

# Módulo 1: Fundamentos de Big Data

<b>MÓDULO 1: FUNDAMENTOS DE BIG DATA.....</b>	<b>1</b>
<b>INTRODUCCIÓN .....</b>	<b>5</b>
PÓSTER DEL MAPA MENTAL .....	6
PÓSTER DE RELACIONES.....	7
<b>ENTENDER QUÉ ES BIG DATA .....</b>	<b>8</b>
INTRODUCCIÓN A BIG DATA.....	8
LECTURAS OPCIONALES .....	10
<b>TERMINOLOGÍA Y CONCEPTOS FUNDAMENTALES .....</b>	<b>14</b>
DATASETS .....	15
ANÁLISIS DE DATOS (DATA ANALYSIS).....	16
ANALÍTICA.....	16
INTELIGENCIA DE NEGOCIOS (BI) .....	17
INDICADORES CLAVE DE DESEMPEÑO (KPI).....	17
UNIDADES DE TAMAÑO DE LOS DATOS.....	18
EJERCICIO 1.1: COMPLETE LOS ESPACIOS EN BLANCO.....	20
<b>FACTORES EMPRESARIALES Y TECNOLÓGICOS DE BIG DATA.....</b>	<b>24</b>
ANALÍTICA Y CIENCIA DE DATOS .....	25
DIGITALIZACIÓN.....	25
TECNOLOGÍA ASEQUIBLE Y HARDWARE BÁSICO .....	26
SOCIAL MEDIA .....	27
COMUNIDADES Y DISPOSITIVOS HIPERCONECTADOS.....	27
CLOUD COMPUTING.....	27
<b>TECNOLOGÍAS EMPRESARIALES TRADICIONALES RELACIONADAS CON BIG DATA....</b>	<b>32</b>
PROCESAMIENTO DE TRANSACCIONES EN LÍNEA (OLTP).....	33
PROCESAMIENTO ANALÍTICO EN LÍNEA (OLAP) .....	34
SISTEMAS DE OLTP Y OLAP .....	34
EXTRAER – TRANSFORMAR – CARGAR (ETL).....	36
BODEGAS DE DATOS DIGITALES (DATA WAREHOUSE).....	36
DATA MARTS .....	37
HADOOP .....	39
LECTURAS OPCIONALES .....	39

<b>CARACTERÍSTICAS DE LOS DATOS EN LOS ENTORNOS BIG DATA .....</b>	<b>43</b>
VOLUMEN .....	44
VELOCIDAD .....	45
VARIEDAD .....	46
VERACIDAD .....	47
VALOR .....	47
<b>TIPOS DE DATOS EN LOS ENTORNOS BIG DATA .....</b>	<b>52</b>
DATOS ESTRUCTURADOS .....	53
DATOS SIN ESTRUCTURAR .....	53
DATOS SEMIESTRUCTURADOS .....	54
METADATA .....	55
TIPOS DE DATOS Y VERACIDAD .....	55
EJERCICIO 1.2: COMPLETE LOS ESPACIOS EN BLANCO .....	57
<b>ANÁLISIS FUNDAMENTAL, ANALÍTICA Y TIPOS DE APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING) .....</b>	<b>61</b>
TIPOS DE ANÁLISIS DE DATOS (DATA ANALYSIS) .....	62
ANÁLISIS CUANTITATIVO .....	62
ANÁLISIS CUALITATIVO .....	63
MINERÍA DE DATOS (DATA MINING) .....	63
ANÁLISIS Y ANALÍTICA .....	63
TIPOS DE ANALÍTICA .....	64
ANALÍTICA DESCRIPTIVA .....	66
ANALÍTICA DIAGNÓSTICA .....	66
ANALÍTICA PREDICTIVA .....	67
ANALÍTICA PRESCRIPTIVA .....	68
APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING) .....	69
TIPOS DE APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING) .....	70
COMPARACIÓN ENTRE EL APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING) Y LA MINERÍA DE DATOS (DATA MINING) .....	70
EJERCICIO 1.3: COMPLETE LOS ESPACIOS EN BLANCO .....	72
<b>INTELIGENCIA DE NEGOCIOS (BI) Y BIG DATA .....</b>	<b>76</b>
LA INTELIGENCIA DE NEGOCIOS (BI) TRADICIONAL .....	77
INTELIGENCIA DE NEGOCIOS (BI) TRADICIONAL: REPORTES ESPECIALIZADOS .....	77
INTELIGENCIA DE NEGOCIOS (BI) TRADICIONAL: TABLEROS DE CONTROL (DASHBOARDS) .....	79

INTELIGENCIA DE NEGOCIOS (BI) DE BIG DATA .....	80
<b>VISUALIZACIÓN DE DATOS Y BIG DATA .....</b>	<b>85</b>
VISUALIZACIÓN DE DATOS .....	86
HERRAMIENTAS DE VISUALIZACIÓN DE DATOS .....	86
CARACTERÍSTICAS DE VISUALIZACIÓN DE DATOS .....	86
HERRAMIENTAS AVANZADAS DE VISUALIZACIÓN .....	87
EJERCICIO 1.4: COMPLETE LOS ESPACIOS EN BLANCO.....	87
<b>ELEMENTOS A TENER EN CUENTA AL PLANEAR Y ADOPTAR BIG DATA.....</b>	<b>92</b>
JUSTIFICACIÓN EMPRESARIAL .....	93
PRERREQUISITOS ORGANIZACIONALES.....	93
APROVISIONAMIENTO DE DATOS .....	93
PRIVACIDAD .....	95
SEGURIDAD .....	95
PROCEDENCIA .....	96
SOPORTE LIMITADO EN TIEMPO REAL .....	97
DIFERENTES PROBLEMAS DE RENDIMIENTO .....	97
DIFERENTES REQUISITOS DE GESTIÓN .....	98
METODOLOGÍA DIFERENCIAL .....	99
CLOUD COMPUTING.....	99
LECTURAS OPCIONALES .....	100
<b>RESPUESTAS A LOS EJERCICIOS .....</b>	<b>104</b>
EJERCICIO 1.1: RESPUESTAS .....	105
EJERCICIO 1.2: RESPUESTAS .....	105
EJERCICIO 1.3: RESPUESTAS .....	105
EJERCICIO 1.4: RESPUESTAS .....	106
<b>EXAMEN B90.01 .....</b>	<b>107</b>
<b>KIT DE AUTOAPRENDIZAJE DEL MÓDULO 1.....</b>	<b>107</b>
<b>INFORMACIÓN Y RECURSOS DE CONTACTO .....</b>	<b>108</b>
COMUNIDAD DE AITCP.....	108
INFORMACIÓN GENERAL DEL PROGRAMA .....	108
INFORMACIÓN GENERAL ACERCA DE LOS MÓDULOS DEL CURSO Y LOS KITS DE AUTOAPRENDIZAJE ...	108
INQUIETUDES ACERCA DEL EXAMEN DE PEARSON VUE .....	108
PROGRAMACIÓN DE TALLERES DIRIGIDOS AL PÚBLICO Y GUIADOS POR INSTRUCTORES .....	108

TALLERES PRIVADOS GUIADOS POR INSTRUCTORES .....	109
CONVERTIRSE EN UN ENTRENADOR CERTIFICADO .....	109
INQUIETUDES GENERALES SOBRE BDSCP .....	109
NOTIFICACIONES AUTOMÁTICAS.....	109
RETROALIMENTACIÓN Y COMENTARIOS .....	109

# Introducción

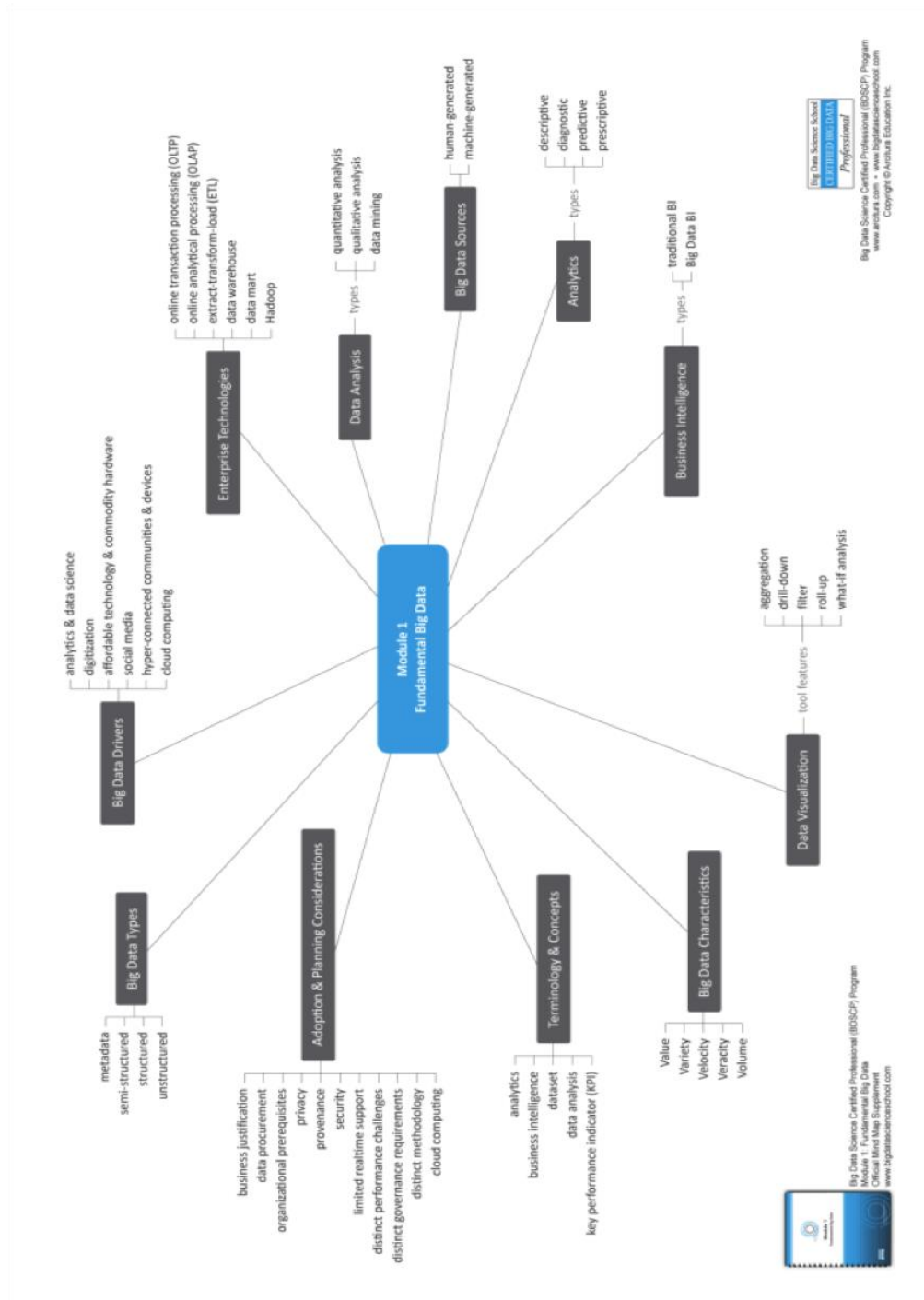
Este es el cuadernillo oficial del **Módulo 1: Fundamentos de Big Data** para el curso del BDSCP y su respectivo **Examen B90.01** de Pearson VUE.

El objetivo de este documento es brindar conocimientos sobre conceptos fundamentales relacionados con Big Data, los cuales incluyen, entre otros:

- Entender qué es Big Data
- Conceptos y terminología fundamentales relacionados con Big Data
- Factores empresariales y tecnológicos de Big Data
- Tecnologías empresariales tradicionales relacionadas con Big Data
- Características de los datos en los entornos Big Data
- Tipos de datos en los entornos Big Data
- Análisis fundamental, analítica y tipos de aprendizaje automático (Machine Learning)
- Inteligencia de negocios (BI) y Big Data
- Visualización de datos y Big Data
- Elementos a tener en cuenta al planear y adoptar Big Data

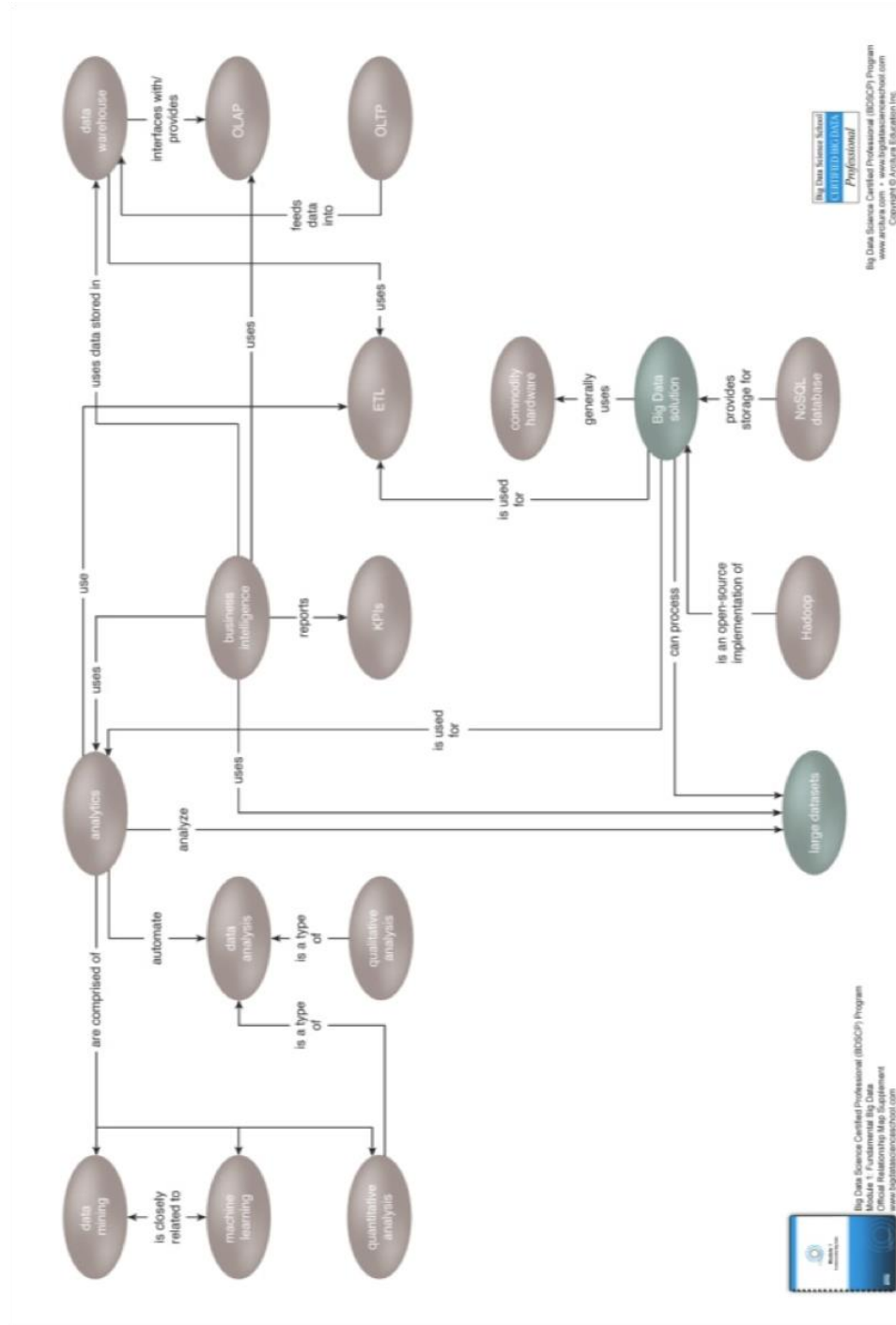
## Póster del mapa mental

El *Póster del mapa mental del Módulo 11 del BDSCP* incluido en este cuadernillo del curso ofrece una representación visual alternativa de todos los principales temas abordados en este curso.



## Póster de relaciones

El *Póster de relaciones del Módulo 1 del BDSCP* incluido en este cuadernillo del curso ofrece una representación visual alternativa de los principales temas abordados en este curso.



# Entender qué es Big Data

## Introducción a Big Data

Big Data es un campo orientado al análisis, procesamiento y almacenamiento de grandes colecciones de datos que, con frecuencia, provienen de distintas fuentes. Por lo general, se requieren soluciones y prácticas de Big Data cuando la tecnología tradicional de análisis, procesamiento y almacenamiento de datos no es suficiente. Particularmente, Big Data aborda distintos requisitos, como la combinación de múltiples datasets no relacionados, el procesamiento de grandes cantidades de datos sin estructurar y la recopilación de información oculta con plazos de tiempo definidos.

Las cualidades que diferencian los datos procesados por medio de soluciones de Big Data son conocidas comúnmente como las “Cinco V”, y serán presentadas en la sección *Características de los datos en los entornos Big Data*, más adelante. Por medio de las soluciones de Big Data se pueden realizar tareas complejas de análisis, con el fin de obtener resultados de análisis profundamente significativos e intuitivos para el beneficio de la empresa. Las soluciones de Big Data pueden procesar grandes cantidades de datos que son recibidos a distintas velocidades, son ampliamente variados y tienen numerosas incompatibilidades.

En los entornos Big Data, los datos son acumulados **al interior de la empresa** por medio de aplicaciones o a partir de fuentes externas, y posteriormente son almacenados en la solución de Big Data. Los datos procesados por una solución de Big Data pueden ser utilizados directamente por las aplicaciones empresariales, o pueden ser enviados a una bodega de datos digital (Data Warehouse), a fin de enriquecer los datos ya existentes. Estos datos **generalmente son analizados y sometidos a analítica**.

Por lo general, los datos procesados, así como los resultados de los análisis, **son utilizados para tareas significativas y complejas de reporte y evaluación, y además son retroalimentados en las aplicaciones** para mejorar el rendimiento de las mismas; por ejemplo, cuando se muestran recomendaciones de productos online. Los datos procesados por las soluciones de Big Data pueden ser generados por máquinas o por humanos, aunque finalmente la generación de los resultados del procesamiento es responsabilidad de las computadoras. Los **datos generados por humanos** son el resultado de la interacción entre las personas y los sistemas; por ejemplo, servicios online y dispositivos digitales. La Figura 1.1 ilustra ejemplos de datos generados por humanos, los cuales pueden ser datos estructurados, video y datos de texto.



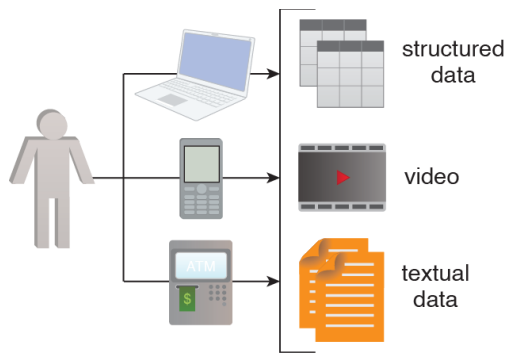


Figura 1.1 – Los ejemplos de datos generados por humanos incluyen social media, microblogging, correos electrónicos, fotos compartidas y mensajería.

Los **datos generados por máquinas** son el resultado de la generación de datos automatizada y determinada por eventos, ejecutada por programas de software o dispositivos de hardware. La Figura 1.2 ofrece una representación visual de ejemplos de datos generados por máquinas de servidores web, medidores inteligentes y dispositivos GPS.

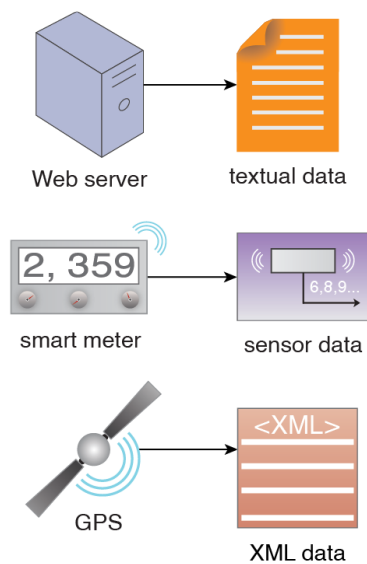


Figura 1.2 – Los ejemplos de datos generados por máquinas incluyen logs web, datos de sensores, datos de telemetría, datos de medidores inteligentes y datos de uso de dispositivos.

Los resultados del procesamiento de la solución de Big Data pueden generar una gran variedad de conocimientos y beneficios, por ejemplo:

- optimización operativa
- inteligencia accionable
- identificación de nuevos mercados
- predicciones precisas
- detección de errores y fraudes

- registros más detallados
- mejor toma de decisiones
- descubrimientos importantes

Existen muchas preocupaciones, limitaciones y consideraciones a la hora de adoptar una solución de Big Data, y todas deben ser comprendidas y sopesadas frente a los correspondientes beneficios esperados. En la sección *Elementos a tener en cuenta al planear y adoptar Big Data*, se analizan muchos de dichos elementos.

### **Lecturas opcionales**

El libro *Analítica de Big Data*, incluido en el *Módulo 2: Conceptos de análisis y tecnología de Big Data*, presenta ejemplos de casos de estudio, desde puntos de vista del mundo real, en la sección *Demasiado grande para ignorarlo: el caso empresarial de Big Data*, del capítulo 5.

[illegible]

[illegible]

[illegible]

## Notas / Bocetos

# Terminología y conceptos fundamentales

Como preparación para las secciones posteriores que abordan el siguiente conjunto de temas introductorios, las páginas a continuación presentan definiciones concisas de estos términos básicos:

- Datasets
- Análisis de datos (Data Analysis)
- Analítica
- Inteligencia de negocios (BI)
- Indicadores Clave de Desempeño (KPI)

La sección finaliza con los términos y las abreviaturas que son usados como parte de la terminología relacionada con el tamaño de los datos.

## Datasets

En los módulos de este curso, a los conjuntos o grupos de datos relacionados se les conoce comúnmente como **datasets**. Cada grupo o miembro de un dataset (**dato**) comparte los mismos atributos con otros dentro de un dataset.

La Figura 1.3 muestra tres datasets basados en tres formatos distintos. Entre los ejemplos están:

- tuits almacenados en un archivo plano
- una colección de archivos de imágenes
- un extracto de filas almacenadas en una tabla
- observaciones climáticas históricas almacenadas como archivos XML

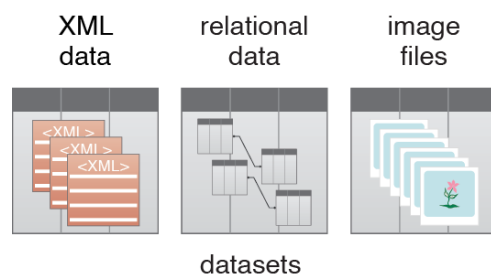


Figura 1.3 – Los datasets pueden estar basados en datos XML, datos relacionales y/o archivos de imágenes.

## Análisis de datos (Data Analysis)

El análisis de datos (Data Analysis) es el proceso de **examinación de los datos** con el fin de hallar hechos, relaciones, patrones, explicaciones y/o tendencias. El objetivo final del análisis de datos (Data Analysis) es **respaldar la toma de decisiones**. Un ejemplo sencillo es el análisis de los datos (Data Analysis) de las ventas de helados, con el objetivo de determinar cómo la cantidad de conos de helado vendidos se encuentra relacionada con la temperatura diaria. Esto sirve para respaldar las decisiones que tienen que ver con cuánto helado y cuántos conos puede pedir y surtir una tienda en relación con la información del pronóstico del clima. La realización de análisis de datos (Data Analysis) permite **establecer patrones y relaciones entre los datos analizados**.



Figura 1.4 – Símbolo utilizado para representar el análisis de datos (Data Analysis).

## Analítica

La analítica es la disciplina encargada comprender los datos, **analizándolos mediante una variedad de técnicas científicas y herramientas automatizadas, enfocada en el descubrimiento de patrones y correlaciones ocultos**. En los entornos Big Data, la analítica por lo general es aplicada usando tecnologías y frameworks distribuidos y altamente escalables para analizar grandes volúmenes de datos provenientes de distintas fuentes.



Figura 1.5 – Símbolo utilizado para representar la analítica.

Por lo general, el proceso de analítica implica filtrar grandes cantidades de datos sin procesar ni estructurar, con el fin de extraer información significativa que pueda servir como datos de entrada para identificar patrones, enriquecer los datos empresariales actuales o realizar búsquedas a gran escala.

Distintas organizaciones utilizan técnicas y herramientas de analítica en formas diferentes; por ejemplo, estos tres sectores:

- En los **entornos orientados a los negocios**, los resultados de la analítica pueden disminuir los costos operativos y facilitar la toma estratégica de decisiones.



- En el **ámbito científico**, la analítica puede ayudar a identificar la causa de un fenómeno y mejorar la precisión de las predicciones.
- En los **entornos basados en servicios** —como en las organizaciones del sector público—, la analítica puede ayudar a mejorar el enfoque orientado en la prestación de servicios de alta calidad, disminuyendo los costos.

En general, la analítica facilita la toma de decisiones determinadas por datos, con un respaldo científico, de manera que estas decisiones puedan estar basadas en datos concretos y no solamente en la experiencia o la intuición.

## Inteligencia de negocios (BI)

La Inteligencia de negocios (BI) es el **proceso de comprender el funcionamiento de una empresa —para mejorar la toma de decisiones— al analizar los datos externos y los datos generados por sus procesos empresariales**. En la Inteligencia de negocios (BI), la analítica es aplicada a grandes cantidades de datos en toda la empresa. Además, la Inteligencia de negocios (BI) puede utilizar los datos consolidados que se encuentran almacenados en la bodega de datos digital (Data Warehouse) para ejecutar consultas analíticas. Como se muestra en la Figura 1.6, la Inteligencia de negocios (BI) puede ser usada por medio de un mecanismo de tablero de control (Dashboard) para analizar y acceder a las consultas sobre estos datos.

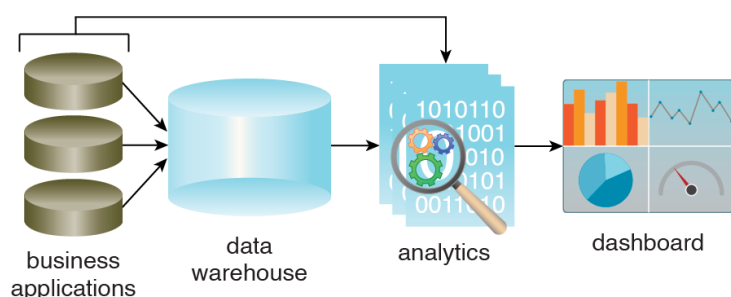


Figura 1.6 – La Inteligencia de negocios (BI) puede ser utilizada para mejorar las aplicaciones empresariales, consolidar los datos en las bodegas de datos digitales (Data Warehouse) y analizar las consultas por medio de un tablero de control (Dashboard).

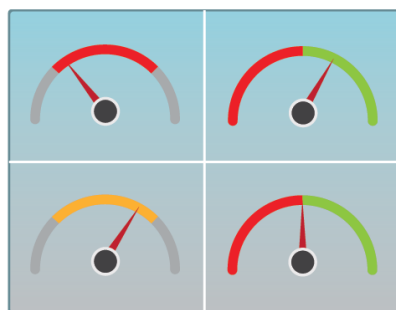
## Indicadores Clave de Desempeño (KPI)

Un indicador clave de desempeño (**KPI**, por sus siglas en inglés) es una forma de medir el éxito dentro de un contexto particular. Los KPI están estrechamente relacionados con los objetivos estratégicos de una empresa y generalmente son utilizados para:

- **identificar áreas problemáticas, con el fin de adoptar medidas correctivas**
- **lograr el cumplimiento normativo**

Los KPI sirven como puntos de referencia rápida para medir el desempeño general de la empresa por medio de los tableros de control (Dashboard) de KPI, ilustrados en la Figura 1.7.

Cada KPI está basado en un indicador cuantificable que es identificado y acordado de antemano.



KPI dashboard

Figura 1.7 – Un tablero de control (Dashboard) de KPI puede medir las llamadas atendidas por día y la cantidad de unidades generadas al mes.

## Unidades de tamaño de los datos

Cuando se analizan los distintos tamaños de los datos, es necesario comprenderlos en unidades de cuantificación de datos. Los siguientes tamaños de datos están enumerados en la Tabla 1.1, donde el byte es utilizado como la unidad fundamental de medida con prefijos decimales, no binarios.

Data Size	Number of Bytes
Kilobyte (KB)	1,000
Megabyte (MB)	1,000,000
Gigabyte (GB)	1,000,000,000
Terabyte (TB)	1,000,000,000,000
Petabyte (PB)	1,000,000,000,000,000
Exabyte (EB)	1,000,000,000,000,000,000
Zettabyte (ZB)	1,000,000,000,000,000,000,000
Yottabyte (YB)	1,000,000,000,000,000,000,000,000

Tabla 1.1 – Unidades de tamaño de los datos

#### NOTA

Los principales tipos de análisis de datos (Data Analysis) pertinentes a Big Data son estudiados en la sección *Análisis fundamental, analítica y tipos de aprendizaje automático (Machine Learning)*. La inteligencia de negocios (BI) y los KPI son analizados con más detalle en la sección *Inteligencia de negocios (BI) y Big Data*.

### **Ejercicio 1.1: complete los espacios en blanco**

1. Un \_\_\_\_\_ es un conjunto de datos relacionados, en el cual todos los miembros del grupo poseen el mismo conjunto de atributos.
2. El objetivo \_\_\_\_\_ es respaldar la toma de decisiones al establecer patrones y relaciones en los datos que son analizados.
3. La \_\_\_\_\_ se enfoca en filtrar grandes cantidades de datos sin \_\_\_\_\_, con el fin de extraer información significativa que pueda ser útil para enriquecer los datos empresariales actuales.
4. El proceso de la Inteligencia de negocios (BI) puede aplicar la \_\_\_\_\_ a grandes cantidades de datos.

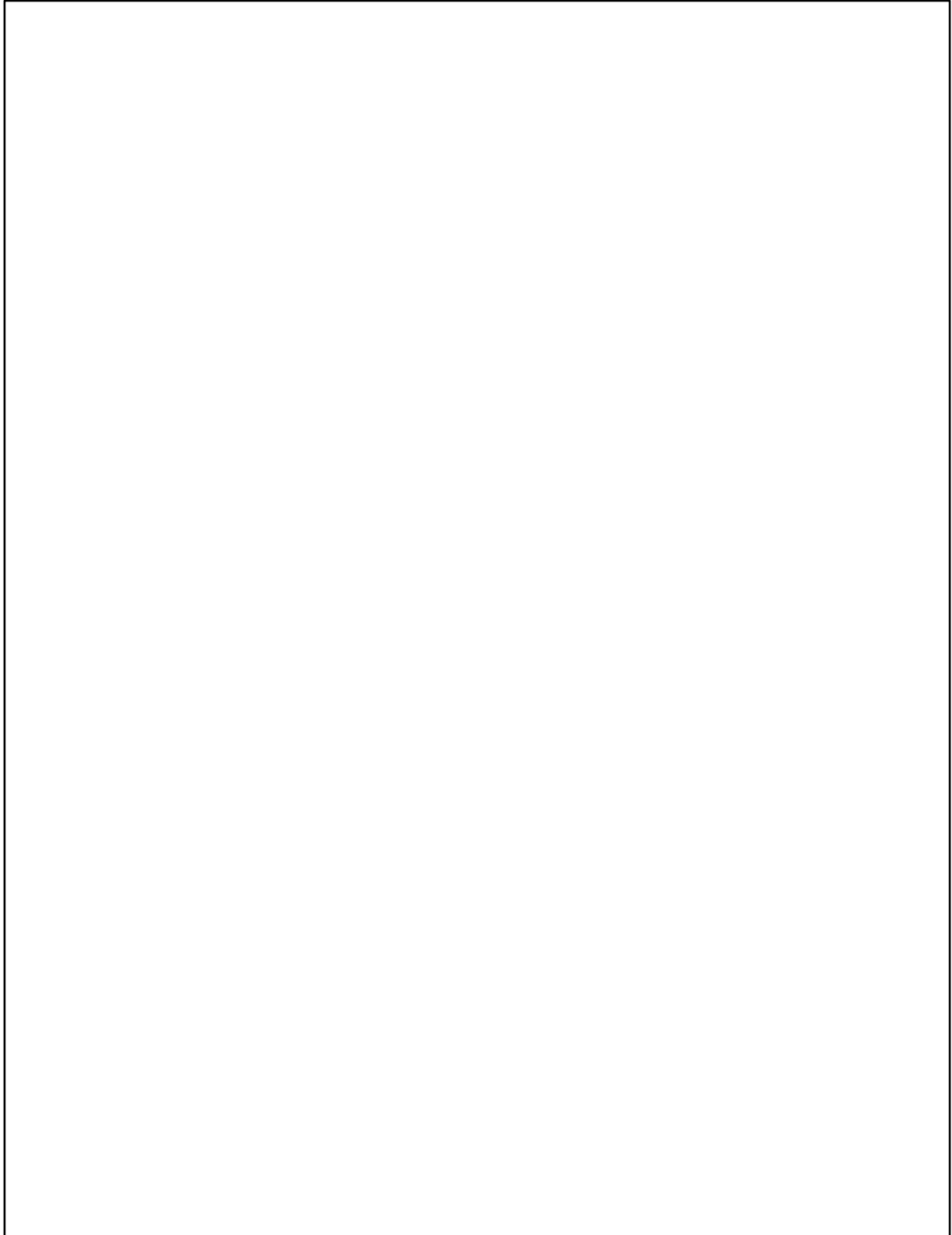
*Las respuestas al ejercicio se encuentran al final de este cuadernillo.*

[illegible]

[illegible]

[illegible]

## Notas / Bocetos





# Factores empresariales y tecnológicos de Big Data

Big Data surge como una combinación de las necesidades empresariales y las innovaciones tecnológicas. Esta sección examinará los principales factores empresariales y tecnológicos que permitieron que Big Data se convirtiera en una especialidad en sí:

- Analítica y ciencia de datos
- Digitalización
- Tecnología asequible y hardware básico
- Social media
- Comunidades y dispositivos hiperconectados
- Cloud Computing

## Analítica y ciencia de datos

A medida que las empresas en crecimiento recopilan y almacenan cada vez más datos, buscando posibles nuevos conocimientos y ventajas competitivas, aumenta la necesidad de contar con técnicas y tecnologías que permitan la extracción de información y conocimientos significativos. Los algoritmos de aprendizaje automático (Machine Learning), las técnicas estadísticas y el bodegaje de datos digitales (Data Warehouse) han permitido que la ciencia de datos y la analítica avancen hasta tal punto que han emergido como disciplinas en sí, con técnicas y herramientas especializadas para realizar análisis complejos y únicos. **La madurez de estos campos prácticos inspiró y posibilitó gran parte de la funcionalidad esencial que se espera de las soluciones y herramientas de Big Data hoy en día.**

## Digitalización

Para muchas empresas, los medios digitales han reemplazado los medios físicos como las comunicaciones y el mecanismo estándar de entrega. Los datos digitalizados brindan la **oportunidad de recopilar datos “secundarios” adicionales**; por ejemplo, cuando las personas realizan búsquedas o completan encuestas. La recopilación de datos secundarios puede ser importante para las empresas, ya que la extracción de este tipo de datos posibilita el mercadeo personalizado, las recomendaciones automatizadas y el desarrollo de características optimizadas de productos. La Figura 1.8 ofrece una representación visual de algunos ejemplos de digitalización.

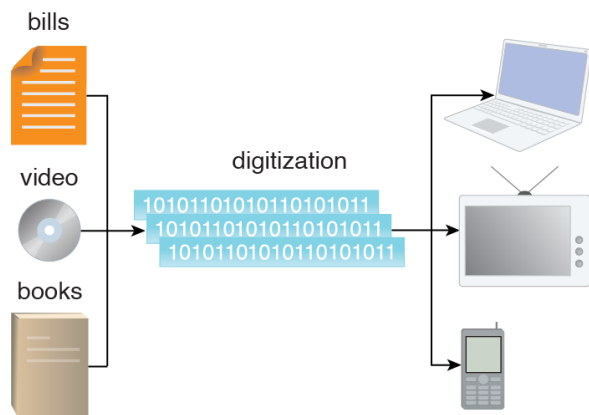


Figura 1.8 – Entre los ejemplos de digitalización están la banca online, la televisión por demanda y el video por streaming.

### Tecnología asequible y hardware básico

La tecnología relacionada con la recopilación y el procesamiento de grandes cantidades de diversos datos es cada vez más asequible. Las soluciones típicas de Big Data están basadas en **software de código abierto que requiere hardware básico**.

El uso de este tipo de hardware permite que la adopción de soluciones de Big Data sea asequible para aquellas empresas que carecen de grandes inversiones de capital. La Figura 1.9 ilustra un ejemplo de los ahorros relacionados con los precios de almacenamiento de datos.

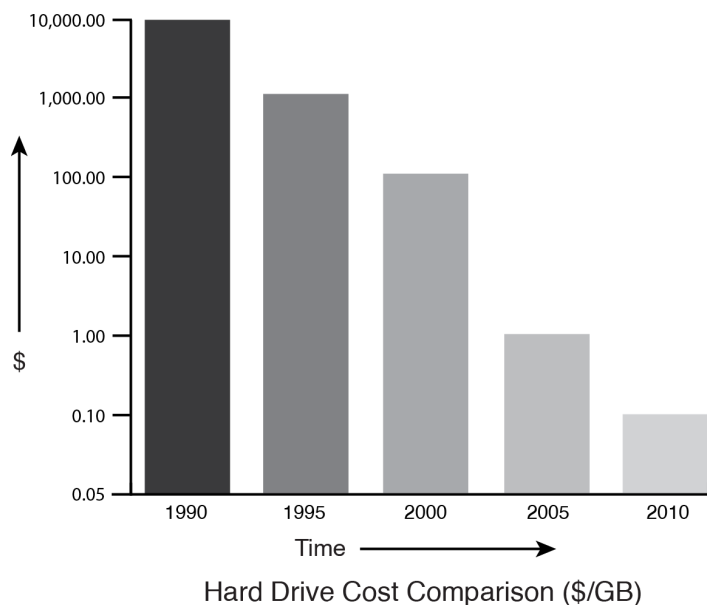


Figura 1.9 – A través de las décadas, el precio del almacenamiento de datos ha disminuido significativamente, de USD 10.000 a menos de USD 0,10 por GB.

## Social media

El surgimiento de social media ha permitido que los clientes **suministren retroalimentación en tiempo (prácticamente) real a través de medios públicos y privados**, un cambio que ha obligado a las empresas a tener en cuenta en su planeación estratégica la retroalimentación que los clientes hacen de sus ofertas. Como resultado, **las empresas almacenan cada vez más datos sobre las interacciones de los clientes y a través de social media** en un intento de recopilar los datos para aumentar las ventas, posibilitar un mercadeo dirigido y crear nuevos productos y servicios. Asimismo, las empresas están cada vez más interesadas en incorporar datasets disponibles al público provenientes de social media y otras fuentes externas de datos.

## Comunidades y dispositivos hiperconectados

El alcance cada vez mayor de la internet y la proliferación de redes de telefonía celular y wifi ha permitido que cada vez más personas estén activas de forma constante en las comunidades virtuales, ya sea directamente por medio de la interacción online, o indirectamente a través del uso de dispositivos conectados. Esto ha traído como resultado **flujos masivos de datos**. Algunos flujos de datos son públicos, mientras que otros flujos están dirigidos directamente a los proveedores y a las empresas. En la Figura 1.10 se ilustran los distintos tipos de comunidades y dispositivos hiperconectados.

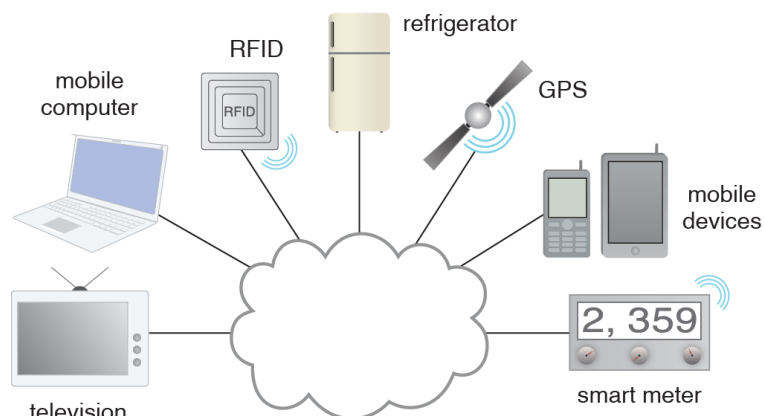


Figura 1.10 – Las comunidades y dispositivos hiperconectados incluyen la televisión, la informática móvil, RFID, refrigeradores, dispositivos GPS, dispositivos móviles y medidores inteligentes.

## Cloud Computing

Los avances en la tecnología de Cloud Computing han llevado a la creación de entornos remotos, a los que se les conoce como “nubes.” Estos entornos proporcionan **alta escalabilidad y recursos de TI por demanda que pueden ser arrendados bajo los modelos de “pago por uso”**. Las empresas tienen la oportunidad de mejorar la infraestructura y la capacidad de almacenamiento y procesamiento que proporcionan estos entornos, con el fin de crear soluciones de Big Data de gran escala que pueden ejecutar grandes tareas de procesamiento.

El ejemplo de la Figura 1.11 ilustra cómo se pueden mejorar las capacidades de escalabilidad de un entorno de nube para realizar tareas de procesamiento de Big Data. El hecho de que los recursos de TI basados en la nube puedan ser alquilados reduce de forma considerable la inversión inicial de los proyectos de Big Data.

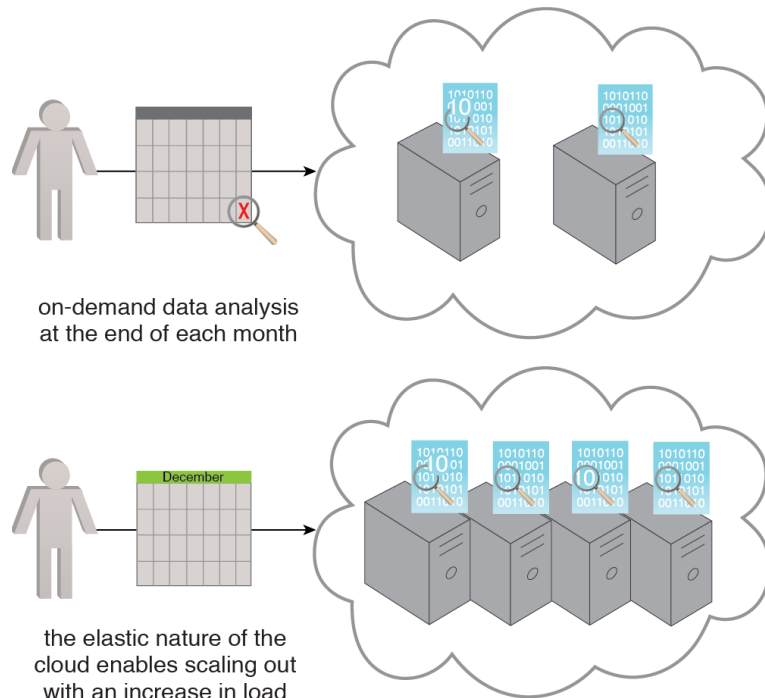


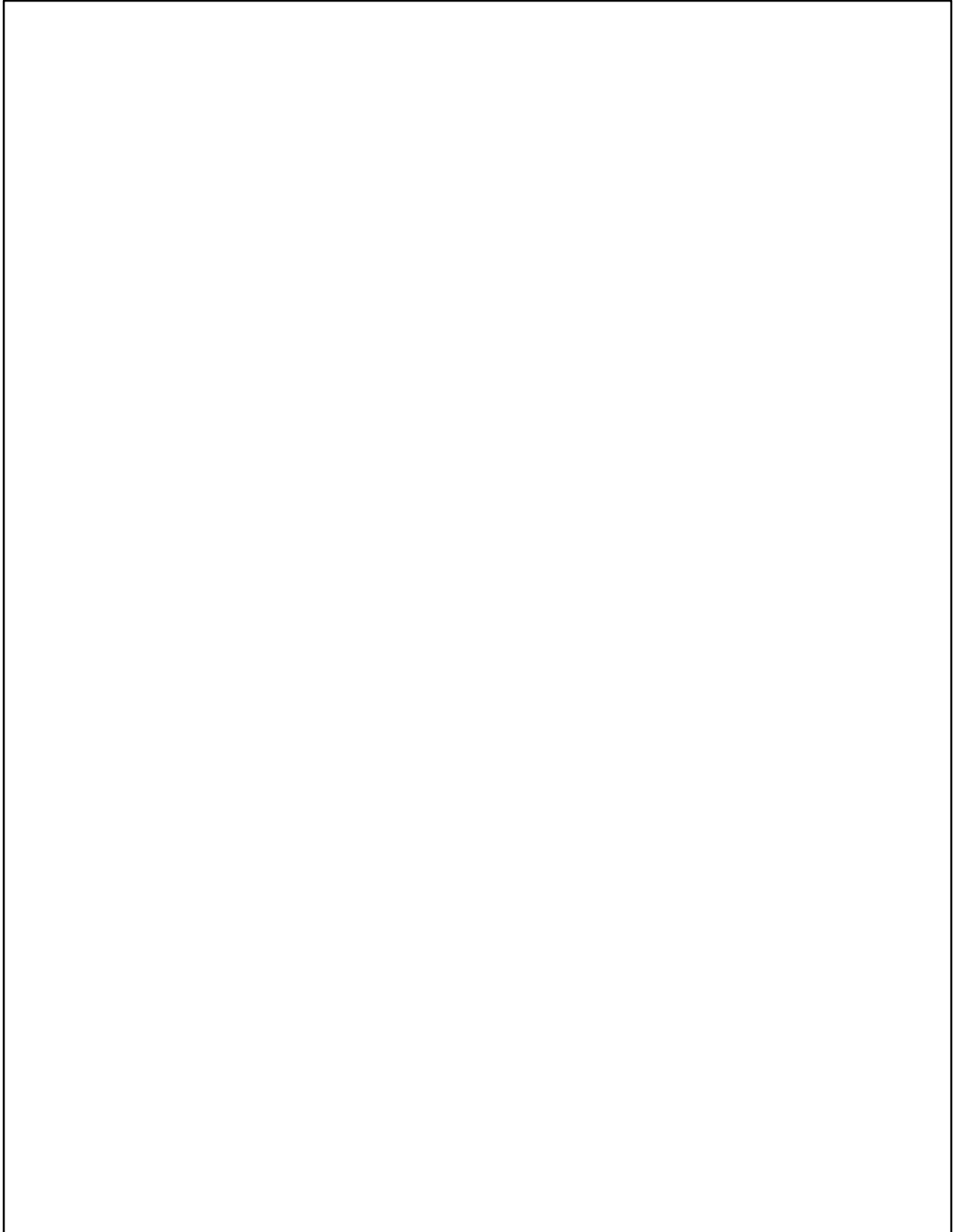
Figura 1.11 – La nube puede ser utilizada, por ejemplo, para completar un análisis de datos (Data Analysis) por demanda a fin de mes, o mejorar la escalabilidad de los sistemas, aumentando la carga.

[illegible]

[illegible]

[illegible]

## Notas / Bocetos





# Tecnologías empresariales tradicionales relacionadas con Big Data

En esta sección se describen brevemente las siguientes tecnologías:

- Procesamiento de Transacciones en Línea (OLTP)
- Procesamiento Analítico en Línea (OLAP)
- Extraer - Transformar - Cargar (ETL)
- Bodegas de datos digitales (Data Warehouse)
- Data marts
- Hadoop

La mayor parte de estas tecnologías se encuentran bien establecidas en la industria de las TI y son anteriores al surgimiento de Big Data. Dichas tecnologías son analizadas en esta sección ya que **cada una de ellas es relevante de forma única para las soluciones y ecosistemas actuales de Big Data.**

## NOTA

Si usted ya se encuentra familiarizado con las tecnologías enumeradas, puede omitir esta sección.

## Procesamiento de Transacciones en Línea (OLTP)

El **Procesamiento de Transacciones en Línea (OLTP)** es un sistema de software que procesa los datos orientados a las transacciones. El término “transacción online” se refiere a la finalización de una actividad en tiempo real y no mediante el procesamiento por lotes (Batch Processing). Los sistemas de OLTP almacenan datos operacionales que están completamente normalizados, y que en el ámbito de Big Data, son importantes para representar una **fuentes común de datos analíticos de entrada**. Los resultados de los análisis de Big Data también pueden ser retroalimentados en los sistemas de OLTP.

Los sistemas de OLTP son compatibles con consultas compuestas por operaciones simples de inserción, eliminación y actualización, con tiempos de respuesta menores a un segundo. Algunos ejemplos incluyen los sistemas de reservas de boletos y las transacciones bancarias y en puntos de venta (POS, por sus siglas en inglés).

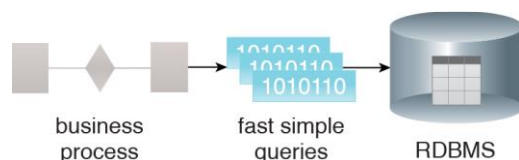


Figura 1.12 – Símbolo utilizado para representar el OLTP.

## Procesamiento Analítico en Línea (OLAP)

El **Procesamiento Analítico en Línea (OLAP)** es un sistema utilizado para el procesamiento de consultas de análisis de datos (Data Analysis). El OLAP es una parte esencial de los procesos de Inteligencia de negocios (BI), minería de datos (Data Mining) y aprendizaje automático (Machine Learning). Estos procesos son relevantes para Big Data ya que actúan como una fuente de datos y un sink con la capacidad de recibir datos. Son utilizados en las analíticas **diagnóstica**, **predictiva** y **prescriptiva**, las cuales son presentadas más adelante en este curso.

Los sistemas de OLAP almacenan datos históricos que son agregados y desnormalizados para ofrecer una capacidad rápida de reportes. Asimismo, estos sistemas utilizan las bases de datos que almacenan los datos históricos en arrays multidimensionales, y pueden solucionar consultas complejas con base en múltiples dimensiones de los datos.

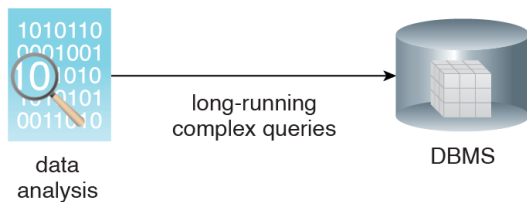


Figura 1.13 – Símbolo utilizado para representar el OLAP.

## Sistemas de OLTP y OLAP

Un sistema de OLAP siempre es alimentado con datos de múltiples sistemas de OLTP por medio de trabajos regulares de procesamiento por lotes (Batch Processing). A diferencia de los sistemas de OLTP, el tiempo de respuesta de las consultas de OLAP puede ser de varios minutos o más, dependiendo de la complejidad de la consulta y de la cantidad de registros solicitados.

En la Figura 1.14, los datos relacionales de dos sistemas OLTP son importados a un sistema OLAP cada cierto tiempo por medio de tareas de importación masiva de datos. Los datos relacionales son almacenados en el sistema OLAP como datos desnormalizados en forma de cubos. Esto permite que los datos sean consultados durante cualquier tarea de análisis de datos (Data Analysis) realizada posteriormente.

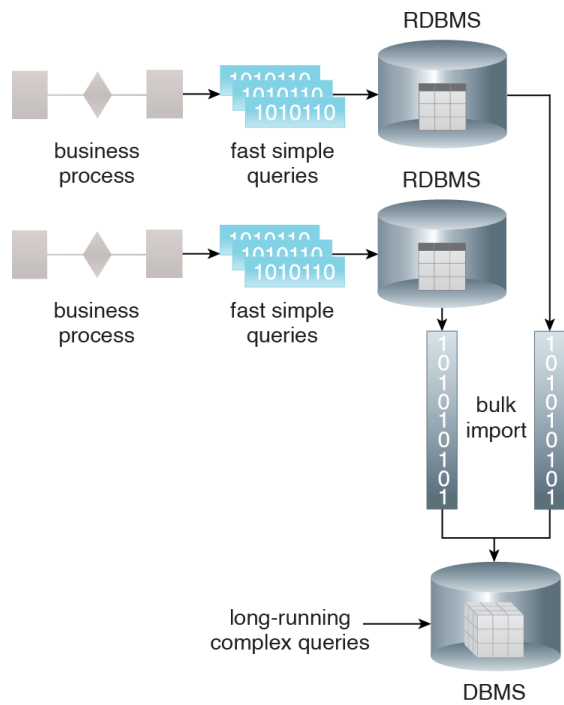


Figura 1.14 – El sistema OLAP almacena datos relacionales que son desnormalizados para futuros análisis de datos (Data Analysis).

## Extraer – transformar – cargar (ETL)

**Extraer – transformar – cargar (ETL)** es un proceso mediante el cual los datos son cargados desde un sistema origen hacia un sistema destino. El sistema origen puede ser una base de datos, un archivo plano o una aplicación. De igual forma, el sistema destino puede ser una base de datos o cualquier otro sistema de información.

ETL representa la principal operación por medio de la cual las bodegas de datos digitales (Data Warehouse) reciben datos. Una solución de Big Data abarca el conjunto de características de ETL con el fin de convertir datos de distintos tipos. La Figura 1.15 muestra que primero los datos requeridos son obtenidos o **extraídos** del origen, luego son modificados o **transformados** mediante la aplicación de reglas. Por último, los datos son insertados o **cargados** al sistema destino.

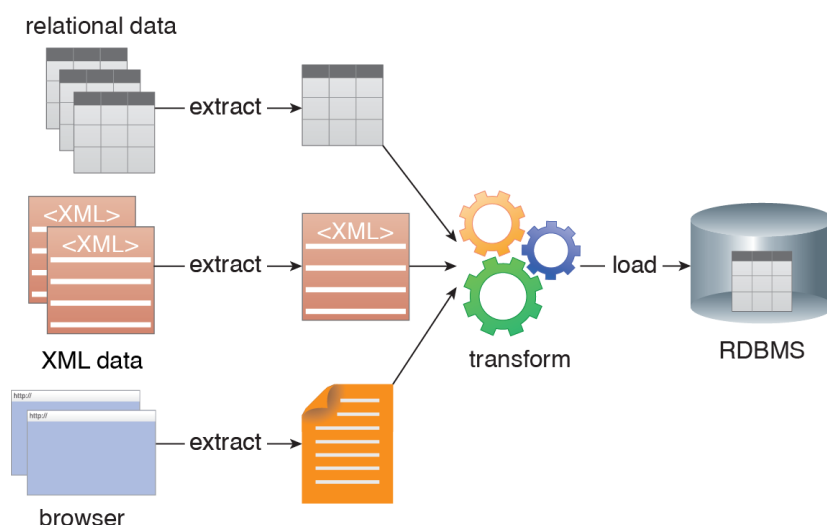


Figura 1.15 – El proceso ETL puede extraer datos de navegador, XML y relacionales.

## Bodegas de datos digitales (Data Warehouse)

Una **bodega de datos digital (Data Warehouse)** es un **repositorio central a nivel empresarial** que contiene datos históricos y actuales. Las bodegas de datos digitales (Data Warehouse) son usadas considerablemente por la Inteligencia de negocios (BI) para realizar distintas consultas analíticas, y por lo general tienen interfaces con el sistema de OLAP para tener compatibilidad de consulta analítica, como se muestra en la Figura 1.16.

Los datos relacionados con distintas entidades empresariales y que provienen de diferentes sistemas operacionales son extraídos, validados, transformados y consolidados periódicamente en una sola base de datos. Gracias a las importaciones periódicas de datos provenientes de toda la empresa, la cantidad de datos contenidos en una sola bodega de datos digital (Data Warehouse) seguirá aumentando. Como resultado, los tiempos de respuesta de las tareas de consulta de análisis de datos (Data Analysis) ejecutadas como parte de la inteligencia de negocios (BI) pueden verse afectados.

A fin de solucionar esta deficiencia, por lo general, las bodegas de datos digitales (Data Warehouse) contienen bases de datos optimizadas, llamadas bases de datos analíticas, para gestionar las tareas de reporte y análisis de datos (Data Analysis). Una base de datos analítica puede existir como una RDBMS, como en el caso de una base de datos de OLAP.

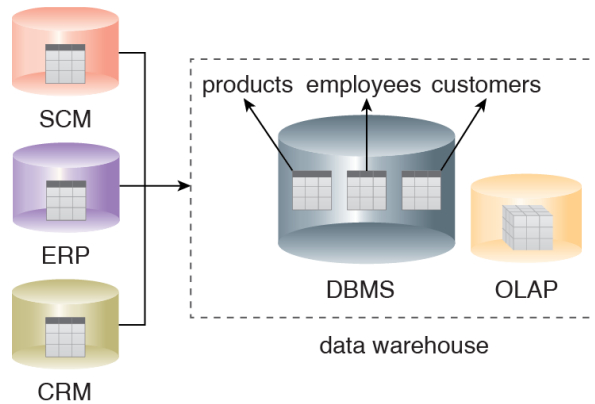


Figura 1.16 – Una bodega de datos digital (Data Warehouse) extrae periódicamente datos de otras fuentes, como sistemas de OLTP, ERP, CRM y SCM, para consolidarlos en un dataset.

## Data marts

Un **data mart** es un subconjunto de datos almacenados en una bodega de datos digital (Data Warehouse) que, por lo general, pertenece a un departamento, división o línea de negocio específica. Las bodegas de datos digitales (Data Warehouse) pueden tener múltiples data marts. Como se muestra en la Figura 1.17, se recopilan y posteriormente se extraen los datos provenientes de toda la empresa y de las entidades empresariales. Las entidades particulares de un dominio son guardadas en la bodega de datos digital (Data Warehouse) mediante un proceso de ETL.

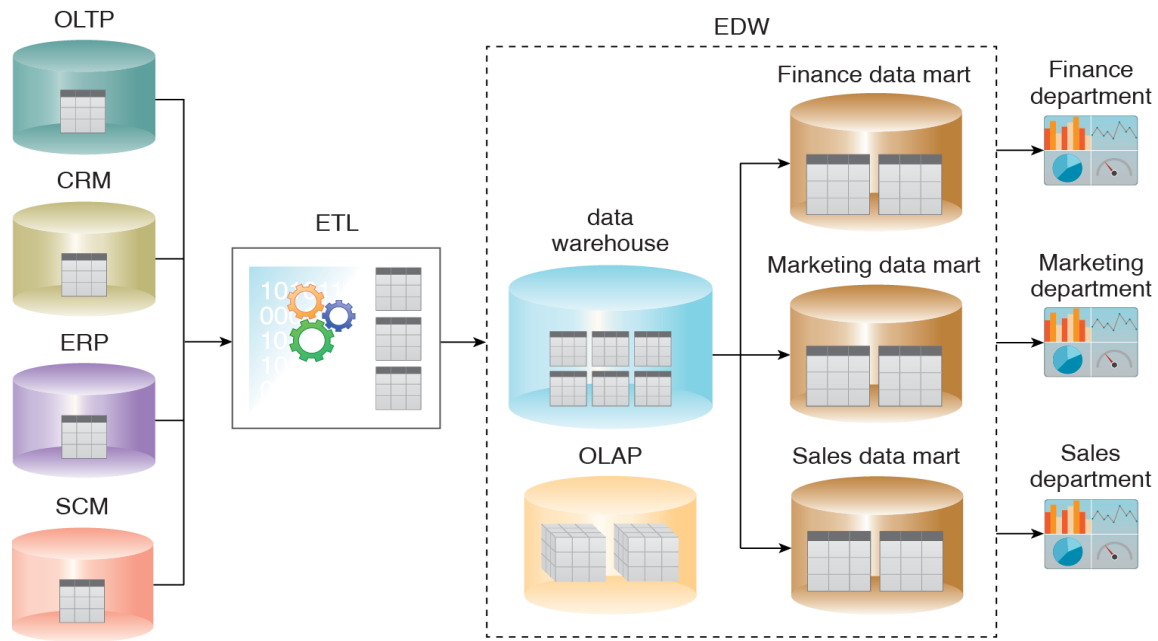


Figura 1.17 – La única base de datos centralizada de una bodega de datos digital (Data Warehouse) está basada en datos limpios (Cleansing), lo cual es un prerequisite para los reportes precisos y libres de error, de acuerdo con el resultado que se muestra a la derecha.

## Hadoop

Hadoop es un **framework de código abierto** para el almacenamiento y procesamiento de datos a gran escala que técnicamente es ejecutado en hardware básico. El framework de Hadoop **se ha establecido como la plataforma predeterminada de la industria para las soluciones modernas de Big Data**. Puede ser utilizado como un motor de ETL o analítico para procesar grandes cantidades de datos estructurados, semiestructurados y sin estructurar. La Figura 1.18 ilustra algunas de las características de Hadoop.

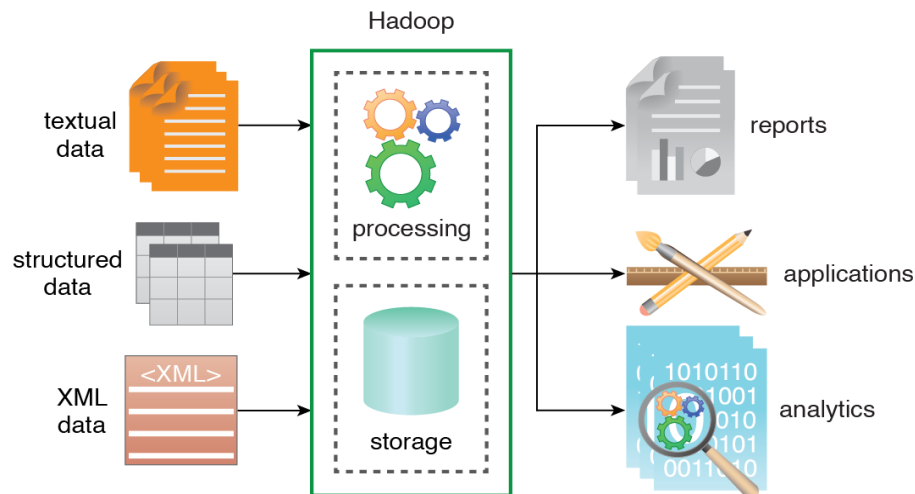


Figura 1.18 – Símbolos utilizados para representar a Hadoop y sus características.

### NOTA

El *Módulo 2: Conceptos de análisis y tecnología de Big Data* presenta más información relacionada con las características de Hadoop y los mecanismos relacionados.

## Lecturas opcionales

El libro *Analítica de Big Data*, incluido con el *Módulo 2: Conceptos de análisis y tecnología de Big Data*, examina la relación entre Big Data y las bodegas de datos digitales (Data Warehouses) en la sección *Analítica de Big Data* del *Capítulo 3: El surgimiento de las opciones de Big Data*.

[illegible]



[illegible]

[illegible]

## Notas / Bocetos



# Características de los datos en los entornos Big Data

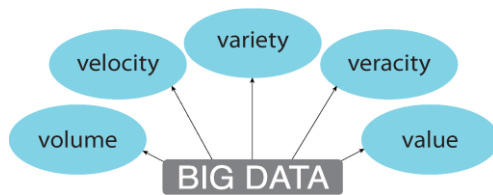


Figura 1.19 – Las cinco "V" de Big Data.

Esta sección examina las cinco características de Big Data que pueden ser utilizadas para diferenciar los datos categorizados como “Big Data” de otros tipos de datos. Los cinco rasgos de Big Data son comúnmente conocidos como las **Cinco "V"**:

- Volumen
- Velocidad
- Variedad
- Veracidad
- Valor

## Volumen



El **volumen anticipado de los datos que son procesados por las soluciones de Big Data** es importante y cada vez mayor. Un gran volumen de datos implica demandas específicas de almacenamiento y procesamiento, al igual que procesos de gestión y acceso. La Figura 1.20 ofrece una representación visual del gran volumen de datos empleados por las organizaciones y los usuarios en todo el mundo.

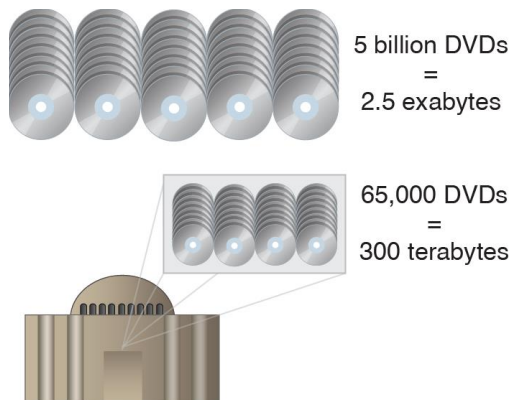
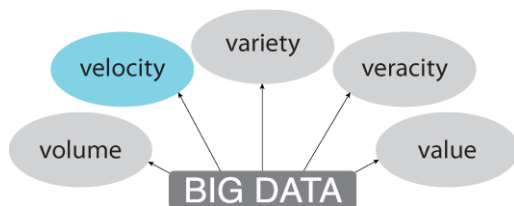


Figura 1.20 – Las organizaciones y los usuarios en todo el mundo crean 2,5 EB de datos diariamente, mientras que la Biblioteca del Congreso de los EE. UU. actualmente contiene cerca de 300 TB.

Entre las fuentes típicas de datos que son responsables de la generación de grandes volúmenes de datos están:

- las transacciones online (punto de venta, bancarias)
- datos científicos y de investigación (Gran Colisionador de Partículas, Telescopio del Atacama Large Millimeter/submillimeter Array (ALMA))
- datos de sensores (RFID, medidores inteligentes, sensores GPS)
- social media (Facebook, Twitter)

## Velocidad



Los datos de Big Data son recibidos con tal velocidad que se pueden acumular enormes datasets en periodos de tiempo cortos. Desde el punto de vista empresarial, la velocidad de los datos se traduce en la cantidad de tiempo necesaria para que los datos sean procesados una vez que llegan a la empresa. Lidar con el veloz flujo de entrada de datos requiere que la empresa diseñe soluciones altamente flexibles y disponibles de procesamiento, que cuenten con la correspondiente capacidad de almacenamiento de datos.

La velocidad no siempre será alta, dependiendo de la fuente de los datos. Por ejemplo, las imágenes de resonancias magnéticas usualmente no son generadas tan frecuentemente como las entradas en un log de un servidor web con mucho tráfico. Como se ilustra en la Figura 1.21, la posible velocidad de los datos se pone en perspectiva si consideramos que, en la actualidad, los siguientes datos son generados cada minuto: 100.000 tuits, 48 horas de video, 171 millones de correos electrónicos y 330 GB generados por un motor de base de datos Jet promedio.

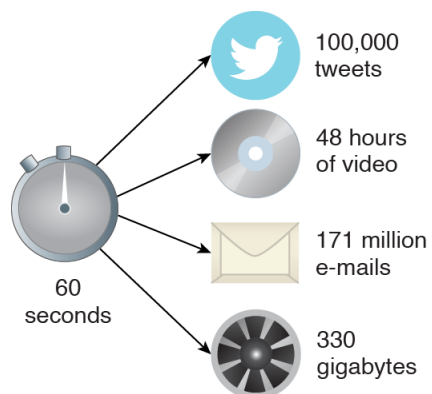
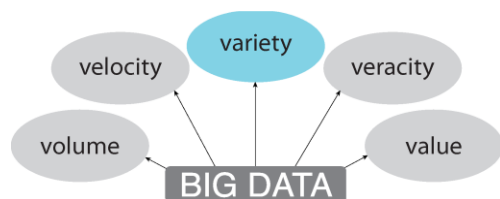


Figura 1.21 – Entre los ejemplos de datasets altamente veloces de Big Data que son producidos cada minuto están tuits, video, correos electrónicos y GB generados por un motor de base de datos Jet promedio.

## Variedad



La variedad de los datos se refiere a los múltiples formatos y tipos de datos que deben ser compatibles con las soluciones de Big Data, como datos estructurados, semiestructurados y sin estructurar, los cuales son descritos detalladamente en la sección *Tipos de datos en los entornos Big Data*, más adelante. La variedad de datos presenta desafíos para las empresas en términos de integración, transformación, procesamiento y almacenamiento de los datos. La Figura 1.22 ofrece una representación visual de la variedad de los datos, incluyendo datos estructurados como transacciones financieras, datos semiestructurados como correos electrónicos y datos sin estructurar como imágenes.

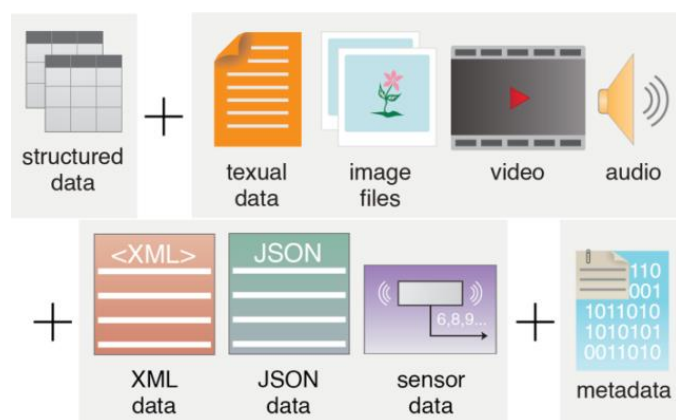
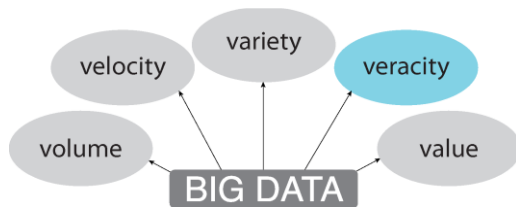


Figura 1.22 – Entre los ejemplos de datasets de Big Data de gran variedad se encuentran datos estructurados, textuales, de imagen, de video, de audio, XML, JSON, de sensores o metadata.

## Veracidad

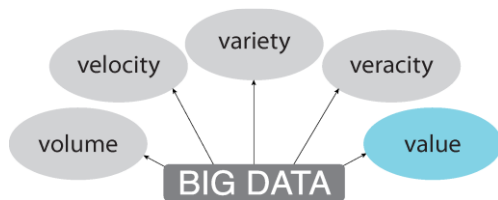


La veracidad se refiere a la calidad o fidelidad de los datos. Dentro de los entornos Big Data, existen datos que pueden ser significativos o que simplemente ocupan espacio. Cuando se evalúan en función de su veracidad, los datos pueden ser de dos tipos:

- **Ruido**; datos que no tienen valor alguno
- **Señal**; datos que tienen valor que conduce a información importante

Por lo general, los datos que son adquiridos de forma controlada —por ejemplo, mediante registros de clientes online— contienen menos ruido que los datos adquiridos mediante fuentes no controladas, como las publicaciones de blog. La cantidad de ruido (datos sin valor) o la proporción entre ruido y señal varía de acuerdo al tipo de datos presentes.

## Valor



El valor se define como la utilidad que los datos tienen para una empresa. La característica de valor está directamente relacionada con la característica de veracidad, en la medida en que, entre más alta sea la fidelidad de los datos, mayor será el valor de los mismos para la empresa. El valor también depende de qué tanto tiempo consuma el procesamiento de los datos, ya que el valor y el tiempo de procesamiento son inversamente proporcionales. Cuanto más tiempo se tardan los datos en ser convertidos en información significativa, menor será el valor que tengan para la empresa, ya que afecta la velocidad con la cual se pueden tomar decisiones informadas. Las Figuras 1.23 y 1.24 ilustran una comparación entre el valor potencial que los datos podrían llegar a tener y el tiempo necesario para el análisis de los mismos.

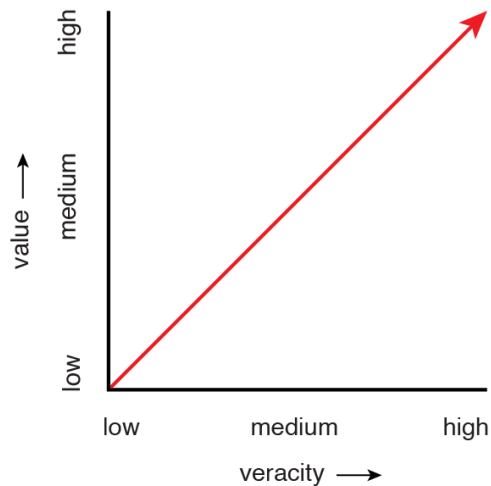


Figura 1.23 – Cuanto más confiables sean los datos, mayor será el posible valor que tengan para la empresa.

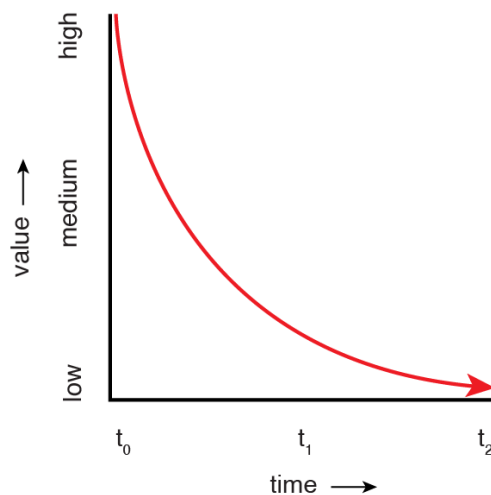


Figura 1.24 – Cuanto más tiempo tarde el análisis de los datos, menor será el posible valor que tengan para la empresa.

Al igual que la veracidad y el tiempo, el valor también está determinado por las siguientes consideraciones:

- ¿Qué tan bien fueron almacenados los datos?
- ¿Los datos fueron despojados de cualquier atributo valioso?
- Durante el análisis de datos (Data Analysis), ¿se hacen las preguntas correctas?
- ¿Los resultados del análisis de datos (Data Analysis) son comunicados con precisión a las personas correctas encargadas de la toma de decisiones?

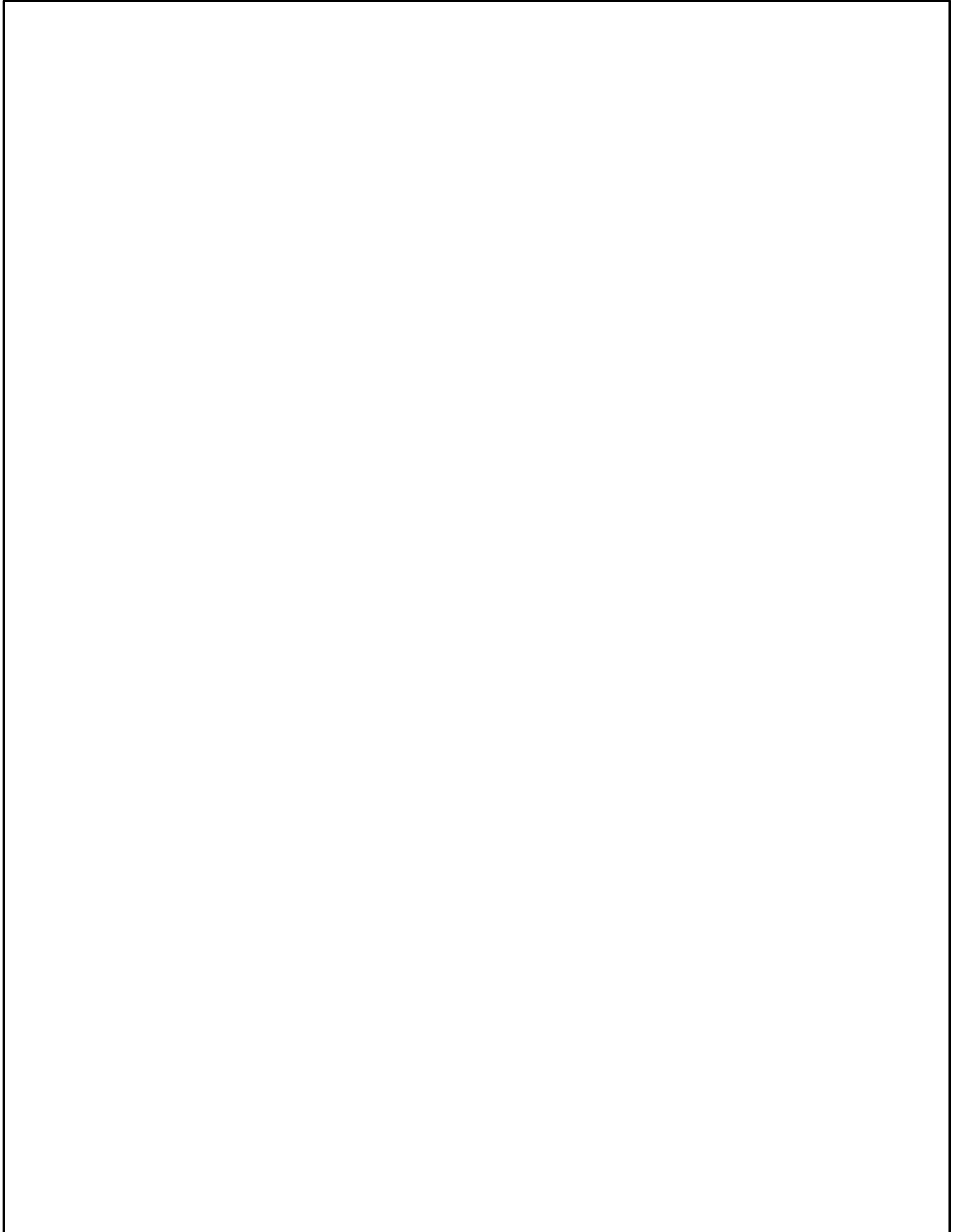


[illegible]

[illegible]

[illegible]

## Notas / Bocetos



# Tipos de datos en los entornos Big Data

Esta sección examina los tipos de datos que son procesados por las soluciones de Big Data, los cuales pueden ser clasificados en las siguientes categorías:

- datos estructurados
- datos sin estructurar
- datos semiestructurados

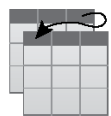
Estos tipos hacen referencia a la organización interna de los datos, lo que también se conoce como **formatos de datos**. Los **metadata** serán presentados brevemente en esta sección, a pesar de que técnicamente no son datos, sino que corresponden a otra forma de datos cuya propia estructura puede variar.

## Datos estructurados

Los datos estructurados:

- cumplen un modelo de datos o esquema
- son almacenados de forma tabular
- pueden ser relacionales

Por lo general, los datos estructurados son almacenados en bases de datos relacionales, y con frecuencia son generados por aplicaciones empresariales personalizadas, los sistemas de Planificación de Recursos Empresariales (ERP) y los sistemas de Relación con los Clientes (CRM). Estos datos normalmente no tienen ningún requisito especial de preprocesamiento o almacenamiento. Algunos ejemplos son las transacciones bancarias, los registros de los sistemas de OLTP y los registros de clientes.



relational/  
tabular  
data

Figura 1.25 - Símbolo utilizado para representar datos estructurados relacionales almacenados de forma tabular.

## Datos sin estructurar

Los datos sin estructurar:

- no cumplen un modelo de datos o esquema
- generalmente no son consistentes ni relacionales

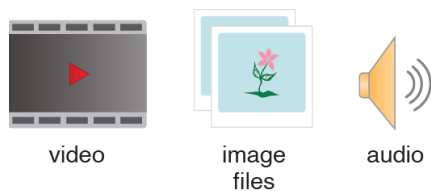


Figura 1.26 - Símbolos utilizados para representar video, archivos de imágenes y audio.

Los datos sin estructurar se encuentran en forma textual o binaria. Algunos ejemplos son archivos de imágenes, audio y video. Técnicamente, los archivos de texto y binarios tienen una estructura que está definida por el formato del archivo en sí. Omitimos este hecho para concentrarnos exclusivamente en el formato de los datos contenidos en el archivo. Por lo general, los datos sin estructurar comprenden hasta el 80% de los datos al interior de una empresa, y tienen un índice de crecimiento mayor que el de los datos estructurados. La Figura 1.27 contiene un gráfico circular de las proporciones generales de los datos estructurados y sin estructurar al interior de una empresa.

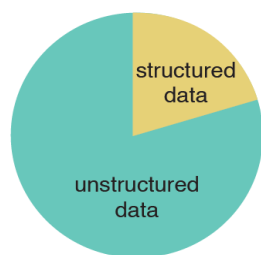


Figura 1.27 – Al interior de una empresa, generalmente el 80% de los datos está compuesto por datos sin estructurar, mientras que el 20% restante está compuesto por datos estructurados

A diferencia de los datos estructurados, los datos sin estructurar normalmente requieren una lógica especial o personalizada al momento de ser preprocesados y almacenados. No pueden ser procesados ni consultados intrínsecamente por medio de SQL ni de otras características tradicionales de programación, y usualmente no se corresponden bien con las bases de datos relacionales. Una base de datos NoSQL es una base de datos no relacional que puede ser utilizada para almacenar datos sin estructurar y datos estructurados.

#### NOTA

NoSQL es examinado a fondo en el *Módulo 2: Conceptos de análisis y tecnología de Big Data*.

## Datos semiestructurados

Los datos semiestructurados tienen un nivel definido de estructura y consistencia, pero no son relacionales. En su mayoría se encuentran en formatos textuales, como archivos XML o JSON, y generalmente, su procesamiento es más sencillo que el de los datos sin estructurar.

Entre los ejemplos de fuentes comunes de datos semiestructurados están los intercambios electrónicos de datos (EDI), los correos electrónicos, las hojas de cálculo, los canales RSS y los datos de sensores. A menudo, los datos semiestructurados tienen requisitos especiales de preprocesamiento y almacenamiento, especialmente si el formato subyacente no está basado en texto.



Figura 1.28 - Símbolos utilizados para representar datos de XML, JSON y de sensor.

## Metadata

**Los metadata proporcionan información sobre las características y la estructura de un dataset.** En su mayoría, este tipo de datos son generados por máquinas y anexados automáticamente a los datos. Son esenciales para el procesamiento, almacenamiento y análisis de Big Data. Algunos ejemplos de metadata son:

- las etiquetas XML que brindan información sobre el autor y la fecha de creación de un documento
- los atributos que proporcionan información sobre el tamaño del archivo y la resolución de una fotografía digital

Las soluciones de Big Data dependen de los metadata, particularmente durante el procesamiento de datos semiestructurados y sin estructurar.



Figura 1.29 – Símbolo utilizado para representar metadata.

## Tipos de datos y veracidad

Los datos semiestructurados y sin estructurar tienen una mayor proporción entre ruido y señal que los datos estructurados. Debido a esta mayor cantidad de ruido, se requiere la limpieza (Cleansing) automatizada y la verificación de los datos al momento de realizar procesos ETL. La Figura 1.30 ilustra la proporción entre ruido y señal para los distintos tipos de datos.

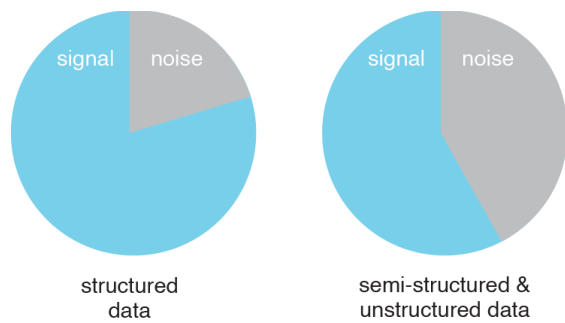


Figura 1.30 – Resumen de las proporciones entre ruido y señal para los datos estructurados, semiestructurados y sin estructurar.



## Ejercicio 1.2: complete los espacios en blanco

1. Una \_\_\_\_\_ puede contener bases de datos analíticas que pueden mejorar los tiempos de respuesta de las consultas.
2. \_\_\_\_\_ es un proceso utilizado para cargar los datos desde un sistema origen hasta un sistema destino, y es la principal operación utilizada para enviar datos a las bodegas de datos digitales (Data Warehouse).
3. Las cinco principales características de Big Data que la diferencian de los datos tradicionales son volumen, velocidad, \_\_\_\_\_, \_\_\_\_\_ y \_\_\_\_\_.
4. La característica de valor de Big Data \_\_\_\_\_ de qué tanto tiempo consume el procesamiento de los datos.
5. En general, los datos que son procesados por las soluciones de Big Data pueden estar clasificados en los siguientes tipos de datos o formatos: \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_ y \_\_\_\_\_.

*Las respuestas al ejercicio se encuentran al final de este cuadernillo.*

[illegible]

[illegible]

[illegible]

## Notas / Bocetos

# Análisis fundamental, analítica y tipos de aprendizaje automático (Machine Learning)

## Tipos de análisis de datos (Data Analysis)

En la sección *Terminología y conceptos fundamentales* se presentó el término “análisis de datos (Data Analysis)” y se mostró un ejemplo sencillo. Las secciones a continuación describen con mayor profundidad los siguientes tipos básicos de análisis de datos (Data Analysis):

- Análisis cuantitativo
- Análisis cualitativo
- Minería de datos (Data Mining)

Cada descripción incluye un ejemplo sencillo basado en el escenario de ventas de conos de helado usado en la descripción inicial del análisis de datos (Data Analysis).



Figura 1.31 – Símbolo utilizado para representar el análisis de datos (Data Analysis).

## Análisis cuantitativo

El análisis cuantitativo es una técnica de análisis de datos (Data Analysis) orientada a cuantificar patrones y correlaciones hallados en los datos. Esta técnica implica el análisis de un gran número de observaciones de un dataset con base en técnicas estadísticas. Debido al amplio tamaño de la muestra, los resultados pueden aplicarse de manera general a todo el dataset.

Los resultados del análisis cuantitativo son de naturaleza absoluta y, por lo tanto, pueden ser usados para realizar comparaciones numéricas. Por ejemplo, en un análisis cuantitativo de las ventas de helados, se puede encontrar que un aumento de 5 grados en la temperatura incrementa las ventas en un 15%.

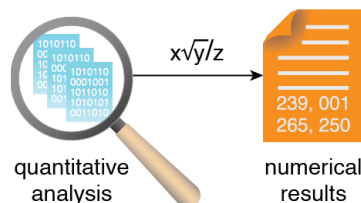


Figura 1.32 - Símbolos utilizados para representar el análisis cuantitativo y los resultados numéricos.

## Análisis cualitativo

El análisis cualitativo es una técnica de análisis de datos (Data Analysis) orientada a describir cualidades de varios datos por medio de palabras. En contraste con el análisis de datos (Data Analysis) cuantitativo, esto implica analizar una pequeña muestra con mayor profundidad.

Los resultados de este análisis no se pueden aplicar de forma general a todo un dataset debido al pequeño tamaño de la muestra. Tampoco pueden ser medidos numéricamente o usados para comparaciones numéricas. Por ejemplo, un análisis de las ventas de conos de helado puede indicar que las cifras de las ventas en mayo **no fueron tan altas comparadas** con el mes de junio. Los resultados del análisis solo muestran que las cifras “no fueron tan altas comparadas con”, mas no indican ninguna diferencia numérica.

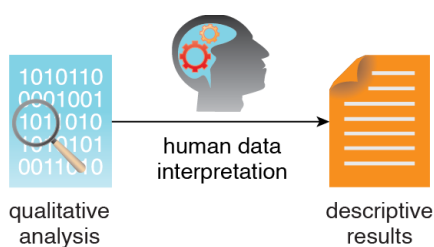


Figura 1.33 - Símbolos usados para representar el análisis cualitativo, la interpretación humana de los datos y los resultados descriptivos.

## Minería de datos (Data Mining)

La minería de datos (Data Mining), también conocida como exploración de datos, es una forma especializada de análisis de datos (Data Analysis) dedicada a los datasets grandes. En relación con el análisis de Big Data, la minería de datos (Data Mining) por lo general se refiere a técnicas automáticas basadas en software que filtran los datasets masivos para identificar patrones y tendencias. Específicamente, implica extraer patrones ocultos o desconocidos en los datos con la intención de identificar patrones antes desconocidos. La minería de datos (Data Mining) constituye la base para la analítica predictiva y la Inteligencia de negocios (BI).



Figura 1.34 – Símbolo utilizado para representar la minería de datos (Data Mining).

## Análisis y analítica

El tiempo y el esfuerzo requeridos para llevar a cabo un análisis manual aumentan considerablemente con Big Data. Es posible que estas técnicas no brinden hallazgos

exactos ni de manera oportuna debido al volumen, velocidad, y/o variedad de los datos. Estas ineficiencias potenciales se agravan aún más si se debe repetir el análisis.

Las herramientas de analítica pueden automatizar el análisis de datos (Data Analysis) usando tecnologías computacionales altamente escalables que aplican técnicas de análisis estadísticos cuantitativos automatizados, minería de datos (Data Mining) y aprendizaje automático (Machine Learning), como se muestra en la Figura 1.35.

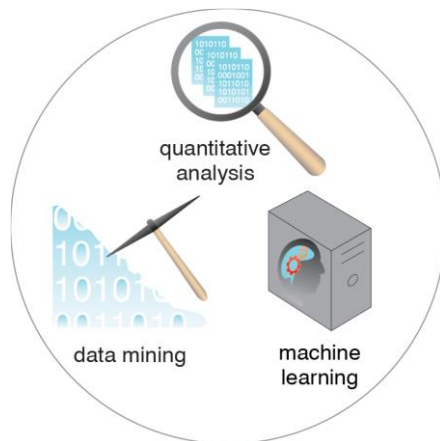


Figura 1.35 – Automatizar el análisis cuantitativo, la minería de datos (Data Mining) y el aprendizaje automático (Machine Learning) puede minimizar las ineficiencias en caso de que se deba repetir el análisis.

## Tipos de analítica

El término “analítica” se describió inicialmente en la sección *Terminología y conceptos fundamentales*. Esta sección explora más a fondo la analítica, describiendo a continuación cuatro tipos comunes de analítica:

- Analítica descriptiva
- Analítica diagnóstica
- Analítica predictiva
- Analítica prescriptiva

La Figura 1.36 ilustra estos tipos de analítica según su valor y complejidad.



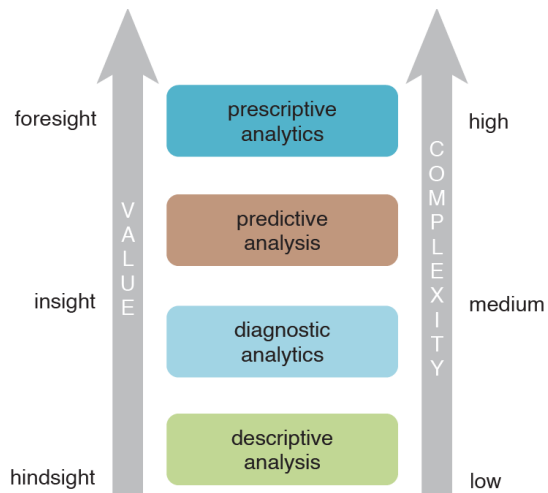


Figura 1.36 – El valor y la complejidad aumentan progresivamente, comenzando por la analítica descriptiva y terminando en la analítica prescriptiva.

## Analítica descriptiva

La **analítica descriptiva** se ejecuta para responder preguntas sobre eventos que ocurrieron.

Algunas preguntas de ejemplo pueden ser:

- *¿Cuáles son los datos de las ventas de los últimos 12 meses?*
- *¿Cuántas llamadas de soporte técnico fueron recibidas y categorizadas según la gravedad y ubicación geográfica?*
- *¿Cuál es la comisión mensual que gana cada agente de ventas?*

Alrededor del 80% de la analítica es de naturaleza descriptiva. En términos de valor, la analítica descriptiva proporciona un valor mínimo y requiere un conjunto relativamente básico de habilidades.

A menudo, la analítica descriptiva se ejecuta mediante reportes o tableros de control (Dashboards) especializados, como se muestra en la Figura 1.37. Por lo general, los reportes son de naturaleza estática y muestran datos históricos que son presentados en forma de grillas de datos o gráficos. Las consultas son ejecutadas en los sistemas de OLTP o en los datos obtenidos a partir de una variedad de otros sistemas de información, tales como CRM Y ERP.

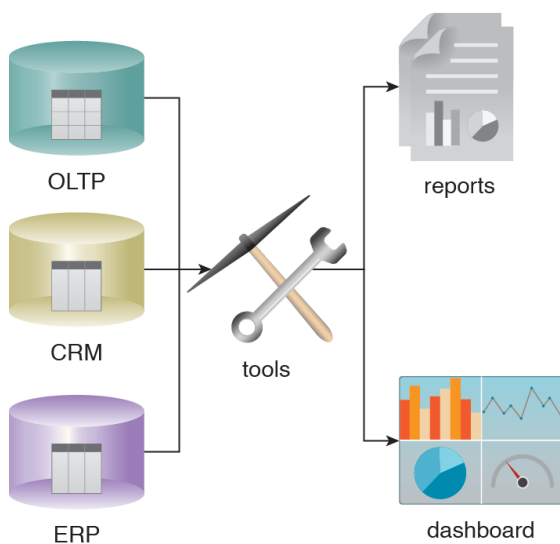


Figura 1.37 – Las consultas desde los sistemas en el costado izquierdo se pueden ejecutar por medio de herramientas de analítica descriptiva que son comunicadas mediante reportes o tableros de control (Dashboards), en el costado derecho.

## Analítica diagnóstica

La **analítica diagnóstica** tiene como objetivo determinar la causa de un fenómeno que ocurrió en el pasado, usando preguntas que se enfocan en la razón del evento.

Algunas preguntas de ejemplo pueden ser:

- *¿Por qué las ventas del segundo trimestre fueron menores que las de primer trimestre?*
- *¿Por qué se han recibido más llamadas de soporte técnico de la región este que de la región oeste?*
- *¿Por qué hubo un incremento en las tasas de readmisión de pacientes en los últimos tres meses?*

Se considera que la analítica diagnóstica proporciona más valor que la analítica descriptiva, y que requiere un conjunto de habilidades más avanzadas. Normalmente, también requiere recopilar datos a partir de múltiples fuentes y almacenarlos en una estructura que se presta para realizar drill-downs y roll-ups, como se muestra en la Figura 1.38. Los resultados de esta analítica pueden ser observados por medio de herramientas interactivas de visualización, que permiten que los usuarios identifiquen tendencias y patrones. Las consultas ejecutadas son más complejas en comparación con la analítica descriptiva y son realizadas sobre datos multidimensionales almacenados en sistemas de OLAP.

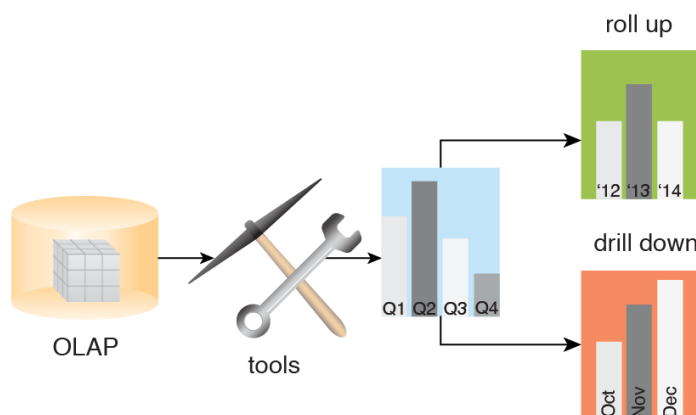


Figura 1.38 – La analítica diagnóstica puede producir datos que son adecuados para realizar drill-downs y roll-ups.

## Analítica predictiva

La analítica predictiva se ejecuta en un intento por determinar el resultado de un evento que podría ocurrir en el futuro.

Las preguntas normalmente se formulan usando una lógica condicional *qué tal sí*, como en los siguientes ejemplos:

- *¿Cuáles son las probabilidades de que un cliente incurra en el incumplimiento de un préstamo si él no ha hecho el pago mensual?*

- Si se administra el medicamento B en vez del medicamento A, ¿cuál será la tasa de supervivencia del paciente?
- Si un cliente compra los Productos A y B, ¿qué posibilidades hay de que también compre el Producto C?

La **analítica predictiva** intenta predecir el resultado de un evento. Las predicciones se hacen con base en patrones, tendencias y excepciones encontradas en datos históricos y actuales. Este análisis permite identificar riesgos y oportunidades.

La analítica predictiva implica el uso de datasets grandes compuestos por datos internos y externos, además de técnicas estadísticas, de análisis cuantitativo, de aprendizaje automático (Machine Learning) y de minería de datos (Data Mining). Se considera que este tipo de analítica proporciona más valor y requiere un conjunto de habilidades más avanzadas que las analíticas descriptiva y diagnóstica. Por lo general, las herramientas usadas abstraen particularidades estadísticas, proporcionando una interfaz de front-end fácil de usar, como se muestra en la Figura 1.39.

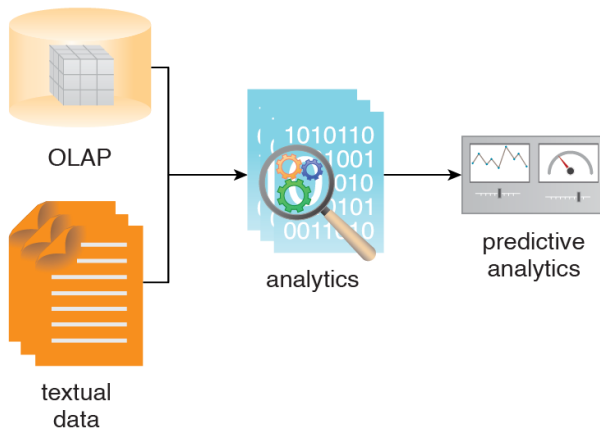


Figura 1.39 – Las herramientas de la analítica predictiva pueden proporcionar interfaces de front-end fáciles de usar.

## Analítica prescriptiva

La **analítica prescriptiva** está basada en los resultados de la analítica predictiva, al indicar acciones que se deberían realizar. Esta analítica se enfoca en qué opción indicada se debe seguir y en por qué y cuándo se debería seguir, con el fin de obtener una ventaja o mitigar un riesgo.

Algunas preguntas de ejemplo pueden ser:

- Entre tres opciones de medicamentos, ¿cuál ofrece los mejores resultados?
- ¿Cuándo es el mejor momento para comercializar una mercancía particular?

La analítica prescriptiva proporciona más valor que cualquier otro tipo de analítica, y en consecuencia, requiere un conjunto de habilidades más avanzadas, además de software

y herramientas especializadas. Se calculan varios resultados y se sugiere el mejor plan de acción para cada uno de ellos. **Se hace un cambio del enfoque explicativo al enfoque consultivo**, y puede incluir la simulación de varias situaciones.

La analítica prescriptiva incorpora **datos internos**, incluidos datos de ventas actuales e históricos, información de los clientes, datos del producto, normas comerciales y también **datos externos**, incluidos datos de social media, datos climatológicos y datos demográficos. La analítica prescriptiva implica el uso de normas comerciales y grandes cantidades de datos internos y/o externos para simular los resultados y prescribir el mejor plan de acción, como se muestra en la Figura 1.40.

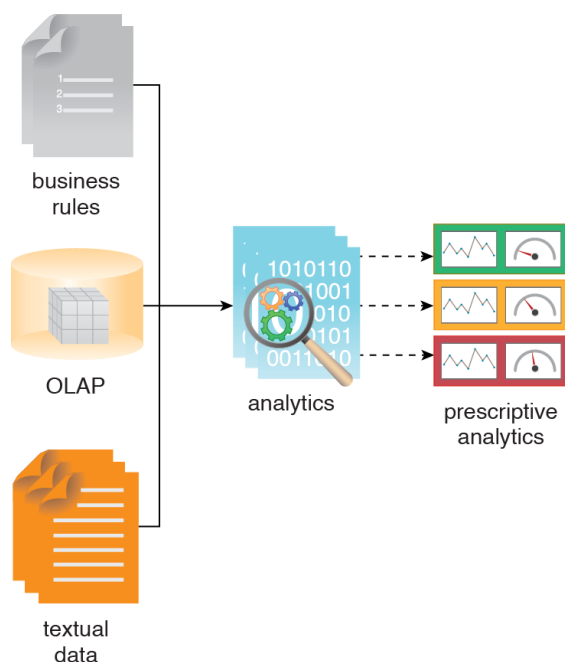


Figura 1.40 – La analítica prescriptiva implica el uso de normas comerciales y de datos internos y externos con el fin de realizar un análisis más profundo.

## Aprendizaje automático (Machine Learning)

El aprendizaje automático (Machine Learning) es el proceso de enseñar a las computadoras a aprender a partir de datos existentes y a aplicar el conocimiento adquirido para formular predicciones sobre datos desconocidos. Esto implica identificar patrones en los datos de entrenamiento y clasificar datos nuevos y no mostrados con base en patrones conocidos. Los algoritmos de aprendizaje automático (Machine Learning) también permiten ajustar los comportamientos utilizando un loop de retroalimentación, a la vez que funcionan con datasets nuevos. Normalmente, estos algoritmos se dividen en los siguientes dos tipos:

- **Aprendizaje supervisado**
- **Aprendizaje no supervisado**

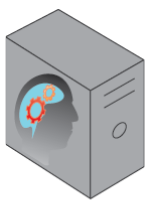


Figura 1.41 – Símbolo utilizado para representar el aprendizaje automático (Machine Learning).

### **Tipos de aprendizaje automático (Machine Learning)**

Primero, los datos de muestra son enviados al algoritmo de **aprendizaje supervisado**, donde ya se conocen las categorías de los datos. Con base en los datos introducidos, el algoritmo comprende qué datos corresponden a qué categoría. Posteriormente, el algoritmo puede aplicar el comportamiento que aprendió para categorizar los datos desconocidos.

En un algoritmo de **aprendizaje no supervisado**, no se conocen las categorías de los datos y no se envía ningún dato de muestra. En vez de eso, el algoritmo intenta categorizar los datos agrupándolos según atributos similares.

### **Comparación entre el aprendizaje automático (Machine Learning) y la minería de datos (Data Mining)**

A pesar de que la minería de datos (Data Mining) y el aprendizaje automático (Machine Learning) están estrechamente relacionados, tienen diferencias notables. Considerando que la minería de datos (Data Mining) encuentra patrones **ocultos** y relaciones basadas en atributos de datos antes desconocidos, el aprendizaje automático (Machine Learning) hace predicciones categorizando datos basados en patrones **conocidos**.

**La minería de datos (Data Mining) puede emplear algoritmos de aprendizaje automático (Machine Learning)** —como el aprendizaje no supervisado— para extraer atributos antes desconocidos. Esto se logra mediante la categorización de datos, lo cual lleva a la identificación de patrones. El aprendizaje automático (Machine Learning) puede usar el resultado de la minería de datos (Data Mining) o los patrones identificados para clasificarlos aún más mediante el aprendizaje supervisado, como se muestra en la Figura 1.42.

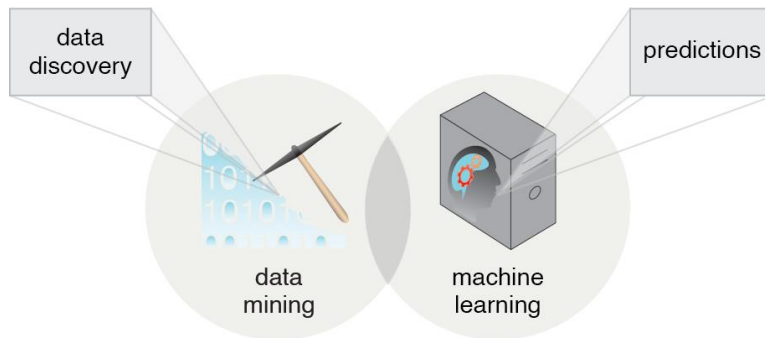


Figura 1.42 – El aprendizaje automático (Machine Learning) hace predicciones para clasificar datos y la minería de datos (Data Mining) explora los patrones usados para automatizar el aprendizaje automático (Machine Learning).

### Ejercicio 1.3: complete los espacios en blanco

1. La analítica se puede categorizar en cuatro tipos diferentes de acuerdo con el valor de su atributo: \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_ y \_\_\_\_\_.
2. Los resultados de la analítica \_\_\_\_\_ se pueden observar por medio de herramientas interactivas de visualización que permiten identificar fácilmente tendencias y patrones.
3. La analítica \_\_\_\_\_ tiene más valor para las empresas porque esta técnica recomienda un plan de acción que puede ser seguido.
4. Los algoritmos de aprendizaje automático (Machine Learning) se pueden categorizar en aprendizaje \_\_\_\_\_ y \_\_\_\_\_.

*Las respuestas al ejercicio se encuentran al final de este cuadernillo.*

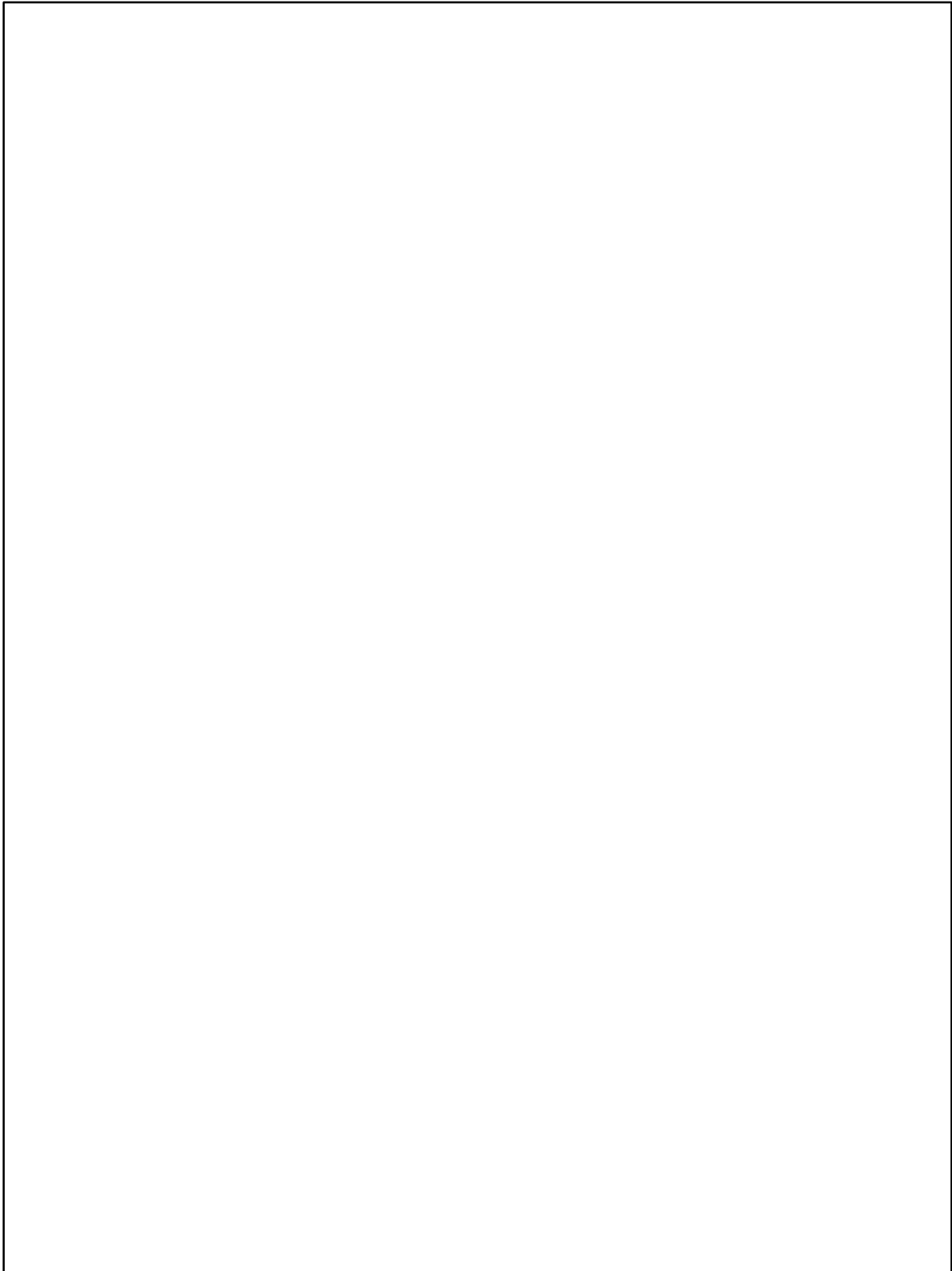


[illegible]

[illegible]

[illegible]

## Notas / Bocetos



# Inteligencia de negocios (BI) y Big Data

Las modernas soluciones empresariales de Big Data dependen de la Inteligencia de negocios (BI) y las bodegas de datos digitales (Data Warehouses) como componentes centrales de los entornos y ecosistemas de Big Data. Por otro lado, la llegada de Big Data conlleva al surgimiento de tecnologías avanzadas de Inteligencia de negocios (BI) y de bodegas de datos digitales (Data Warehouses), hasta el punto en que ha surgido una nueva generación de estas plataformas.

Esta sección compara la Inteligencia de negocios (BI) tradicional y de última generación y los entornos de bodegas de datos digitales (Data Warehouses); además, define la relación de los dos con las soluciones de Big Data.

## La Inteligencia de negocios (BI) tradicional

La Inteligencia de negocios (BI) tradicional usa la analítica descriptiva y diagnóstica para proporcionar información sobre eventos históricos y actuales. No es realmente “inteligente” porque únicamente proporciona respuestas a preguntas formuladas correctamente. Para formular preguntas correctamente es necesario entender los problemas e inconvenientes de la empresa y de los datos en sí. La Inteligencia de negocios (BI) realiza reportes sobre diferentes KPI por medio de:

- reportes especializados
- tableros de control (Dashboards)

## Inteligencia de negocios (BI) tradicional: reportes especializados

Preparar reportes especializados es un proceso que implica procesar datos manualmente para generar reportes personalizados, como se muestra en la Figura 1.43. Por lo general, los reportes especializados se enfocan en área específica de la empresa, como mercadeo o gestión de la cadena de suministro. Se generan reportes personalizados muy detallados que a menudo son tabulares.

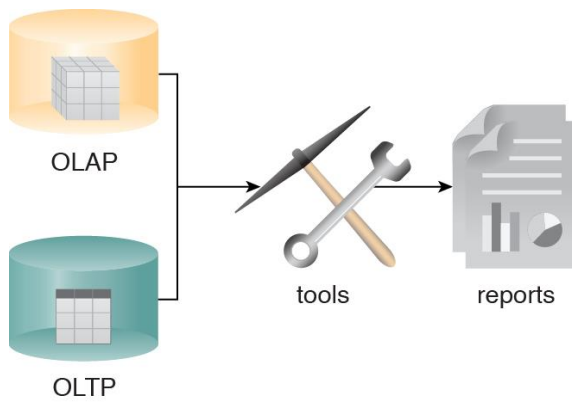


Figura 1.43 – Las fuentes de datos de OLTP Y OLAP pueden ser usadas por las herramientas de Inteligencia de negocios (BI) para la elaboración de reportes especializados y tableros de control (Dashboards).

## Inteligencia de negocios (BI) tradicional: tableros de control (Dashboards)

Los tableros de control (Dashboards) ofrecen una perspectiva holística de las áreas clave de la empresa. La información que se muestra en los tableros de control (Dashboards) es generada en intervalos periódicos en tiempo real o prácticamente real. La presentación de los datos en los tableros de control (Dashboard) es de naturaleza gráfica; utiliza gráficos, gráficos de barras, circulares e indicadores, según se muestra en la Figura 1.44.

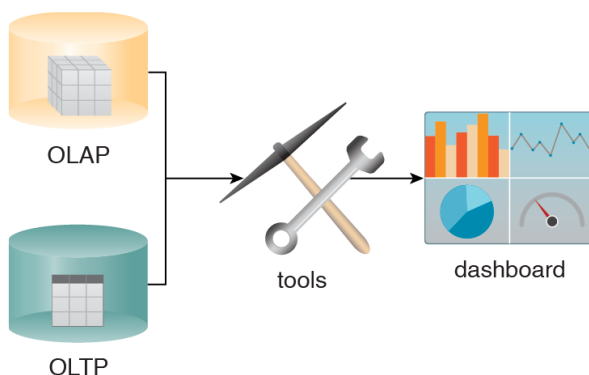


Figura 1.44 – Las herramientas de Inteligencia de negocios (BI) usan tanto el OLAP como el OLTP para mostrar la información en los tableros de control (Dashboards).

Como se explicó anteriormente, las bodegas de datos digitales (Data Warehouses) y los data marts contienen información consolidada y validada sobre entidades de negocio en toda la empresa. La Inteligencia de negocios (BI) tradicional no puede funcionar eficientemente sin los data marts, puesto que estos contienen los datos optimizados y segregados que la Inteligencia de negocios (BI) necesita para realizar reportes. Sin los data marts, los datos deben ser extraídos de la bodega de datos digital (Data Warehouse) por medio de un proceso ETL especializado cada vez que se deba hacer una consulta. Esto aumenta el tiempo y los esfuerzos de ejecución de consultas y generación de reportes.

La Inteligencia de negocios (BI) tradicional usa bodegas de datos digitales (Data Warehouses) y data marts para realizar reportes y analizar datos (Data Analysis), porque esto permite realizar consultas complejas de análisis de datos (Data Analysis) con múltiples joins y agregaciones, como se muestra en la Figura 1.45.

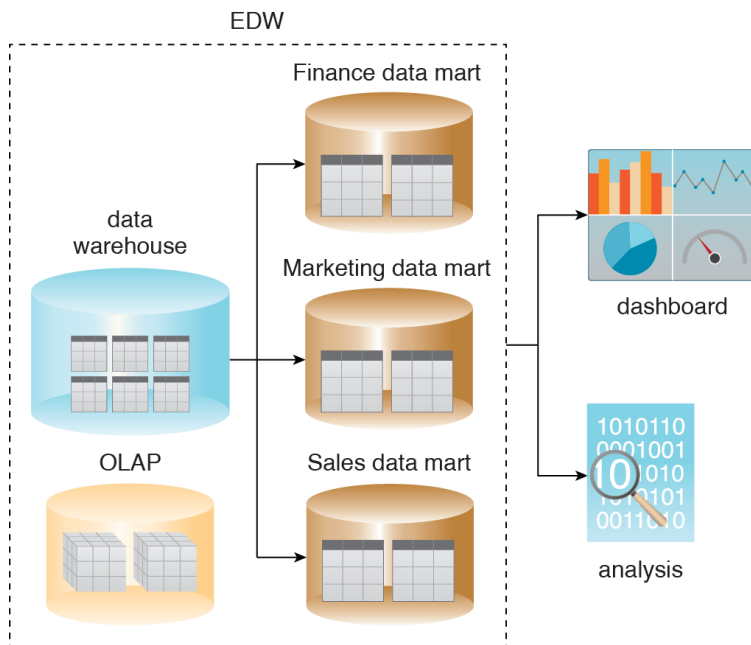


Figura 1.45 – Ejemplo de Inteligencia de negocios (BI) tradicional.

## Inteligencia de negocios (BI) de Big Data

La Inteligencia de negocios (BI) de Big Data está basada en la Inteligencia de negocios (BI) tradicional ejecutando acciones sobre los datos limpiados y consolidados de toda la empresa en la bodega de datos digital (Data Warehouse) y combinándolos con fuentes de datos semiestructurados y sin estructurar. Comprende tanto la analítica predictiva como la prescriptiva para facilitar la **comprensión, a nivel empresarial**, de la forma en que funciona un negocio.

Si bien los análisis de la Inteligencia de negocios (BI) tradicional generalmente se enfocan en procesos individuales de la empresa, los análisis de la Inteligencia de negocios (BI) de Big Data se enfocan simultáneamente en múltiples procesos de negocio. Esto permite hallar patrones y anomalías en mayor medida dentro de la empresa. También permite la exploración de datos, al identificar los conocimientos e información que posiblemente antes no estaban disponibles o eran desconocidos.

La Inteligencia de negocios (BI) de Big Data requiere el análisis de datos estructurados, semiestructurados y sin estructurar en la bodega de datos digital (Data Warehouse); lo cual, a su vez, requiere una bodega de datos digital (Data Warehouse) de última generación que use nuevas características y tecnologías para almacenar datos limpiados provenientes de una variedad de fuentes en un formato único y uniforme de datos. Cuando una bodega de datos digital (Data Warehouse) tradicional se acopla con estas nuevas tecnologías, el resultado es una bodega de datos digital (Data Warehouse) híbrida que actúa como un depósito uniforme y central de datos estructurados, semiestructurados y sin estructurar, que puede suministrarles todos los datos necesarios a las herramientas de Inteligencia de negocios (BI) de Big Data. Esto elimina la necesidad



de que las herramientas Inteligencia de negocios (BI) de Big Data se conecten a múltiples fuentes de datos para que recuperen o accedan a los datos. En la Figura 1.46, una bodega de datos digital (Data Warehouse) de última generación establece una capa de acceso a datos estandarizados a través de una variedad de fuentes de datos.

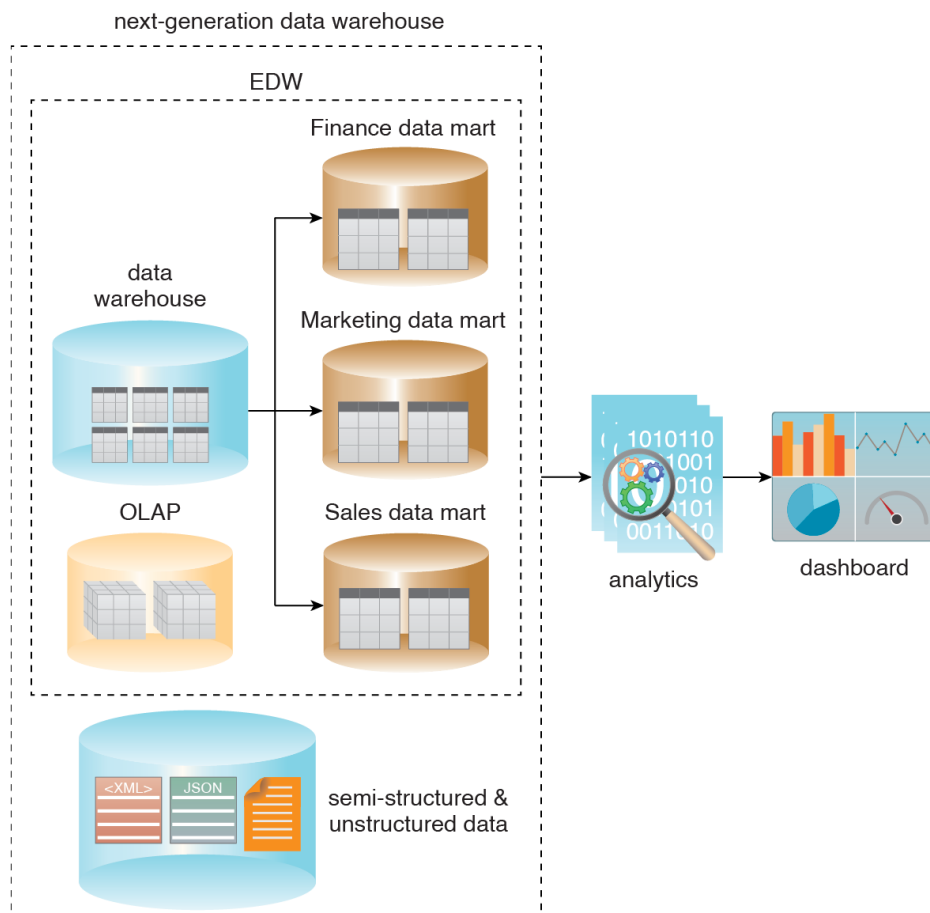


Figura 1.46 – Una bodega de datos digital (Data Warehouse) de última generación.

[illegible]

[illegible]

[illegible]

## Notas / Bocetos

# Visualización de datos y Big Data

Esta sección ofrece un vistazo a la visualización de datos, particularmente en relación con Big Data.

## Visualización de datos

La visualización de datos es una técnica a través de la cual los resultados de analítica son comunicados gráficamente utilizando gráficos, mapas, grilla de datos, infografías y alertas. Los datos representados gráficamente pueden facilitar el entendimiento de los reportes, la visualización de tendencias y la identificación de patrones.

Casi siempre, la visualización tradicional de datos proporciona cuadros estáticos y gráficos en los reportes y tableros de control (Dashboards), mientras que las herramientas modernas de visualización de datos son interactivas y ofrecen vistas de los datos de forma resumida o detallada, y están diseñadas para que quienes no tienen habilidades estadísticas o matemáticas comprendan mejor los resultados analíticos, sin tener que recurrir a las hojas de cálculo.

## Herramientas de visualización de datos

Las herramientas tradicionales de visualización de datos consultan los datos de las bases de datos relacionales, sistemas de OLAP, bodegas de datos digitales (Data Warehouses) y hojas de cálculo para presentar los resultados de la analítica descriptiva y de la analítica diagnóstica. Las soluciones de Big Data requieren herramientas de visualización de datos que se puedan conectar sin problemas a las fuentes de datos estructurados, semiestructurados y sin estructurar, y además sean capaces de manejar millones de registros de datos. Las herramientas de visualización de datos para las soluciones de Big Data generalmente utilizan tecnologías analíticas en memoria que reducen la latencia que se atribuye normalmente a las herramientas tradicionales de visualización de datos basadas en disco.

## Características de visualización de datos

Las características comunes de las herramientas de visualización usadas en Big Data son:

- **Agregación:** proporcionan una vista holística y resumida de los datos a través de múltiples contextos
- **Drill-Down:** proporcionan una vista detallada de los datos de interés al enfocarse en un subgrupo de datos de la vista resumida
- **Filtrado (filtering):** se enfocan en un conjunto particular de datos al filtrar los datos que no son de interés inmediato

- **Roll-Up**: agrupan datos en todas las múltiples categorías, para mostrar totales y subtotales
- **Análisis “qué-tal-si”**: facilitan la visualización de múltiples resultados al permitir el cambio dinámico de factores relacionados

## Herramientas avanzadas de visualización

Las herramientas avanzadas de visualización de datos de Big Data comprenden la analítica de datos **predictiva** y **prescriptiva**, y las características de transformación de datos. Estas herramientas eliminan la necesidad de usar métodos de preprocesamiento de datos (como ETL) y se conectan directamente a las fuentes de datos estructurados, semiestructurados y sin estructurar. Como parte de las soluciones de Big Data, las herramientas avanzadas de visualización de datos pueden unir datos estructurados y sin estructurar que son guardados en memoria para tener acceso rápido. Luego, las consultas y fórmulas estadísticas se pueden aplicar como parte de varias tareas de análisis de datos (Data Analysis) para visualizar datos en un formato fácil de usar; por ejemplo, un tablero de control (Dashboard).

### Ejercicio 1.4: complete los espacios en blanco

1. La Inteligencia de negocios (BI) tradicional utiliza analítica \_\_\_\_\_ y \_\_\_\_\_.
2. La \_\_\_\_\_ comunica resultados de analítica utilizando una variedad de herramientas gráficas e interactivas.
3. Las herramientas tradicionales de visualización de datos presentan los resultados tanto de la analítica \_\_\_\_\_ como de la \_\_\_\_\_.
4. La Inteligencia de negocios (BI) de Big Data añade valor a la Inteligencia de negocios (BI) tradicional al utilizar la analítica \_\_\_\_\_ y \_\_\_\_\_.

5. Las herramientas avanzadas de visualización de Big Data utilizan herramientas de analítica de datos \_\_\_\_\_ y \_\_\_\_\_.

*Las respuestas al ejercicio se encuentran al final de este cuadernillo.*

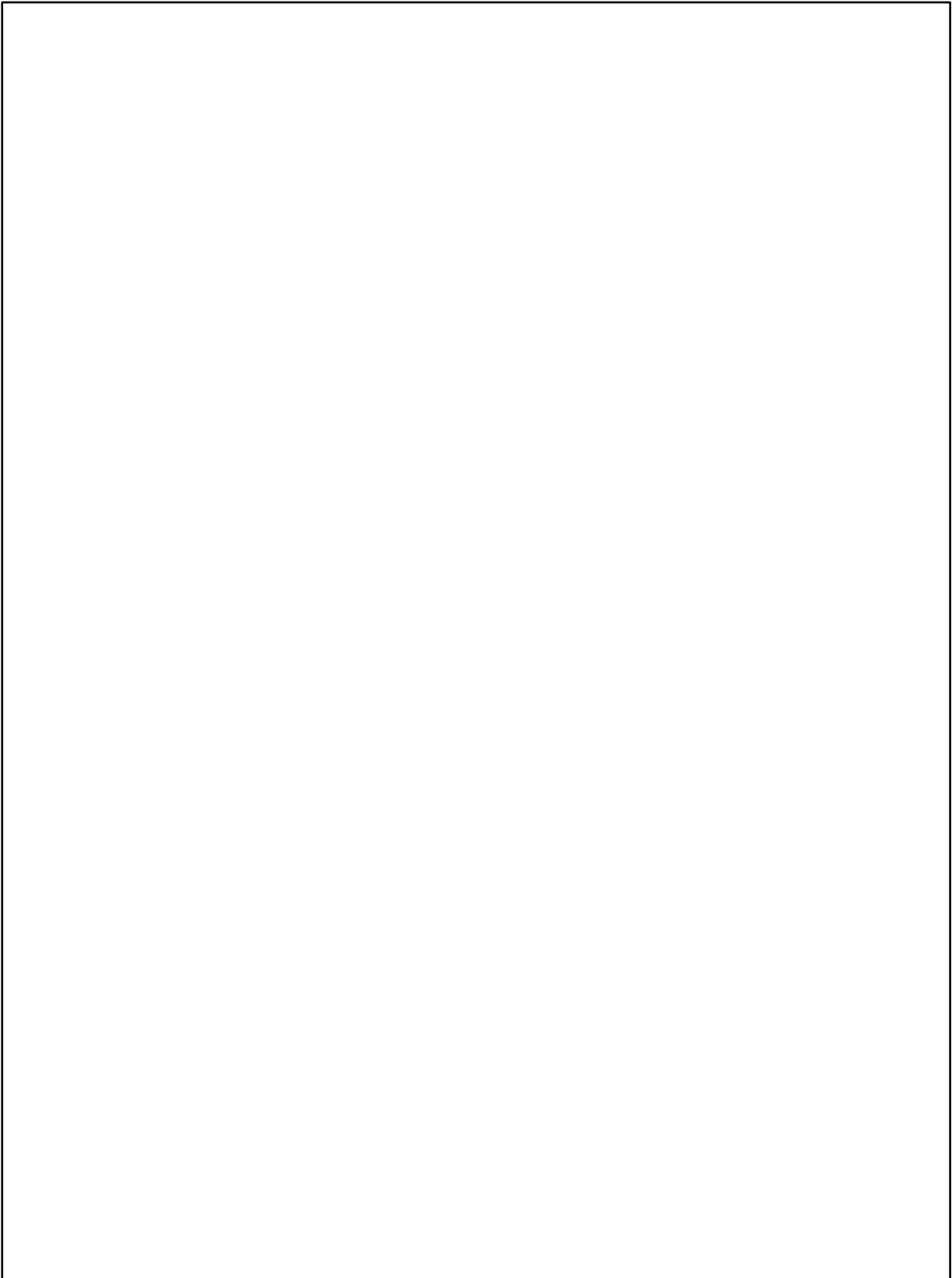


[illegible]

[illegible]

[illegible]

## Notas / Bocetos



# Elementos a tener en cuenta al planear y adoptar Big Data

## Justificación empresarial

Como las iniciativas de Big Data están orientadas inherentemente a las empresas, debe existir un caso empresarial claro para adoptar una solución de Big Data y garantizar que está justificada y cumple las expectativas. **Es necesario establecer metas claras en relación con el valor del negocio proporcionado por una solución de Big Data empresarial. Se deben sopesar los beneficios esperados en relación con los riesgos e inversiones.**

Por ejemplo, la meta puede ser crear una vista completa de la base de clientes de una empresa. Para cumplirla, puede ser necesario que todos los datos internos de los clientes sean consolidados a partir de varios sistemas. Se deben identificar y cuantificar los riesgos relacionados con la recopilación de datos exactos y relevantes, y su integración en el entorno Big Data. Es importante aceptar que **las soluciones de Big Data no son necesarias para todas las empresas**. Por ejemplo, es posible que algunas empresas simplemente no generen suficientes datos para garantizar el entorno Big Data.

## Prerrequisitos organizacionales

Los frameworks de Big Data no son soluciones integrales. Para que el análisis de datos (Data Analysis) y analítica tengan éxito y ofrezcan valor, las empresas necesitan tener **frameworks de gestión de datos y gestión de Big Data**. También se requiere que los responsables de implementar, personalizar, alimentar y usar las soluciones de Big Data cuenten con **procesos robustos y conjuntos de habilidades adecuadas**. Además, se debe evaluar la calidad de los datos que las soluciones de Big Data van procesar.

Los datos obsoletos, inválidos o mal identificados generarán entradas de baja calidad, y sin importar qué tan buena sea la solución de Big Data, esto producirá resultados de baja calidad. También se debe planificar el período de vida del entorno Big Data. Se debe definir una hoja de ruta para garantizar que cualquier expansión o aumento necesario del entorno esté planeado para estar alineado con las exigencias de la empresa.

## Aprovisionamiento de datos

Adquirir soluciones de Big Data puede ser económico debido a la disponibilidad de plataformas de código abierto y las oportunidades de aprovechar un hardware básico. Sin embargo, tal vez sea necesario contar con un presupuesto considerable para **obtener datos externos**. La naturaleza de la empresa puede hacer que los datos externos sean muy valiosos. Cuanto mayor sea el volumen y la variedad de los datos, mayores son las oportunidades de encontrar información oculta en los patrones.

Las fuentes de datos externos incluyen mercados de datos y el gobierno. Los datos que proporciona el gobierno, como los datos geoespaciales, pueden ser gratuitos. No

obstante, se deben comprar los datos comercialmente más relevantes. Este tipo de inversión puede ser recurrente, con el fin de obtener versiones actualizadas de los datasets.

## Privacidad

Realizar procesos de analítica sobre los datasets puede revelar información confidencial sobre las organizaciones o las personas naturales. Incluso el análisis de datasets separados que contienen datos aparentemente inofensivos puede revelar información privada si llegan a ser analizados conjuntamente. Esto puede generar **violaciones intencionales o accidentales de la privacidad**.

Para tratar estos asuntos de privacidad es necesario entender la naturaleza de los datos que están siendo acumulados y las regulaciones pertinentes sobre la protección de datos, así como las técnicas especiales para etiquetado y anonimato de datos. Por ejemplo, los datos de telemetría —como el log GPS de un automóvil o la lectura de datos de medidores inteligentes— recopilados en un cierto periodo de tiempo pueden revelar la ubicación o el comportamiento de una persona, como se muestra en la Figura 1.47.

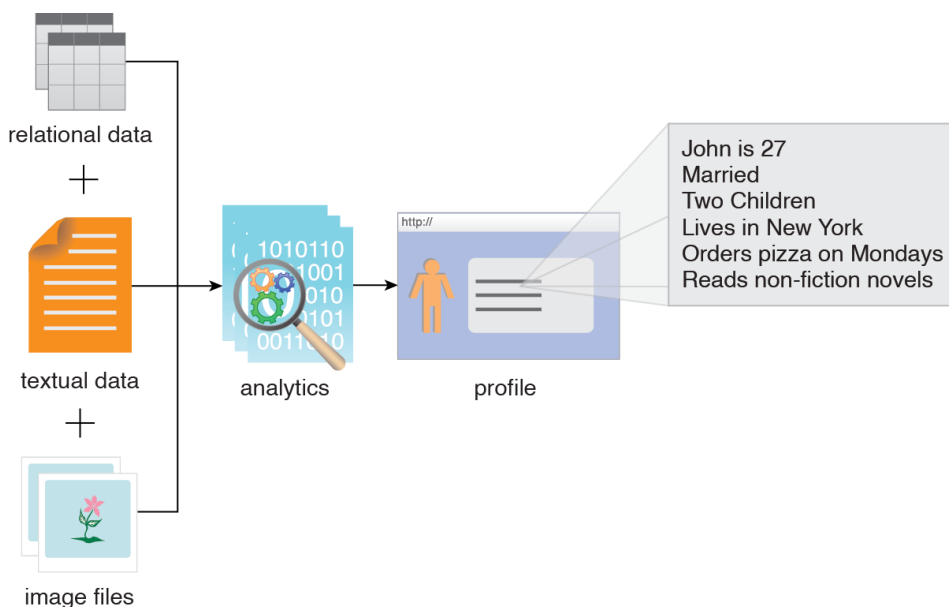


Figura 1.47 – La información recopilada a partir de la analítica ejecutada en archivos de imagen y datos relacionales y textuales es utilizada para crear el perfil de John.

## Seguridad

Algunos componentes de las soluciones de Big Data carecen de la solidez de los entornos de soluciones empresariales tradicionales en términos del control del acceso y la seguridad de los datos. La seguridad en Big Data implica garantizar que las redes de datos proporcionen acceso a repositorios lo suficientemente seguros por medio de mecanismos personalizados de autenticación y autorización.

La seguridad de Big Data además implica establecer niveles de acceso a los datos para diferentes categorías de usuarios. Por ejemplo, a diferencia de los sistemas tradicionales de gestión de bases de datos relacionales, las bases de datos NoSQL generalmente no ofrecen mecanismos incorporados sólidos de seguridad. En vez de eso, estas bases de

datos cuentan con API basadas en HTTP simple, en las que los datos se intercambian en texto plano, haciendo que los datos estén propensos a ataques basados en la red, como se muestra en la Figura 1.48.

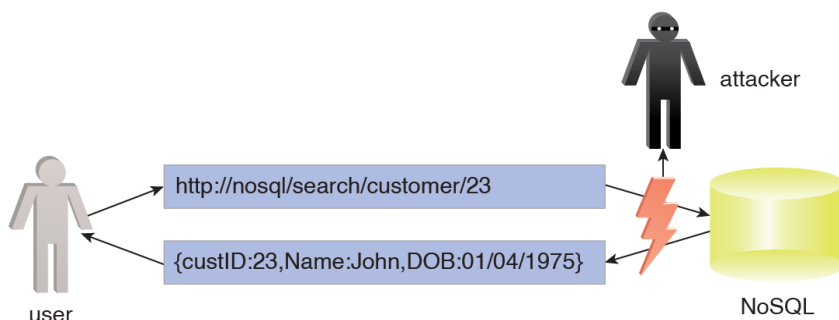


Figura 1.48 – Las bases de datos tradicionales pueden ser susceptibles a ataques basados en la red.

## Procedencia

La procedencia **se refiere a la información sobre el origen de los datos que ayuda a determinar su autenticidad y calidad**. También se utiliza con fines de auditoría. Mantener la procedencia en forma de largos volúmenes de datos que son adquiridos, combinados y procesados en múltiples etapas de procesamiento puede ser una tarea compleja. Para resolver los problemas de procedencia puede ser necesario hacer anotaciones de la información del origen y otros metadata en los datos, a medida que son generados o recibidos, como se muestra en la Figura 1.49.



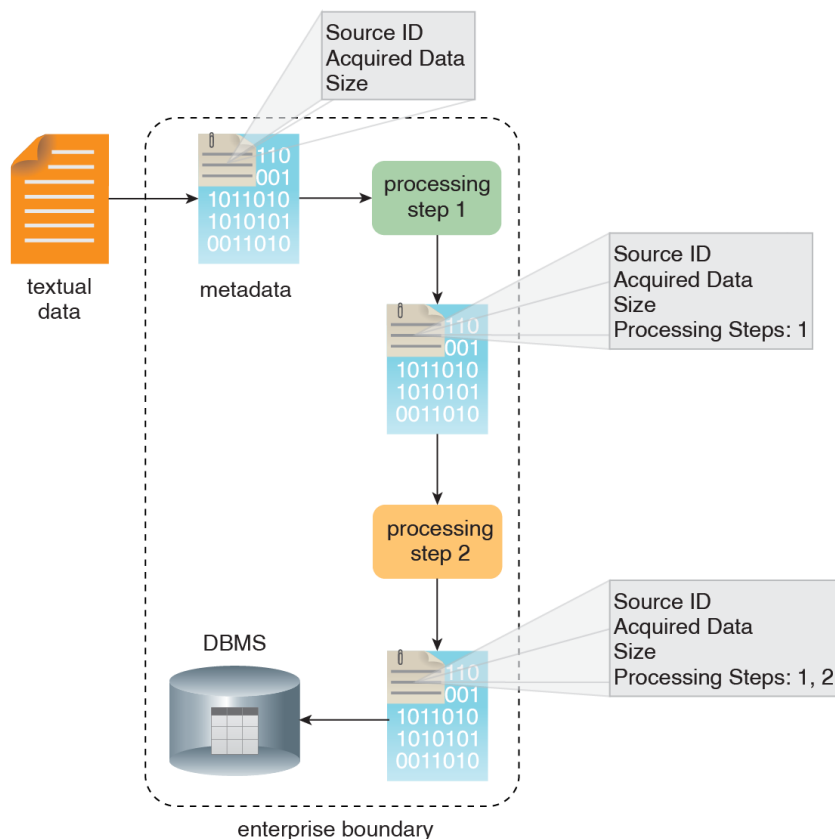


Figura 1.49 – Puede ser necesario hacer anotaciones de los atributos del origen del dataset y los detalles de los pasos del proceso a medida los datos avanzan por las etapas de transformación.

## Soporte limitado en tiempo real

Los tableros de control (Dashboards) y otras aplicaciones que requieren datos de transmisión y alertas a menudo exigen transmisión de datos en tiempo real o prácticamente real. Muchas soluciones y herramientas modernas de Big Data de código abierto están orientadas en lotes, lo que significa que **pueden tener soporte limitado para la transmisión del análisis de datos (Data Analysis), o no tener soporte alguno**. Existen soluciones propietarias de análisis de datos (Data Analysis) en tiempo real. Se puede lograr el procesamiento de datos en tiempo prácticamente real al procesar datos transaccionales a medida que son recibidos y combinarlos con datos de procesamiento por lotes (Batch Processing) que ya están resumidos.

## Diferentes problemas de rendimiento

Debido a que algunas soluciones de Big Data deben procesar altos volúmenes de datos, en ocasiones el rendimiento puede convertirse en un problema. Por ejemplo, casos en los que grandes datasets están acompañados por algoritmos complejos de búsqueda, lo que puede generar mayores tiempos de consulta. Otro ejemplo tiene que ver con el ancho de banda disponible. Con los volúmenes de datos cada vez mayores, el tiempo para

transferir una unidad de datos puede exceder el tiempo real de procesamiento de datos, como se muestra en la Figura 1.50.

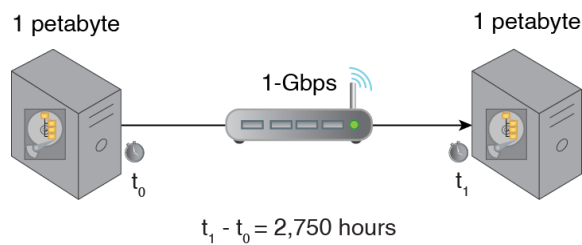


Figura 1.50 – Transferir 1 PB de datos por medio de una conexión LAN de 1 Gigabit al 80% del rendimiento puede tomar aproximadamente 2.750 horas.

### Diferentes requisitos de gestión

Las soluciones de Big Data acceden y generan datos que son convertidos en activos de la empresa. **Se requiere un framework de gestión para garantizar que los datos y el entorno de solución mismo están regulados, estandarizados y se desarrollan de manera controlada.**

Algunos ejemplos de lo que un framework de gestión de Big Data puede abarcar son:

- Estandarización sobre cómo se etiquetan los datos y sobre los metadata usados para el etiquetado
- Políticas que regulan el tipo de datos externos que se pueden adquirir
- Políticas para la protección de datos y conservar el anonimato de datos
- Políticas para el archivo de datos provenientes de fuentes de datos y resultados de análisis
- Políticas para la limpieza (cleansing) y filtrado (filtering) de datos

## Metodología diferencial

Es necesaria una metodología para **controlar cómo fluyen los datos hacia dentro y hacia afuera de las soluciones de Big Data y controlar cómo se pueden establecer los loop de retroalimentación**, para facilitar que los datos procesados sean sometidos a mejoras constantes, como se muestra en la Figura 1.51. Por ejemplo, se podría utilizar un enfoque iterativo para que el personal de la empresa haga retroalimentación para el personal de TI, de forma que se hagan mejoras constantes en el sistema. Cada ciclo de retroalimentación puede revelar la necesidad de modificar los pasos existentes o de crear nuevos pasos; por ejemplo, el preprocesamiento para la limpieza (cleansing) de los datos.

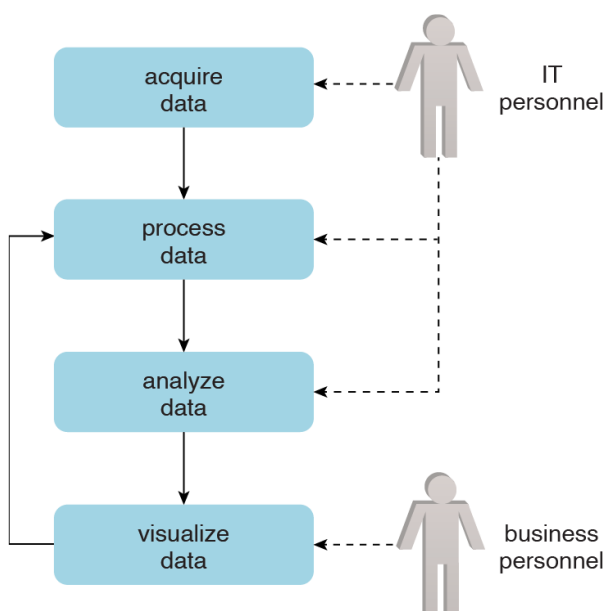


Figura 1.51 – Cada iteración puede ayudar a ajustar los pasos del procesamiento, algoritmos y modelos de datos para mejorar la precisión de los resultados y ofrecer un mayor valor para la empresa.

## Cloud Computing

Como se mencionó en la sección *Factores empresariales y tecnológicos de Big Data*, el Cloud Computing introduce entornos remotos que pueden hospedar una infraestructura de TI para almacenar y procesar datos a gran escala, entre otras cosas. Sin importar si una organización ya cuenta con información basada en la nube, la adopción de un entorno Big Data puede requerir que una parte o todo el entorno esté hospedado en la nube. Por ejemplo, una empresa que ejecuta su sistema de CRM en la nube toma la decisión de implementar una solución Big Data en el mismo entorno de nube, con el fin de ejecutar la analítica en sus datos de CRM. Luego, estos datos pueden ser compartidos con su entorno Big Data principal que se encuentra al interior de la empresa.

#### NOTA

El tema de Cloud Computing en relación con Big Data se explora con mayor profundidad en el *Módulo 2: Conceptos de análisis y tecnología de Big Data*.

Algunas de las justificaciones comunes para incorporar un entorno basado en nube para respaldar una solución de Big Data son:

- existen recursos inadecuados de hardware interno
- no se dispone de inversión inicial de capital
- el proyecto debe estar aislado del resto de la empresa para no afectar los procesos empresariales existentes
- la iniciativa de Big Data es un prototipo
- los datasets que requieren ser procesados están hospedados en la nube
- se ha llegado a los límites de los recursos de informática y almacenamiento disponibles usados por una solución interna de Big Data

#### Lecturas opcionales

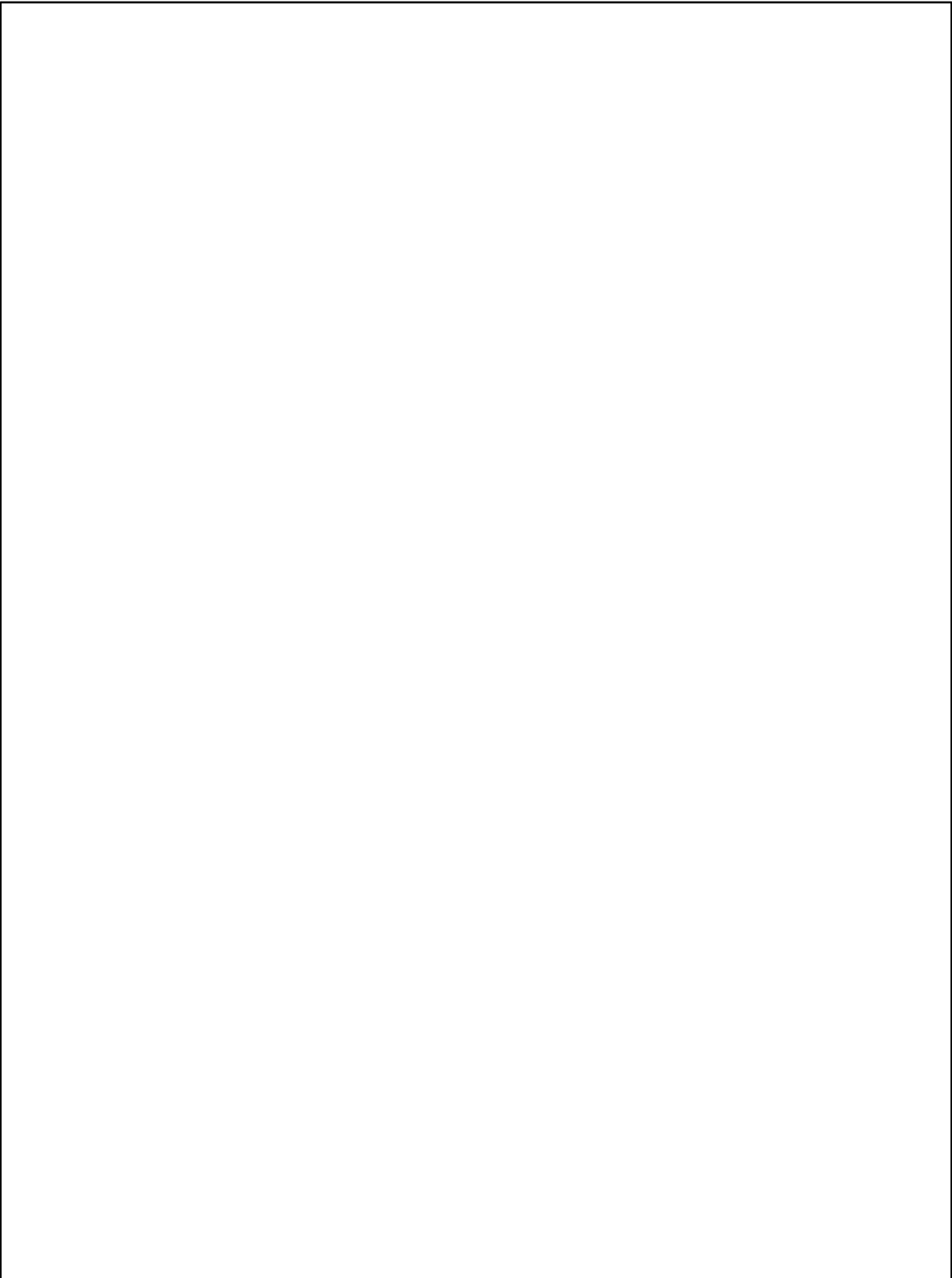
Las consideraciones relacionadas con la adopción y planeación de Big Data se discuten más a fondo en el capítulo 3 del libro *Analítica de Big Data*, incluido con el *Módulo 2: Conceptos de análisis y tecnología de Big Data* y en los capítulos 1, 2, 3, 5, 6 y 7 de la sección *Demasiado grande para ignorarlo: el caso empresarial de Big Data*.

[illegible]

[illegible]

[illegible]

## Notas / Bocetos





# Respuestas a los ejercicios

## Ejercicio 1.1: respuestas

1. Un **dataset** es un conjunto de datos relacionados, en el cual todos los miembros del grupo poseen el mismo conjunto de atributos.
2. El objetivo del **análisis de datos (Data Analysis)** es respaldar la toma de decisiones al establecer patrones y relaciones en los datos que son analizados.
3. La **analítica** se enfoca en filtrar grandes cantidades de datos sin **estructurar**, con el fin de extraer información significativa que pueda ser útil para enriquecer los datos empresariales actuales.
4. El proceso de la Inteligencia de negocios (BI) puede aplicar la **analítica** a grandes cantidades de datos.

## Ejercicio 1.2: respuestas

1. Una **bodega de datos digital (Data Warehouse)** puede contener bases de datos analíticas que pueden mejorar los tiempos de respuesta de las consultas.
2. **Extraer – transformar – cargar** es un proceso utilizado para cargar los datos desde un sistema origen hasta un sistema destino, y es la principal operación utilizada para enviar datos a las bodegas de datos digitales (Data Warehouse).
3. Las cinco principales características de Big Data que la diferencian de los datos tradicionales son volumen, velocidad, **variedad**, **valor** y **veracidad**.
4. La característica de valor de Big Data **depende** de qué tanto tiempo consuma el procesamiento de los datos.
5. En general, los datos que son procesados por las soluciones de Big Data pueden estar clasificados en los siguientes tipos de datos o formatos: **estructurados**, **sin estructurar**, **semiestructurados** y **metadata**.

## Ejercicio 1.3: respuestas

1. La analítica se puede categorizar en cuatro tipos diferentes de acuerdo con el valor de su atributo: **descriptiva**, **diagnóstica**, **predictiva** y **prescriptiva**.
2. Los resultados de la analítica **diagnóstica** se pueden observar por medio de herramientas interactivas de visualización que permiten identificar fácilmente tendencias y patrones.
3. La analítica **prescriptiva** tiene más valor para las empresas porque esta técnica recomienda un plan de acción que puede ser seguido.

4. Los algoritmos de aprendizaje automático (Machine Learning) se pueden categorizar en aprendizaje **supervisado** y **no supervisado**.

#### **Ejercicio 1.4: respuestas**

1. La Inteligencia de negocios (BI) tradicional utiliza analítica **prescriptiva** y **diagnóstica**.
2. La **visualización de datos** comunica resultados de analítica utilizando una variedad de herramientas gráficas e interactivas.
3. Las herramientas tradicionales de visualización de datos presentan los resultados tanto de la analítica **descriptiva** como de la **diagnóstica**.
4. La Inteligencia de negocios (BI) de Big Data añade valor a la Inteligencia de negocios (BI) tradicional al utilizar la analítica **predictiva** y **prescriptiva**.
5. Las herramientas avanzadas de visualización de Big Data utilizan herramientas de analítica de datos **prescriptiva** y **predictiva**.

## Examen B90.01

El curso que acaba de finalizar corresponde al Examen B90.01, que es un examen oficial del Programa Profesional Certificado de Ciencias de Big Data (BDSCP).

PEARSON VUE

Este examen se puede tomar en los Centros de Examinación de Pearson VUE en todo el mundo, o a través de Pearson VUE Proctoring Online, que le permite tomar exámenes desde su casa o estación de trabajo y contar con supervisión en vivo. Si desea obtener más información, visite las siguientes páginas web:

[www.bigdatascienceschool.com/exams/](http://www.bigdatascienceschool.com/exams/)

[www.pearsonvue.com/arcitura/](http://www.pearsonvue.com/arcitura/)

[www.pearsonvue.com/arcitura/op/](http://www.pearsonvue.com/arcitura/op/) (Online Proctoring)

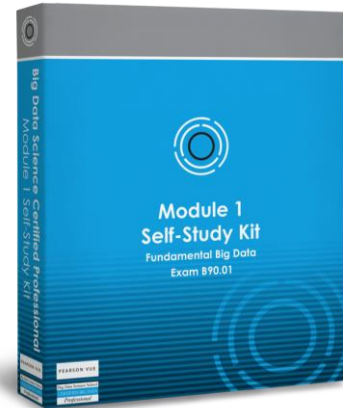
## Kit de autoaprendizaje del Módulo 1

Para este Módulo se encuentra disponible un kit oficial de autoaprendizaje de BDSCP, el cual le ofrece materiales y recursos de estudio adicionales, incluyendo una guía separada de autoaprendizaje, CD de audiotutoría y tarjetas de memorización.

Tenga en cuenta que las versiones de este kit de autoaprendizaje están disponibles con y sin un cupón para el Examen B90.01 de Pearson VUE.

Si desea obtener más información, visite la siguiente página web:

[www.bigdataselfstudy.com](http://www.bigdataselfstudy.com)



## Información y recursos de contacto

### Comunidad de AITCP

Únase a la creciente comunidad internacional de Profesionales Certificados en TI de Arcitura (Arcitura IT Certified Professional, AITCP), conectándose mediante las plataformas oficiales de social media: LinkedIn, Twitter, Facebook y YouTube.

Los links de social media y de la comunidad están disponibles en:

- [www.arcitura.com/community](http://www.arcitura.com/community)
- [www.servicetechbooks.com/community](http://www.servicetechbooks.com/community)



### Información general del programa

Si desea conocer información general acerca del programa de BDSCP y los requisitos de certificación, visite las siguientes páginas web:

[www.bigdatascienceschool.com](http://www.bigdatascienceschool.com) y [www.bigdatascienceschool.com/matrix/](http://www.bigdatascienceschool.com/matrix/)

### Información general acerca de los módulos del curso y los kits de autoaprendizaje

Si desea conocer información general acerca de los módulos del curso BDSCP y los kits de autoaprendizaje, visite las siguientes páginas web:

[www.bigdatascienceschool.com](http://www.bigdatascienceschool.com) y [www.bigdataselfstudy.com](http://www.bigdataselfstudy.com)

### Inquietudes acerca del examen de Pearson VUE

Si desea conocer información relacionada con la presentación de los exámenes del BDSCP en los centros de examen de Pearson VUE o mediante la supervisión online de Pearson VUE, visite las siguientes páginas web:

[www.pearsonvue.com/arcitura/](http://www.pearsonvue.com/arcitura/)  
[www.pearsonvue.com/arcitura/op/](http://www.pearsonvue.com/arcitura/op/) (Online Proctoring)

### Programación de talleres dirigidos al público y guiados por instructores

Si desea conocer la más reciente programación de los talleres del BDSCP guiados por instructores que están abiertos al público, visite la siguiente página web:

[www.bigdatascienceschool.com/workshops](http://www.bigdatascienceschool.com/workshops)

## **Talleres privados guiados por instructores**

Los entrenadores certificados pueden realizar los talleres directamente en sus instalaciones, con la opción de supervisión de exámenes en el sitio. Si desea saber más acerca de las opciones y tarifas, envíe un correo electrónico a la siguiente dirección:

[info@arcitura.com](mailto:info@arcitura.com)

o llame a la línea

1-800-579-6582

## **Convertirse en un entrenador certificado**

Si usted está interesado en alcanzar el rango de Entrenador Certificado para este o cualquier otro curso o programa de Arcitura, puede obtener más información visitando la siguiente página web:

[www.arcitura.com/trainerdevelopment/](http://www.arcitura.com/trainerdevelopment/)

## **Inquietudes generales sobre BDSCP**

En caso de que tenga otras preguntas relacionadas con este Curso, o cualquier otro Módulo, Examen o Certificación que haga parte del programa BDSCP, envíe un correo electrónico a la siguiente dirección:

[info@arcitura.com](mailto:info@arcitura.com)

o llame a la línea

1-800-579-6582

## **Notificaciones automáticas**

Si desea que se le notifique automáticamente sobre cambios o actualizaciones al programa BDSCP y a los sitios de recursos relacionados, envíe un mensaje de correo electrónico en blanco a la siguiente dirección:

[notify@arcitura.com](mailto:notify@arcitura.com)

## **Retroalimentación y comentarios**

Ayúdenos a mejorar este curso. Envíe su retroalimentación o comentarios a la dirección de correo electrónico:

[info@arcitura.com](mailto:info@arcitura.com)