Minería de Datos II

Clase 3 - La era de los Datos

Los sistemas informáticos utilizados hoy en el mundo empresarial, para automatizar los procesos de negocio, tuvieron sus orígenes en el área militar. Entre 1950 y 1960, después de la Segunda Guerra Mundial, Estados Unidos comenzó a promover el uso de programas informáticos para la administración de tareas y organización de su ejército.
De hecho, hasta la década de los 50, las instituciones militares eran prácticamente las únicas que tenían acceso a los primeros equipos informáticos. Estos sistemas de organización han sido declarados como los precursores del ERP (Enterprise Resource Planning).
A partir de 1960 salieron al mercado las primeras computadoras comerciales y se sentaron las bases de la gestión automatizada, herramientas de planificación, gestión de inventarios, etc.
Por supuesto eran sistemas “primitivos”, pero suponían un gran avance en aquella época.
La década de 1970 se vio marcada por la irrupción de los MRP ó Sistemas de Planificación y Programación de la Producción de la mano de IBM. Estos sistemas ayudaban a planificar los requerimientos de la materia prima que se utilizaba en la fabricación de artículos y productos.
En 1972 se fundó la empresa SAP en Alemania. Las iniciales de la empresa significaban “Sistemas, Aplicaciones y Productos”. El objetivo de SAP era crear software empresarial que funcionara en tiempo real. SAP lanzó su primer programa de contabilidad financiera en 1973.
En 1975, el software MRP ya se utilizaba en multitud de grandes empresas. El sistema funcionaba en enormes ordenadores centrales que eran muy caros, aunque su potencia de cálculo no era comparable ni siquiera a la de algunos de los ordenadores portátiles de hoy en día.
Al entrar en los 80 estos sistemas evolucionaron y se denominaron MRP- II. Incluían un avance importante: ya no solo se encargaban de la gestión de materiales, sino que también administraban recursos económicos. De esta forma, incluyeron elementos financieros presentes en la producción como los costes de la materia prima, mano de obra y organización.
El término ERP (Enterprise Resource Planning) se utilizó por primera vez en la década de 1990. Se centraba en funciones empresariales como la producción, finanzas y contabilidad, recursos humanos, gestión de proyectos, etc. Estaba compuesto por sistemas modulares con características avanzadas.
Con el comienzo del nuevo siglo los ERP se popularizaron. A partir del año 2000 Gartner Group proporcionó funcionalidades como la gestión de la cadena de suministro, la gestión de las relaciones con los clientes (CRM) y la inteligencia de negocio.
A partir del 2005 la tendencia se ha orientado hacia soluciones de software en la nube y se ha alejado de los modelos tradicionales de instalación en los servidores del cliente. Las soluciones de software Cloud ERP proporcionaron funcionalidades comparables a las del ERP local a un coste mucho menor. Desde ese momento se produce un punto de inflexión en el que ya no solo son utilizados por grandes empresas manufactureras sino que se “democratizan” extendiendo su uso a las PYMES.
En consecuencia, se viene registrando datos desde hace décadas. Almacenándose en diferentes formatos. Al principio, se trató de “tablas libres”, los cuales eran archivos binarios

alojados en el sistema de archivos. Así era como se generaba y guardaba el dato, hasta que a principios de la década del setenta en los laboratorios de IBM, se creó el nuevo software de base de datos System R. Y para gestionar los datos almacenados en System R, se creó el lenguaje SQL. En un principio se llamó SEQUEL, un nombre que todavía se utiliza como una pronunciación alternativa para SQL, pero más tarde fue renombrado a sólo SQL.
Hasta antes del SQL, para realizar una consulta a los datos, era necesario desarrollar un programa con el mismo lenguaje en que había sido creado, por ejemplo Cobol o Fortran, lo cual suponía una tarea que consumía mucho tiempo.
En 1979, una compañía llamada Relational Software, que luego se convirtió en Oracle, vio el potencial comercial del lenguaje SQL y lanzó su propia versión modificada, denominada Oracle V2.
Cuando hablamos de Big Data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.
Aunque el tamaño utilizado para determinar si un conjunto de datos determinado se considera Big Data no está firmemente definido y sigue cambiando con el tiempo, la mayoría de los analistas y profesionales actualmente se refieren a conjuntos de datos que van desde 30-50 Terabytes a varios Petabytes.
La naturaleza compleja del Big Data se debe principalmente a la naturaleza no estructurada de gran parte de los datos generados por las tecnologías modernas, como los logs de servidores web, los sensores incorporados en dispositivos, las búsquedas en Internet, las redes sociales, computadoras portátiles, teléfonos inteligentes y otros teléfonos móviles, dispositivos GPS, registros de centros de llamadas y en general todo lo que refiere a tecnEn la mayoría de los casos, con el fin de utilizar eficazmente el Big Data, debe combinarse con datos estructurados (normalmente de una base de datos relacional) de una aplicación comercial más convencional, como un ERP (Enterprise Resource Planning) o un CRM (Customer Relationship Management).
Lo que hace que Big Data sea tan útil para muchas empresas es el hecho de que proporciona respuestas a muchas preguntas que las empresas ni siquiera sabían que tenían. En otras palabras, proporciona un punto de referencia. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la empresa considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible.
La recopilación de grandes cantidades de datos y la búsqueda de tendencias dentro de los datos permiten que las empresas se muevan mucho más rápidamente, sin problemas y de manera eficiente. También les permite eliminar las áreas problemáticas antes de que los problemas acaben con sus beneficios o su reputación.ología multimedia, es decir, imágenes, audio y video.
El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más felices. Las empresas con más éxito con Big Data consiguen valor de las siguientes formas:

• Reducción de coste. Las grandes tecnologías de datos, como Hadoop y el análisis basado en la nube, aportan importantes ventajas en términos de costes cuando se trata de almacenar grandes cantidades de datos, además de identificar maneras más eficientes de hacer negocios.

• Más rápido, mejor toma de decisiones. Con la velocidad de Hadoop y la analítica en memoria, combinada con la capacidad de analizar nuevas fuentes de datos, las empresas pueden analizar la información inmediatamente y tomar decisiones basadas en lo que han aprendido.

• Nuevos productos y servicios. Con la capacidad de medir las necesidades de los clientes y la satisfacción a través de análisis viene el poder de dar a los clientes lo que quieren. Con la analítica de Big Data, más empresas están creando nuevos productos para satisfacer las necesidades de los clientes.

-Big Data

Es un término que hace referencia a una nueva clase de datos que no pueden ser gestionados por sistemas tradicionales.
3V's de Big Data:

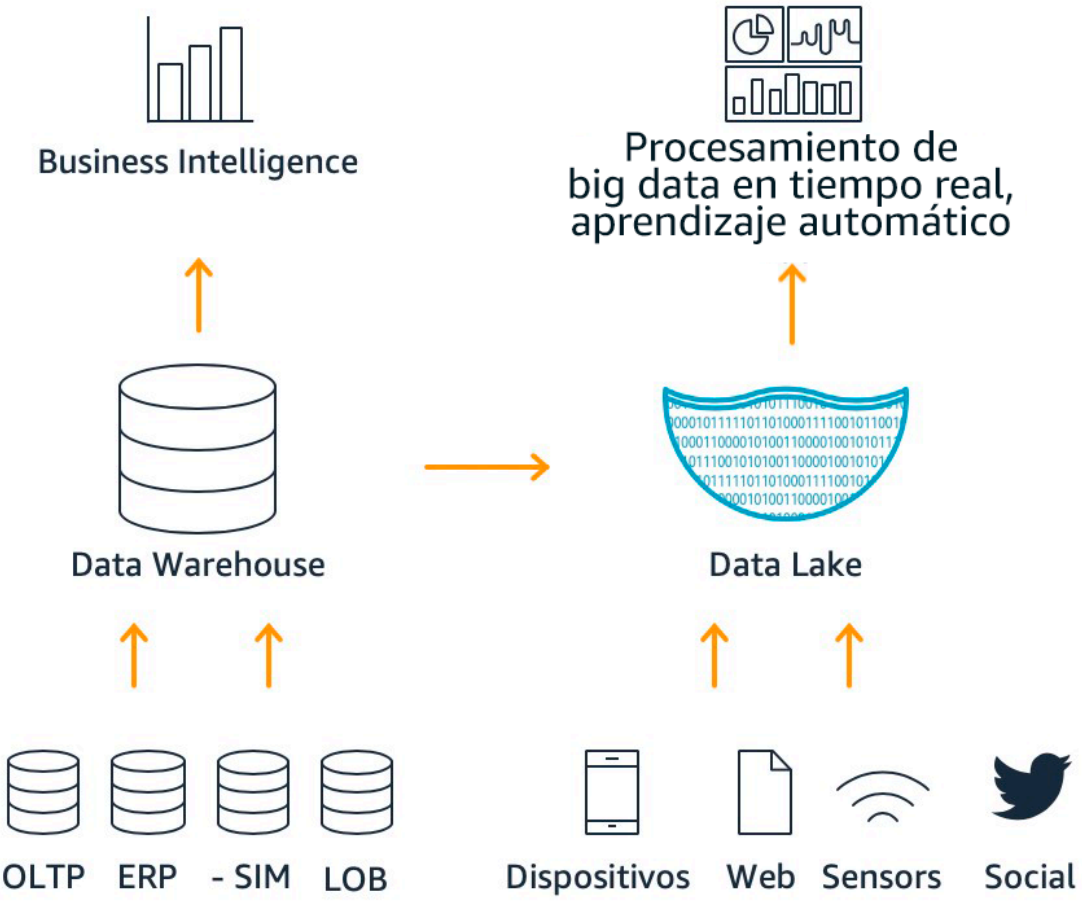
- Volumen.
- Variedad.
- Velocidad.

Las especiales características del Big Data hacen que su calidad de datos se enfrente a múltiples desafíos. Se trata de las conocidas como 3 Vs: Volumen, Velocidad y Variedad que definen la problemática del Big Data.
Estas características del Big Data provocan que las empresas tengan problemas para extraer datos reales y de alta calidad, de conjuntos de datos tan masivos, cambiantes y complicados.

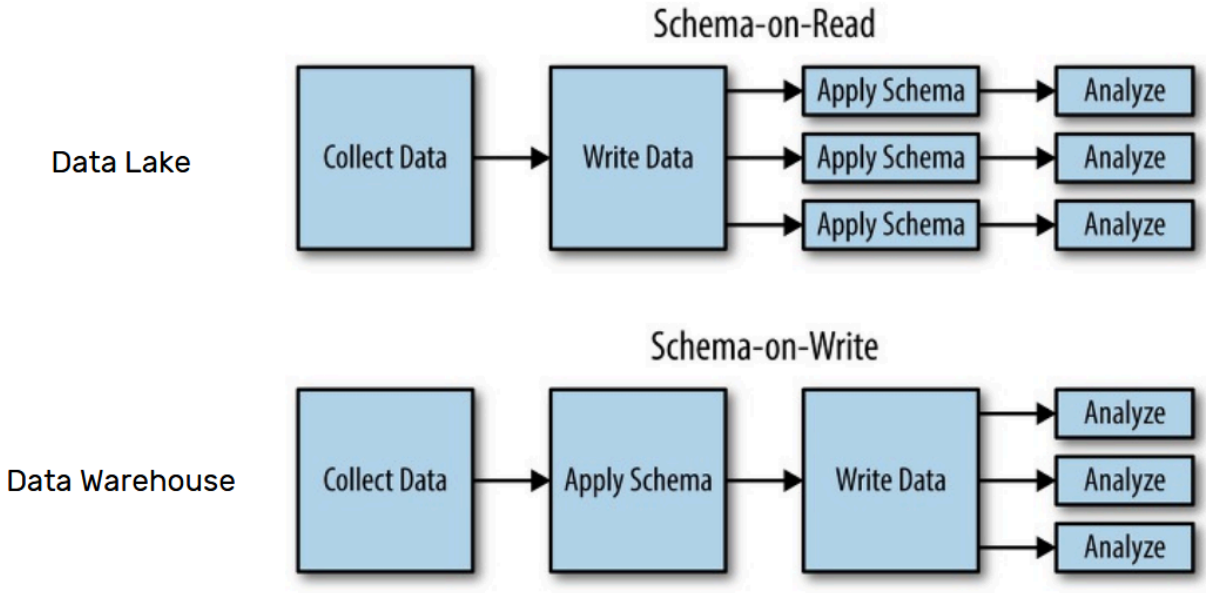
-Casos de Uso

-Data Lake

Es un repositorio unificado de datos, estructurados y no estructurados.
Está diseñado para soportar las cargas de trabajo de Big Data y Machine Learning.
Prioriza el almacenamiento de los datos en su formato original para luego ser procesados de acuerdo a la demanda.



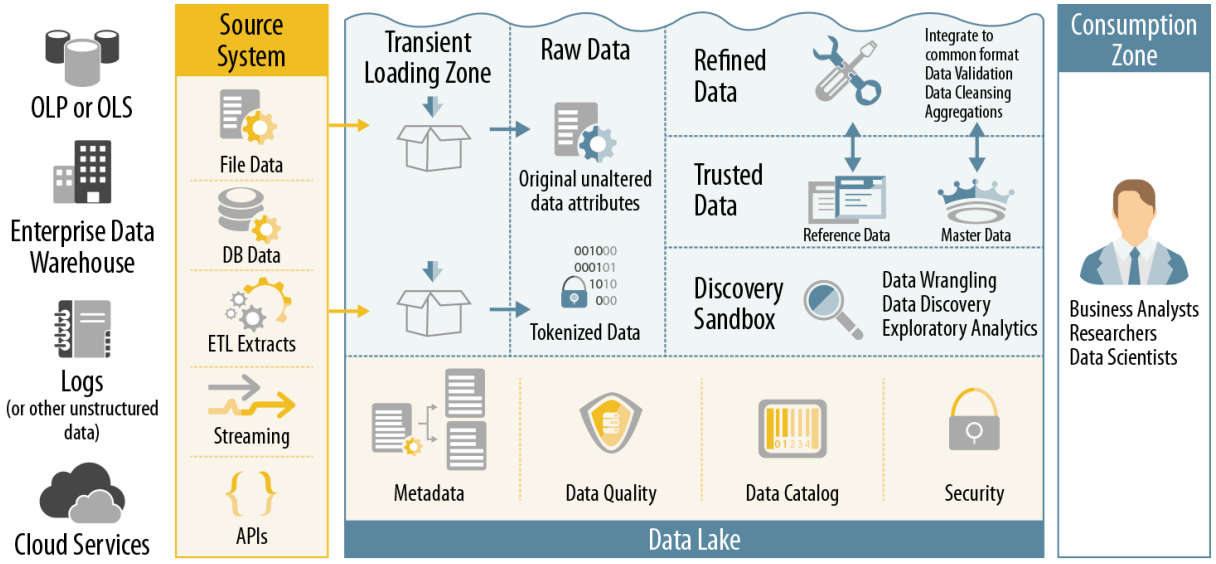
-Estrategias de Procesamiento:



Dada la afluencia en los últimos tiempos, del uso de Internet y la tecnología de redes en general, como por ejemplo sensores o APIs, concretamente el desarrollo de lo que se conoce como IoT (Internet de las Cosas), se comenzó a trabajar datos que no

necesariamente son llevados a una estructura tabular, dentro de un Data Warehouse y se manejan por fuera del mismo, dando lugar al desarrollo de una serie de herramientas conocidas como motores de bases de datos No-SQL y al desarrollo de una arquitectura conocida como Data Lake, la cuál contempla el almacenamiento y disponibilización de todo tipo de datos, estructurados y no estructurados, manejando esa variedad y también soportando grandes volúmenes de datos, que también se genera a gran velocidad. Éstas tres características, son conocidas como las 3 V del Big Data, y una diferencias notoria respecto de un proceso de ETL tradicional, es que esas fases se reordenan, dando lugar a un concepto conocido como ELT, donde primero se realiza la extracción al igual que en un ETL, pero luego se hace la carga de los datos, sin necesariamente pasar por un proceso de transformación, proceso que llega luego bajo la necesidad de analizar ese dato. Por eso se define que el Data Warehouse consiste en un esquema “On Write” y el Data Lake en un esquema “On Read”, en este último, se almacenan todos los datos que se generan, aún si todavía se desconoce si luego no va a utilizarse.

-Arquitectura del Data Lake



Laboratorio:

1) Realizar la instalación de VirtualBox: https://www.virtualbox.org/wiki/Downloads

2) Realizar la instalación de Putty: https://www.putty.org/ - https://www.compuhoy.com/como-descargo-putty-en-linux/

3) Realizar la instalación de WinSCP: https://winscp.net/eng/download.php (FileZilla es una alternativa si no usas sistema operativo Windows)

4) En el archivo "Servidor_Ubuntu.zip" hay disponible un archivo VDI necesario para crear una máquina virtual Linux en VirtualBox. Esta máquina virtual es un servidor Ubuntu:
usuario: ubuntu
contraseña: ubuntu

Adicionalmente:

Para realizar la instalación de Docker:

1) Instalación de Docker en tu sistema operativo: https://hub.docker.com/

Si el sistema operativo usado es Linux: https://docs.docker.com/engine/install/

* sudo apt install -y docker-compose

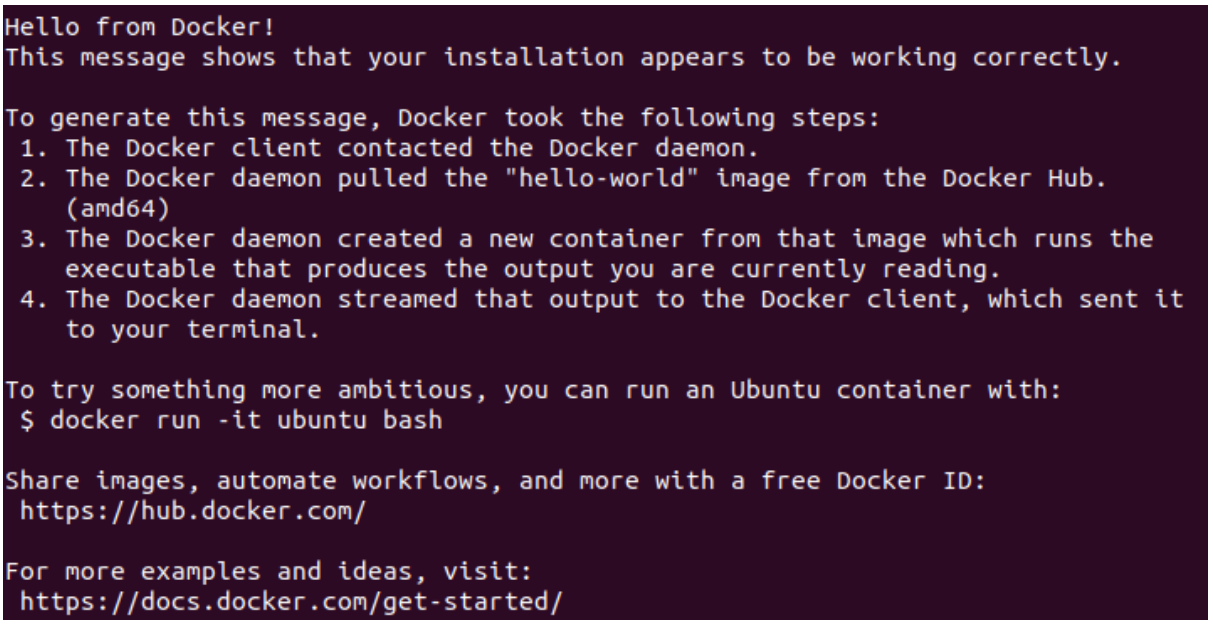
2) Comenzar a familiarizarse con los comandos de Linux:

* Tutorial:https://www.tutorialspoint.com/unix_commands/index.htm

* Interactivo:
https://cli-boot.camp/?id=1dbj970vv4n

3) Una vez que tenemos Docker instalado, habiendo escogido cualquiera de las opciones ya podemos crear el cluster Hadoop siguiendo las siguientes instrucciones:

1-Probemos la instalación de Docker: “sudo docker run hello-world”



4) Descarguemos desde GitHub lo necesario para ejecutar el cluster:

“git clone https://github.com/dariomiguellopez/curso_big_data_modulo1”
“cd curso_big_data_modulo1”

Revisemos el archivo “start-container.sh”, los archivos dentro de la carpeta “config” y la URL “localhost:8088”.

Finalmente sigamos las instrucciones propuestas para ver en funcionamiento HDFS y MapReduce sobre Hadoop.