos con Octoparse. En él se extraen datos de la página web de una inmobiliaria. Se emplean el generador de url y los bucles. https://drive.google.com/open?id=1yYjTcCXQSwvmknKZ9CoXXOgW8L654z6t

7.5. Importación y transformación de los datos

La importación de los datos consiste en incluirlos en las bases en las que los van a tratar las herramientas de BI/BD. Mientras que, la transformación de los datos incluye filtrar, ordenar, agregar, combinar, limpiar, desduplicar y/o validar datos (Tejada, 2018), para que, los datos tengan las estructuras que dichas herramientas son capaces de analizar.

La importación y la transformación son dos fases del proceso que pueden intercambiar su orden. Es posible que se obtengan los datos, se importen y se transformen en el servidor de destino o que se transformen los datos en el servidor de origen y se importen ya listos para su análisis. También se puede dar un híbrido en el que se realiza parte de la transformación en cada servidor.

En ocasiones, estas tareas las hacen las mismas herramientas y en otras herramientas separar. En el caso de las aplicaciones más accesibles para las PYMES en materia de transformación: las hojas de cálculo, estas tienen la capacidad de importar los datos y crear una base de datos. Aunque sigue siendo necesario una importación anterior de los datos al almacenamiento local o en la nube que emplean estos programas.

7.5.1. **Power Pivot**

Para superar ese millón de registros, Excel cuenta con el complemento Power Pivot. En las versiones más actuales de la hoja de cálculo de Microsoft, este complemento ya viene incluido, simplemente se debe añadir a la cinta de opciones. En versiones anteriores es necesario descargarlo, esta descarga es gratuita.

	Se ha grabado un pequeño vídeo que muestra un sencillo caso de importación, transformación y relación de datos con Power Pivot. En él se emplean datos ficticios sobre los proveedores, las facturas y los pagos de una empresa. Empleando esa base de datos relacionada se crea una tabla dinámica para visualizar las facturas pendientes de pago agrupadas por proveedor. https://drive.google.com/open?id=1cfxS26Hgg5NAjz7OeHl_D_24Tj82300w

7.5.2. **Power Query**

Excel también cuenta con una herramienta para la extracción y transformación de datos (ETL) aplicable a sus hojas de cálculo y a los modelos de datos de Power Pivot, denominada Power Query. Esta misma herramienta se integra en la versión de escritorio de Power BI, también de Microsoft, que se analizará más adelante.

	Se ha grabado un pequeño vídeo que muestra un sencillo caso de importación, transformación y relación de datos con Power Query. En él se emplean datos del Índice de Precios de Vivienda extraídos de la web del Instituto Nacional de Estadística (INE). Se emplea la importación desde carpeta para conseguir que al añadir ficheros de más provincias todos los datos se junten en la misma tabla. https://drive.google.com/open?id=1tf1o_d91FW6Z1oO7gjb6M28KSqHfLDZ_

7.6. Almacenamiento

7.6.1. **Data warehouse**

El data warehouse o almacén de datos es un almacén ordenado que contiene toda la información de la empresa (Maheshwari, 2015, p.12). Bill Inmon, el padre de este término, lo define como un conjunto de datos sobre temas concretos, integrados, no volátiles y variantes en el tiempo que sirven para tomar decisiones de negocio (Inmon, 2002). Se diferencia de las bases de datos en que:

- Incluye en un solo almacén, registros de todas las áreas de la empresa.
- Contiene datos depurados, haciendo hincapié en la no duplicidad de los datos.
- Es más útil para realizar el análisis (Phillips-Wren et al., 2015, p.445). Mientras que la base de datos "simple" sirve para las gestiones del día a día (Maheshwari, 2015, p.12).
- Se centra en el largo plazo, por lo que, no se eliminarán valores (Maheshwari, 2015, p.12).

El data warehouse surge precisamente de la necesidad de (Guha et al., 2012, background) (Foster & Godbole, 2014):

- Una visión global de toda la información de la empresa independientemente de su fuente o estructura
- Sistemas de información para gestionar de forma eficiente grandes volúmenes de datos.

7.6.2. **Hadoop**

Almacenar y analizar Big Data excede las capacidades del tradicional almacén de datos. Por ello, surgieron otros sistemas que almacenan la información en diferentes clústeres o grupos de ordenadores conectados por la red (Phillips-Wren et al., 2015, p.445). Es el caso de Hadoop un sistema complementario al de sistema de datos que permite almacenar mayores volúmenes de datos mediante la utilización de MapReduce. Esta última es una herramienta ideada por Google que permite unir diversas líneas de ordenadores, de forma que, si la información no puede llegar por un "camino de ordenadores" lo haga por otro (Kaufman, 2013, p.67-68). El análisis se puede realizar en el Haddop clúster o llevar los datos al data warehouse (Phillips-Wren et al., 2015, p.445).

Hadoop ha sido desarrollado por la empresa Apache. Se trata de una aplicación de código abierto desarrollado en Java. Consta de dos partes, una de procesamiento y otra de almacenamiento (Murillo González, 2016), es decir, una de utilizar los datos y otra de guardarlos.

Sakr y otros indican que Hadoop surgió como una forma de aprovechar las fuentes y la capacidad de grandes clústeres de ordenadores en varios campos de aplicación (Sakr, Maamar, Awad, Benatallah, & Van Der Aalst, 2018, p.77313). Pero ya se han alcanzado sus limitaciones, ya que no soluciona todos los retos de BD. Por ejemplo, en cada paso intermedio entre ordenadores genera información aumentando el volumen de los datos a guardar (Sakr et al., 2018, p.77313). Por ello, han surgido otra serie de herramientas específicas para cada actividad del procesamiento de BD, como Apache Spark, Hive, GraphLab o Flink (Sakr et al., 2018, p.77313).

7.6.3. **Almacenamiento en la nube o local**

Como ya se ha comentado, el coste del almacenamiento es cada vez menor y no se requiere procesamiento local, pues se venden como servicios en la nube. Esto consiste en que las empresas ya no tienen que instalar servidores donde almacenar

las herramientas de sistemas de información y todos los datos, sino que compran ese almacenamiento. De esta manera, no solo se ahorrar los gastos en infraestructura, también el proceso de implantación es más sencillo.

El cloud computing ha reducido los costes y las barreras tecnológicas para las empresas (McKinsey Global Institute, 2011). Además, con las claves adecuadas, permite que se puedan acceder a los datos desde cualquier lugar del mundo con acceso a Internet, a través de una dirección web o conectándose a un ordenador.

La otra alternativa, la de los servidores en la empresa, se denomina local u on-premise y tiene la ventaja de mayor control de los datos.

CAPÍTULO 8. Análisis y visualización

8.1. Introducción

En este capítulo se definen dos actividades muy relacionadas: el análisis y la visualización. También se explican una serie de herramientas empleadas para llevar a cabo estas actividades.

8.2. Definiciones

8.2.1. **Análisis**

Una vez que los datos han sido obtenidos, importados y tratados, es el momento de realizar el análisis. Esta es la fase que permite transformar esta materia prima en información (Gandomi & Haider, 2015, p.4).

Para Minelli y otros (2013, p.100), el análisis de Big Data incluye análisis descriptivo (al cual restringe el Business Intelligence), el análisis predictivo y el análisis prescriptivo (Figura 9).

Figura 9 Big Data Analytics para Minelli y otros

Descriptive Analytics (Business Intelligence)	Predictive Analytics	Prescriptive Analytics
o What and when did it happen? o How much is impacted and how often does it happen? o What is the problem?	o What is likely to happen next? o What if these trends continue? o What if?	o What is the best answer? o What is the best outcome given uncertainty? o What are significantly differing and better choices?
Statistics	Data Mining Predictive Modeling Machine Learning Forecasting Simulation	Constraint-based optimization Multiobjective optimization Global optimization
Information Management		

Fuente: Minelli et al., 2013, p.100

El análisis descriptivo estudia lo que ya ha pasado, el predictivo lo que va a pasar y el prescriptivo lo que se debería cambiar para alcanzar los objetivos (IBM, s. f.).

Un tipo de análisis que se puede realizar es descriptivo y estadístico tradicional. Se trata de realizar cálculos como la media aritmética o la desviación típica, que permiten al usuario de la información conocer mejor los datos y sus tendencias. La propia hoja de cálculo, por ejemplo, Excel, cuenta con una lista de fórmulas estadísticas (suma, subtotal, media, mediana, desviación, producto, suma producto, etc.), permite aplicar filtros y elaborar diversos tipos de gráficos con estos fines. Además, las tablas dinámicas son una herramienta muy útil a la hora de analizar la información, permitiendo calcular totales, filtrar por diversos campos, cambiar los campos mostrados en filas y columnas con facilidad, etc.

Además, el análisis puede incluir la elaboración de ratios y otros indicadores característicos del área en la que se emplea la información. En el apartado 8.3.3 se definen los KPI, una herramienta habitual en este campo.

Los análisis también pueden incluir la predicción, que permite extraer patrones y extrapolarlos al futuro. En el punto 8.3.2 se tratará una herramienta de predicción conocida como data mining o minería de datos. Otras herramientas de predicción son el forecasting, la simulación y el machine learning (Minelli et al., 2013, p.10).

También se elaboran informes y cuadros de mando que permitan visualizar los datos a analizar o los resultados encontrados. A lo largo de este capítulo se detallan diversas herramientas útiles en este campo.

8.2.2. **Visualización**

Habitualmente los datos se almacenan en tablas, por lo que, es la forma más sencilla de mostrarlos. Sin embargo, no es la forma más intuitiva para emplearlos (Dutta & Bose, 2015). Por ello, existe esta tarea de visualización de los datos, que básicamente consiste en exponer los datos de forma más intuitiva, editando las tablas o en otros formatos.

Más técnicamente, la visualización se define como “a technique to facilitate the identification of patterns in data and presenting data” (Minelli et al., 2013, p.110) de forma más intuitiva que facilite su utilización, eficiencia y escalabilidad (Kaufman, 2013, p.122). También, se define como “different ways that information can be represented for the purposes of quick analysis” (Loshin, 2012, p.72) Los gráficos y cuadros de mando se emplean para resumir y mostrar de forma más explicativa la información resultante del proceso de análisis (Minelli et al., 2013, p.110).

Se puede considerar que existen dos usos de las visualizaciones:

- Cuando se utilizan los datos para realizar el análisis.
- Cuando se exponen los resultados obtenidos.

Aunque los fines y los usuarios de la información sean distintos, las visualizaciones, realizadas para estos dos usos, pueden llegar a ser muy similares.

8.3. Herramientas

Tradicionalmente, el análisis de los datos se ha realizado mediante la estadística. Se trataba de un análisis descriptivo basado en datos históricos que indicaba características de una muestra como su valor medio, las relaciones entre variables, etc. Este análisis se puede realizar con papel y calculadora o con software como la hoja de cálculo o el reconocido lenguaje y entorno de programación R (The R Foundation, s. f.). Si lo que se analiza son resultados de encuestas, también destaca SPSS.

Además, como se ha visto en la Figura 9, las herramientas para predecir que va a ocurrir con ciertos valores en el futuro se emplean el machine learning, la minería de datos, el forecasting y la simulación. Las herramientas no se emplean de forma aislada ni solo para un fin. Por ello, el machine learning se comentará en las herramientas de análisis de texto con las que está muy relacionado (apartado 8.3.1). También se analizará, en el apartado 8.3.2, el data mining.

También se calculan diversos indicadores, con las cuatro operaciones básicas. Por ejemplo, restando, el resultado del ejercicio (ingresos menos gastos). Los indicadores o medidas que permiten analizar las actividades más críticas de la empresa son los Key Performance Indicators (KPI), que se estudiarán en el apartado 8.3.3 y se compararán con otros indicadores similares.

Los datos resultantes de estos análisis se pueden resumir y mostrar en informes dinámicos, visuales y fáciles de entender, se trata de los cuadros de mando (apartado 8.3.4), estos se pueden realizar en hojas de cálculo, como en Power View con Excel o con otros programas.

De acuerdo con el cuadrante mágico de Análisis y Business Intelligence de Gartner de 2019 (Figura 10), las empresas más destacadas en esta área son Microsoft (con la herramienta Power BI) y Tableau (con la herramienta del mismo nombre). Estas dos herramientas se tratan en los apartados 8.3.5 y 8.3.6

Figura 10 Cuadrante mágico de Análisis y BI, Gartner 2019

Fuente: Gartner, 2019, p.15

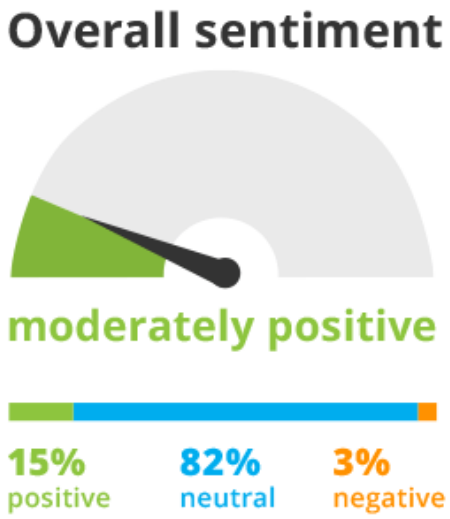
8.3.1. **Herramientas de análisis de texto**

Herramientas de análisis de texto, como computación lingüística y machine learning: permiten interpretar textos como, por ejemplo, los comentarios en una página web o red social. Incluyen herramientas de (Gandomi & Haider, 2015, p.4):

- Extracción de información, como Relation Extraction, por ejemplo, son capaces de extraer nombre de un texto y clasificarlos en categorías predefinidas como nombre, fecha o empresa.
- Respuesta a preguntas: se trata de aplicaciones como Siri de Apple, Cortana de Microsoft o el asistente virtual de la web de la Universidad de Sevilla. Estas herramientas buscan, a partir de las palabras clave que se les proporciona, los textos que versan sobre dicho tema.
- Resumen, que seleccionan de un texto las palabras clave, justo lo contrario que en el caso anterior.
- Análisis de sentimientos, tratan de determinar si un comentario es positivo o negativo, incluso se pueden añadir más matices. Hootsuite Insights (Figura 11)

o Brandwatch son algunos ejemplos de herramientas que realizan este análisis (Hootsuite, s. f.-a).

Figura 11 Hootsuite Insights



Fuente: (Hootsuite, s. f.-b)

8.3.2. **Minería de datos**

Lo que originariamente se denominaba minería de datos o data mining eran las consultas en bases de datos (Loshin, 2012, p.271). Sin embargo, el significado de este término se ha expandido incluyendo tareas más complejas.

Algunos autores definen el Data Mining como el proceso de análisis para obtener patrones tendencias y relaciones entre datos usando machine learning (Minelli et al., 2013, p.84). El machine learning es “an algorithmic technique for learning from empirical data and then using those lessons to predict future outcomes of new data” (Minelli et al., 2013, p.84). Por lo tanto, Minelli y otros reducen el data mining al emplear algoritmos para predecir el futuro, con esta técnica en que las máquinas aprenden de sus errores en las predicciones. Sin embargo, el data mining se considera algo más amplio. Y el machine learning también lo es, considerándose una herramienta al nivel del data mining, como refleja la clasificación de Minelli recogida anteriormente en la Figura 9.

De acuerdo con Loshin (2012, p.271), en la actualidad el data mining se refiere a la generación de conocimiento o knowledge discovery. También se refieren a los mismos las siglas KDD (Knowledge Discovery in Database). Se trata del proceso de búsqueda y hallazgo de patrones en grandes bases de datos empleando una o varias técnicas de data mining, como por ejemplo, el clustering o la clasificación (Loshin, 2012, p.271). El clustering consiste en dividir los datos en pequeños grupos con cierta similitud. Se diferencia de la clasificación en que las clases no están definidas de antemano. Por lo tanto, la clasificación es el reparto de los datos en grupos predefinidos (Loshin, 2012, p.276-277).

Otros autores definen el data mining como la búsqueda de patrones en grandes volúmenes de datos con técnicas procedentes de la estadística y la inteligencia artificial. Por lo tanto, coincide con Loshin en añadir más técnicas a parte del machine learning (Kaufman, 2013, 247-248), que como se ha visto se puede considerar una técnica aparte, aunque contribuya al data mining.

En resumen y de acuerdo con Kaufman (2013, p. 247-248) destaca que las tareas básicas del data mining son la clasificación de los datos y la predicción de los valores futuros.

8.3.3. **Key Performance Indicators**

Definición

Los Key Performance Indicators (KPI), también llamados Key Performance Metrics (KPMs) (Guha et al., 2012, p.16), son medidas que permiten evaluar sucesos de la empresa críticos para su actual y futuro éxito (Parmenter, 2015, p.4). Por ejemplo, el importe pendiente de pago por parte de la empresa a sus proveedores entre el total de deuda es una medida (el saldo de la cuenta de proveedores), pero el KPI sería la prueba ácida, que indica la capacidad de la empresa para solventar sus deudas.

Los KPI proporcionan a los directivos información de cómo está siendo el progreso de la empresa en el seguimiento de su estrategia.

Estos indicadores son información utilizada en los informes de BI y otras visualizaciones de los análisis realizados en el proceso de implantación de BI y BD o de camino al conocimiento. Se consideran información porque provienen del realizar cálculos con diversos datos y, además, como por ejemplo el caso de las ratios, el cociente les da un contexto que permite interpretarlos. El grado de información aumenta si se comparan con otros.

Características

De acuerdo con Parameter (2015, p.12), los KPIs tienen las siete características siguientes:

1. No son financieras, pues no están expresadas en unidades monetarias.
2. Oportunas, porque se miden con frecuencia.
3. Centrados en los directivos, debido a que ellos son los que los van a emplear.
4. Simples, porque las debería entender toda la compañía.
5. Basadas en el equipo, pues llevan al equipo responsable.
6. Gran impacto en la empresa.
7. Motivan una acción apropiada.

No se está se está de acuerdo en la primera característica, pues existen indicadores financieros que no se expresan en unidades monetarias. El plazo de pago a clientes o la prueba ácida son métricas financieras y la primera se muestra en unidades temporales y la segunda no tiene unidad de medida.

Además, que sean opotunas o no (segunda característica), depende de el uso que le de a esta herramienta la empresa.

Tampoco tiene por qué entenderlas en profundidad toda la compañía, aunque sí puede intuir por su nombre de que se trata.

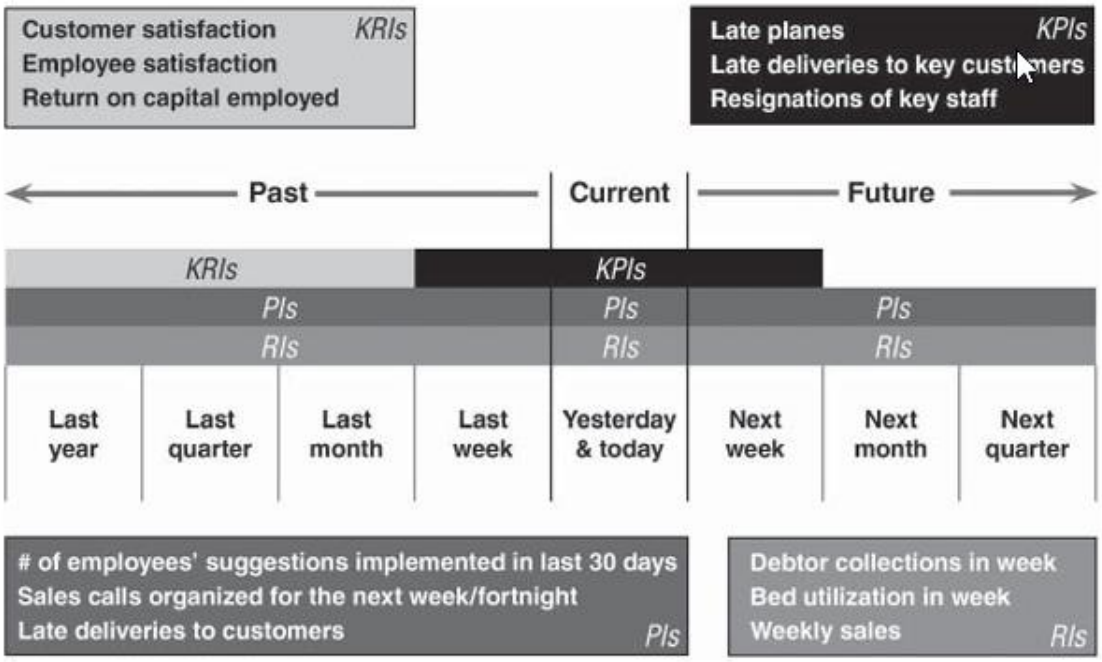
Otros indicadores

Existen otras medidas complementarias a los KPI que se deben diferenciar de estos:

- Key Results Indicators (KRI). Proporcionan a los directivos un resumen de cómo está siendo el rendimiento de la empresa. Muestran los resultados de acciones de trabajo llevadas a cabo por diferentes equipos de trabajo. Permiten comparar los resultados con el objetivo de la estrategia (Parmenter, 2015, p.4), es decir, permiten ver lo que la empresa ha logrado (Badawy, El-Aziz, Idress, Hefny, & Hossam, 2016, p.1). Son KRI el beneficio neto después de impuestos, el beneficio neto en las líneas de producto claves, satisfacción del cliente, beneficios sobre el los recursos empleados y satisfacción de los empleados (Parmenter, 2015, p.4).
- Result indicators (RIs). Muestran cómo se combinan las labores de los equipos para mostrar resultados (Parmenter, 2015, p.5). Con respecto a los KRI, los RI no resumen tanto la información y son más detallados. Son las ventas de ayer, las iniciativas llevadas a cabo a partir de la última encuesta a los clientes, las sugerencias implementadas provenientes de clientes, etc. (Parmenter, 2015, p.6). Parmenter incluye en estos tipos de indicadores todos los financieros, refiriéndose a todos los que tienen unidades monetarias.
- Performance indicators (PIs). Parmenter (2015, p.7) los define como indicadores no financieros que permiten derivar la responsabilidad al equipo que realiza las acciones. Si fuesen financieros, serían RIs. Son PIs el número de llamadas realizadas al call center que se cortan en el tiempo de espera, los pedidos que llegaron tarde, las innovaciones aplicadas por cada equipo de trabajo, etc. (Parmenter, 2015, p.7).

En la Figura 12 se detallan las diferencias a nivel temporal entre estos tres indicadores y los KPIs y se muestra un ejemplo de cada uno.

Figura 12 Diferencias entre KRIs, Pis, RIs y KPIs



Fuente: Parmenter, 2015, p.19

8.3.4. **Cuadros de mando**

Los cuadros de mando, dashboards o scoreboards son herramientas de gestión que recogen un conjunto coherente de indicadores para proporcionar a los encargados de tomar decisiones una visión comprensible del negocio que les facilite la labor (Asociación Española de Normalización y Certificación, 2003, p.5).

El cuadro de mando integral es una evolución del dashboard ideada por Robert Kaplan y David Norton. Este se centra los indicadores claves del negocio muy relacionados con la estrategia. El cuadro de mando integral tiene en cuenta las cuatro perspectivas clave de una organización (Blanco Martinez, 2012, p.464):

- Económico-financiera
- Del cliente
- De los procesos internos
- De los empleados

Crear cuadros de mando también es posible con gestores de hojas de cálculo. Para ello se emplean, tablas dinámicas, gráficos dinámicos y segmentaciones de datos conectadas a estos. También existe un complemento de Excel, denominado Power View, con mayor variedad de representación gráfica de los datos.

8.3.5. **Power BI**

Power BI, de Microsoft, permite importar datos y crear visuales Cuadros de Mando. De acuerdo con la propia empresa creadora del software, Power BI destaca por mostrar los datos de forma impactante visualmente, que permite analizar los datos en modo local o en la nube, personalizar informes interactivos, compartirlos y gestionar quién puede verlos (Microsoft, s.f.).
Además, Power BI se puede conectar con otros servicios empleados por la empresa como MailChimp y Google Analitics. También cabe remarcar, que no es necesario aprender un lenguaje de programación, es sencillo, intuitivo y permite realizar directamente preguntas, como si hablásemos con Cortana o Siri.

Se puede registrar una cuenta gratuita con un correo empresarial de 12 meses para una serie de servicios de Microsoft del paquete de herramientas para empresa Azure.

	Se ha grabado un pequeño vídeo que muestra un sencillo caso de análisis y visualización de datos con Power BI. En él se emplean los índices de precios reales en la agricultora en los diversos países de la Unión Europea. Estos datos han sido extraídos del Instituto Europeo de Estadística (Eurostat). Se emplean, entre otros, gráficos de tipo mapa. https://drive.google.com/open?id=1IMpGda2VbC5UeQzwcNT0Fw8yj6DIIqIt

8.3.6. **Tableau**

La herramienta Tableau sirve para realizar el análisis y la visualización de los datos y, también, para exportarlos, transformarlos y cargarlos. Por lo que, podría formar parte, tanto de este apartado como del de ETL. Se ha incluido aquí porque la propia empresa en su página web destaca de la herramienta el análisis (Tableau, s.f.) y las posibilidades de visualización; y es por ello, por lo que destaca en el mercado.
Tableau es una herramienta de pago (70 dólares por usuario y por mes) que también ofrece 14 días de muestra y licencias para estudiantes (Tableau, s. f.).
Al descargar Tableau se nos ofrecen dos softwares: Tableau Pre Builder y Tableau Desktop. El primero permite la edición de los datos, mientras que, el segundo se centra en su análisis.

CAPÍTULO 9. Conclusiones

En este trabajo se ha cumplido el objetivo primario de estudiar la implantación y utilización de herramientas de Big Data y Business Intelligence en las PYMES con el fin de que obtengan ventajas competitivas. Se ha han definido ambos términos, concluyendo que se refieren a lo mismo, aunque haciendo énfasis en procesos diferentes. También, se ha establecido un proceso de implantación y se ha contribuido a la fase de investigación informado sobre estos procesos y sus demás fases.

Este trabajo se ha realizado centrándose en aquellas herramientas que son accesibles para las PYMES, de forma que, sus limitaciones de recursos no les impidan gestionar sus datos y alcanzar ventajas competitivas a través del BI y BD. Las herramientas analizadas han sido Power Pivot, Power Query, Power BI, Tableau, Import.oi y Octoparse.

Además, para mostrar de manera práctica algunas de las principales herramientas analizadas, se ha optado por la realización de vídeos recogiendo pequeños casos prácticos. En ellos se emplean alguna de las herramientas para exportar, transformar o analizar los datos.

También, se han incluido definiciones y ejemplos de formatos de archivos en un pequeño blog que sirve de complemento a la revisión bibliográfica.

Todo ello, ha permitido mostrar aplicaciones y técnicas, útiles para la pequeña y mediana empresa, que se pueden emplear sin habilidades ni conocimientos informáticos muy específicos, como lenguajes de programación, para tratar los datos y obtener información que mejore las decisiones y pueda conducir a obtener ventajas competitivas.

Este trabajo se ha visto limitado en su parte práctica, a la información disponible en la web y a la dificultad de encontrar en ella un caso realista en el que emplear diversas herramientas. Además, de las típicas limitaciones de los trabajos académicos como el límite de páginas y los plazos de entrega. Por ello, se consideran las siguientes líneas de mejora:

- Aplicación de lo aprendido en una empresa real con un reto de Sistemas de Información real.
- Ampliación del estudio en el debate sobre si las máquinas pueden generar o no conocimiento (machine learning).
- Estudio de las API (Interfaces de programación de aplicaciones) en la obtención de datos.
- Obtención, tratamiento y utilización de datos cualitativos para la toma de decisiones.
- Obtención de datos de correo electrónico y redes sociales y emplearlos en el día a día de la empresa.

Bibliografía

Anuradha J; Ishwarappa. (2015). *A Brief Introduction on Big data 5Vs Characteristics and Hadoop Technology*. *48*, 319-324. https://doi.org/10.1016/j.procs.2015.04.188

Asociación Española de Normalización y Certificación. *UNE 66175:2003 Sistemas de gestión de la calidad. Guía para la implantación de sistemas de indicadores*. , (2003).

Badawy, M., El-Aziz, A. A. A., Idress, A. M., Hefny, H., & Hossam, S. (2016). A survey on exploring key performance indicators. *Future Computing and Informatics Journal*, *1*(1-2), 47-52. https://doi.org/10.1016/j.fcij.2016.04.001

Blanco Martinez, E. (2012). Cuadro De Mando Integral. *Debates Iesa*. Recuperado de https://ebookcentral--proquest--com.us.debiblio.com/lib/uses/detail.action?docID=3228588#

Briones Delgado, J. M. (2014). *Datos, información y conocimiento: promesas y realidades de la red global* (Universidad Complutense de Madrid). Recuperado de http://eprints.ucm.es/27622/7/T35541.pdf

De Pablos Heredero, C., López Hermoso Agius, J. J., Martín-Romo Romero, S., & Medina Salgado, S. (2012). *Organización y transformación de los sistemas de información de la empresa*. Madrid: ESIC Editorial.

Dutta, D., & Bose, I. (2015). Managing a big data project: The case of Ramco cements limited. *International Journal of Production Economics*, *165*, 293-306. https://doi.org/10.1016/j.ijpe.2014.12.032

Espinoza, M., & Secaira, J. (2016). Gestión del conocimiento para el desarrollo de organizaciones inteligentes. *FFL Roca - Revista*, *3*(9), 660-673. Recuperado de https://www.rmlconsultores.com/revista/index.php/crv/article/view/393

Foster, E. C., & Godbole, S. V. (2014). *Database Systems. A Pragmatic Approach*. https://doi.org/https://doi.org/10.1007/978-1-4842-0877-9

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gartner. (s. f.). Big Data. Recuperado de IT Glossary website: https://www.gartner.com/it-glossary/big-data/

Gartner. (2019). *2019 Analytics and BI Magic Quadrant*. Recuperado de http://public2.brighttalk.com/resource/core/228334/apr29rsallamakronzjrichardson_499972.pdf

Guha, A., Wrabetz, J., Wu, S., & Madireddi, V. (2012). *Method and system for business intelligence analytics on unstructured data*. Recuperado de https://patents.google.com/patent/US8266148B2/en

Hootsuite. (s. f.-a). Herramientas de análisis de sentimiento. Recuperado de https://blog.hootsuite.com/es/analisis-de-sentimiento/

Hootsuite. (s. f.-b). Hootsuite Insights. Mejor escucha social. Decisiones empresariales más inteligentes. Recuperado de https://hootsuite.com/es/productos/insights#

IBM. (s. f.). IBM Business Analytics. Recuperado de https://www.ibm.com/analytics/mx/es/business-intelligence/pdf/ibm-business-analytics-soluciones-destacadas-mx.pdf

Import.oi. (2019). Import.io. Recuperado de https://www.import.io/

Inmon, W. H. (2002). *Building the Data Warehouse* (John Wiley & Sons, Ed.). Recuperado de https://books.google.es/books?id=9T6Oe6AujzUC&dq=Building the data warehouse&hl=es&source=gbs_book_other_versions

Kaufman, M. (2013). *Big Data for Dummies.pdf*. Recuperado de https://ebookcentral--proquest--com.us.debiblio.com/lib/uses/detail.action?docID=1160914

Lafuente, A. (2019). Qué es el web scraping. *Aukera web*. Recuperado de https://aukera.es/blog/web-scraping/

Loshin, D. (2012). *Business Intelligence: The Savvy Manager's Guide*. https://doi.org/ISBN : 0-12-385890-9

Maheshwari, A. (2015). *Business intelligence and data mining* (1º; New York New York 222 East 46th Street, Ed.). Recuperado de https://ebookcentral--proquest--com.us.debiblio.com/lib/uses/reader.action?docID=1911815

Martínez López, F. J. (2019). *La era mundial- internacional*. Sevilla: Universidad de Sevilla.

McKinsey Global Institute. (2011). *Big data : The next frontier for innovation , competition , and productivity*. Recuperado de https://www.mckinsey.com/~/media/McKinsey/Business Functions/McKinsey Digital/Our Insights/Big data The next frontier for innovation/MGI_big_data_exec_summary.ashx

Microsoft. (s. f.). Power BI. Recuperado de https://powerbi.microsoft.com/es-es/

Minelli, M., Chambers, M., & Dhiraj, A. (2013). *Big Data, Big Analytics : Emerging Business Intelligence and Analytic Trends for Today's Businesses* (John Wiley & Sons Inc., Ed.). Recuperado de https://ebookcentral--proquest--com.us.debiblio.com/lib/uses/detail.action?docID=818100#

Mochón Morcillo, F., & Gonzálvez Cabañas, J. C. (2016). *Big Data : una gestión inteligente de los datos*. Madrid: García Maroto Editores.

Murillo González, M. (2016). *Sistema Big Data para el análisis de rutas de taxis en NYC*. Recuperado de https://e-archivo.uc3m.es/bitstream/handle/10016/24115/TFG_Montserrat_Murillo_Gonzalez_2016.pdf

Octopus Data Inc. (2019). Octoparse. Recuperado de https://www.octoparse.com/

Ontsi red.es. (2018). *Individuos que han usado internet en los últimos 3 meses por dispositivos utilizados para conectarse a internet Indicators Individuos*. Recuperado de https://www.ontsi.red.es/ontsi/es/indicador/individuos-que-han-usado-internet-en-los-últimos-3-meses-por-dispositivos-utilizados-para-conectarse-a-internet

Parmenter, D. (2015). Key performance indicators : developing, implementing, and using winning KPIs. *Performance Management in Healthcare*. https://doi.org/10.4324/9781315102214-5

Phillips-Wren, G., Iyer, L. S., Kulkarni, U., & Ariyachandra, T. (2015). Business analytics in the context of big data: A roadmap for research. *Communications of the Association for Information Systems*, *34*, 448-472. https://doi.org/10.17705/1CAIS.03723

Rojas Pescio, H. G. (2016). El rol de las empresas basadas en conocimiento (EBC) y las empresas basadas en tecnología (EBT) para la innovación The. *Tecnología: Resultados de Investigación*, *Edición Nº*, 65-80. Recuperado de https://dialnet.unirioja.es/descarga/articulo/5771043.pdf

Ruiz del Castillo, J. C. (2019). *Big Data y Business Intelligence*.

Sakr, S., Maamar, Z., Awad, A., Benatallah, B., & Van Der Aalst, W. M. P. (2018). Business process analytics and big data systems: A roadmap to bridge the gap. *IEEE Access*, *6*, 77308-77320. https://doi.org/10.1109/ACCESS.2018.2881759

Sánchez Sánchez, D., Pan Bermúdez, A., & Viña Castiñeiras, A. (2004). *Eii, Un Nuevo Paradigma Para La Integración De Información Dispersa Y Heterogénea En La Administración. Experiencia:Vixía, Un Servicio De Vigilancia Tecnológica*. 1-14.

Suárez Sánchez, A. (2017). Sistemas para la organización del conocimiento: definición y evolución histórica. *e-Ciencias de la Información*, *7*(2), 1. https://doi.org/10.15517/eci.v7i2.26878

Tableau. (s. f.). Tableau. Recuperado de https://www.tableau.com/es-es

Tejada, Z. (2018). Extracción, transformación y carga de datos (ETL). *Data Architecture Guide de Microsoft Azure*. Recuperado de https://docs.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl

The R Foundation. (s. f.). What is R? Recuperado de https://www.r-project.org/about.html

Anexos

Vídeos complementarios

https://drive.google.com/drive/folders/1OOMT4Y2lc9WRsLxmpuyXdc_Msbevcjhp?usp=sharing

Definiciones blog

A continuación, se recoge el código QR que dirige al blog y los enlaces de los apartados del blog empresait vinculados a lo largo del trabajo.

Access <https://empresait.wordpress.com/2019/05/26/tips-basicos-de-access/>

CSV <https://empresait.wordpress.com/2019/07/02/csv/>

SQL <https://empresait.wordpress.com/2019/05/24/structured-query-language/>

SWIFT <https://empresait.wordpress.com/2019/07/02/swift/>

XML <https://empresait.wordpress.com/2019/07/02/xml/>