Minería de Datos II

Clase 1 - Ingeniería de Datos

-¿Qué es la Ingeniería de Datos?

Es un campo de la Ciencia de Datos enfocado principalmente en el diseño, construcción y mantenimiento de sistemas y arquitecturas para la gestión eficiente, procesamiento y análisis de grandes volúmenes de datos. Esto incluye tareas como la extracción, transformación y carga de datos (ETL), el modelado de datos, la limpieza de datos, la integración de datos, la administración de bases de datos, entre otras.

Al igual que un ingeniero se encarga de diseñar y construir estructuras físicas como puentes o edificios, un ingeniero de datos diseña y construye sistemas y procesos para manejar y gestionar datos de manera eficiente y efectiva. Con la aplicación de metodologías sistemáticas, herramientas y técnicas para resolver problemas relacionados con el manejo y análisis de datos.

La ingeniería de datos se basa en la planificación, diseño, implementación y optimización de sistemas y procesos para garantizar que los datos estén disponibles, sean accesibles y sean utilizables para diferentes propósitos, como análisis, toma de decisiones y desarrollo de aplicaciones.

-Data Governance

Conjunto de procesos, políticas, estándares y controles que aseguran que los datos de una organización sean manejados de manera adecuada, con calidad, seguridad y cumplimiento normativo. Es esencialmente el marco de trabajo que establece quién es responsable de qué datos, cómo se deben utilizar, almacenar, compartir, proteger y mantener.

Calidad de los datos: Ayuda a garantizar que los datos sean precisos, consistentes, completos y confiables, lo que mejora la confianza en la información y facilita la toma de decisiones.

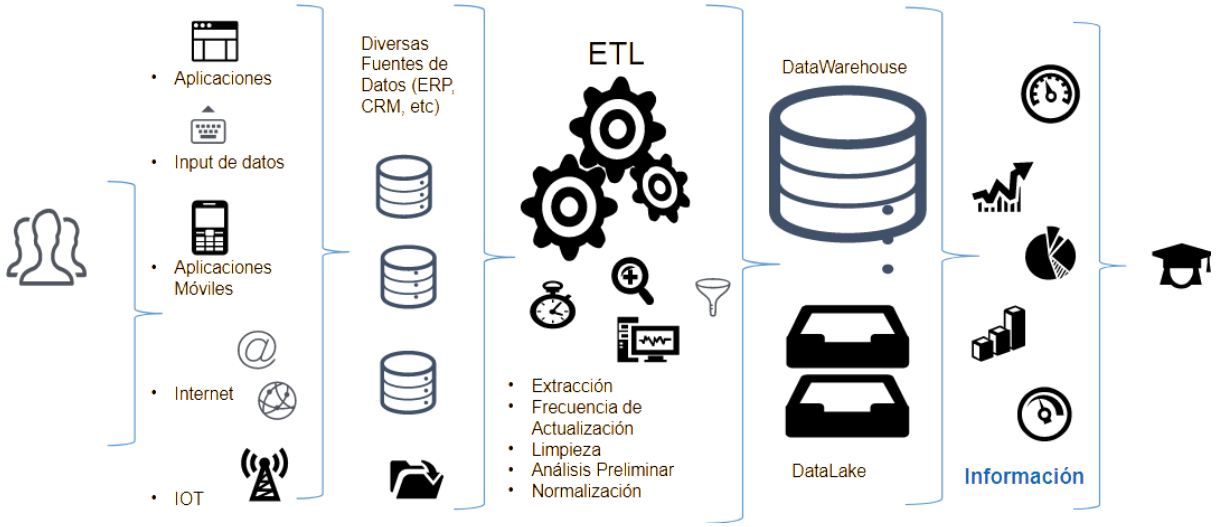
Seguridad y privacidad: Permite proteger los datos sensibles de la organización, evitando filtraciones o accesos no autorizados, lo que reduce el riesgo de pérdida de información y posibles sanciones legales.

Cumplimiento normativo: Ayuda a asegurar que la organización cumpla con las regulaciones y normativas aplicables en cuanto a la privacidad de los datos, como el RGPD en la Unión Europea o HIPAA en Estados Unidos.

Eficiencia y productividad: Facilita el acceso a los datos adecuados en el momento oportuno, lo que agiliza los procesos de análisis, reporting y desarrollo de aplicaciones.

Toma de decisiones fundamentada: Proporciona una visión clara de los datos disponibles, su procedencia y su calidad, lo que permite a los líderes empresariales tomar decisiones informadas y estratégicas.

-Ciclo de Vida del Dato



En una organización, normalmente se trabaja con diversos sistemas transaccionales, es decir, sistemas que se utilizan para la operatoria diaria, que son el punto de ingreso de los datos, a través de clientes, operadores, administrativos, usuarios en general, pero también se generan datos a partir de dispositivos de medición, como sensores o mecanismos de log, por ejemplo un servidor web dedicado a un portal de venta de productos, que en cada click deja un log con los datos del producto y usuario que lo visitó.

Esto trae en consecuencia que en una primera instancia los datos no están en un único repositorio, ni tienen necesariamente una estructura apta para ser almacenados en una base de datos relacional (RDBMS), Por tanto, al momento de querer extraer datos para su posterior análisis, nos encontramos con el problema de que debemos acceder a varias fuentes, unificarlo en un repositorio común, y luego realizar las consultas requeridas.

En este proceso hay una situación muy importante, y tiene que ver con que el hecho de tener que unificar datos, de distintas fuentes y darle una estructura adecuada, deja visibles inconvenientes de incongruencia, incompletitud y falencias varias en los datos, producto de que desde un principio, no se diseñó de manera integral el camino que transcurre el dato desde que se origina hasta que es utilizado y se transforma en información y conocimiento consecuentemente, este camino del dato es conocido como “Ciclo de Vida del Dato”, y es un concepto al que recurriremos frecuentemente.

El punto a cuidar aquí, tiene que ver entonces con el grado de calidad que podemos garantizar que la información que se disponibiliza va a tener, por lo tanto, es importante poder tener una apreciación clara acerca de la forma de cuantificar el grado de calidad, para lo cual, en primer lugar, es necesario tener en cuenta que los datos de la realidad pueden ser impuros debido a incompletitud, ruido o inconsistencias, sin embargo, debemos asegurarnos la fiabilidad del dato, teniendo el camino del dato, es decir, la trazabilidad.

Es una tarea muy importante debido a que la calidad de datos confiabiliza en análisis. Si utilizamos datos no confiables podemos llegar a conclusiones erróneas. Este proceso nos ayuda a recuperar información perdido o incompleta y resolver conflictos en los datos. Así

mismo, las fuentes son críticas en el proceso de selección de datos. Por último, es necesario armar métodos de control, auditoría, corrección y confiabilización de datos.

-Preparación del Dato

La etapa de preparación de datos llega a insumir el 90% del trabajo de una tarea de análisis de datos.
Sin la preparación de datos no es posible realizar análisis veraces.
Las técnicas de preparación son obligatorias para la incorporación de nuevos datos a nuestras bases, las mismas son: integración, limpieza del ruido, transformación e imputación de valores faltantes.

-Integración de Datos

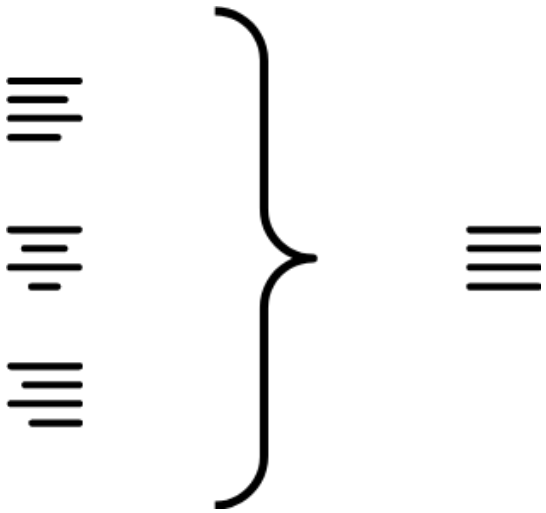
Es el primer paso e implica integrar datos de diferentes fuentes e incorporar los datos a formatos tabulares si no lo estuvieran.

-Limpieza de Datos

Proceso por el cual se discriminan los datos importantes de los accesorios y a la vez los veraces de los erróneos. Luego se procede a desestimar o borrar aquellos datos que no serán utilizados y a validar los que se conservarán.

-Normalización de Datos

En este paso se identifican aquellos valores iguales pero con notaciones diferentes y se los reescribe de una manera uniforme.
Ejemplo: Calle S Martín, Calle Gral. San Martín, Calle José de San Martín…. = Calle Gral. José de San Martín.

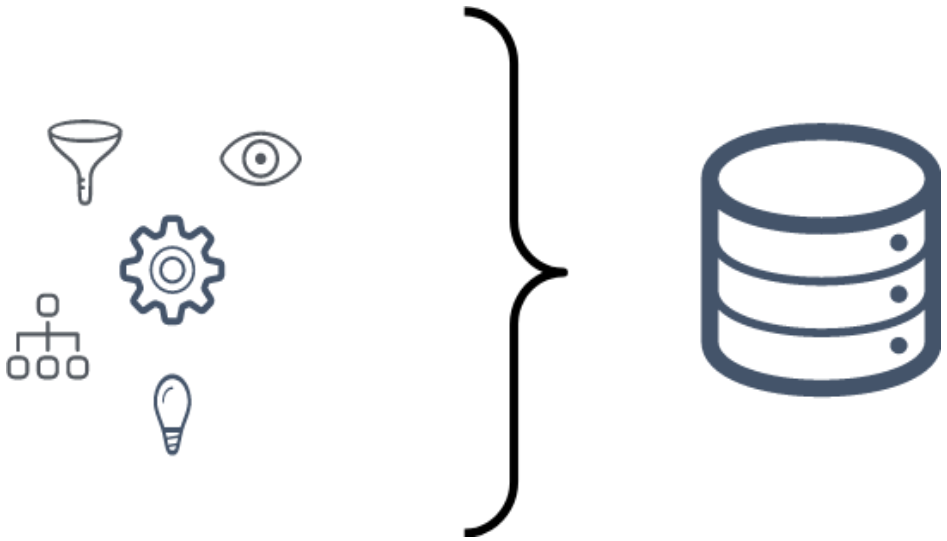


-Imputación de Valores Faltantes

La imputación de valores faltantes está relacionada con la identificación de los casos en donde exista pérdida o ausencia de datos. En estos casos será importante realizar acciones de reconstrucción, aunque no siempre es posible resolverlas.

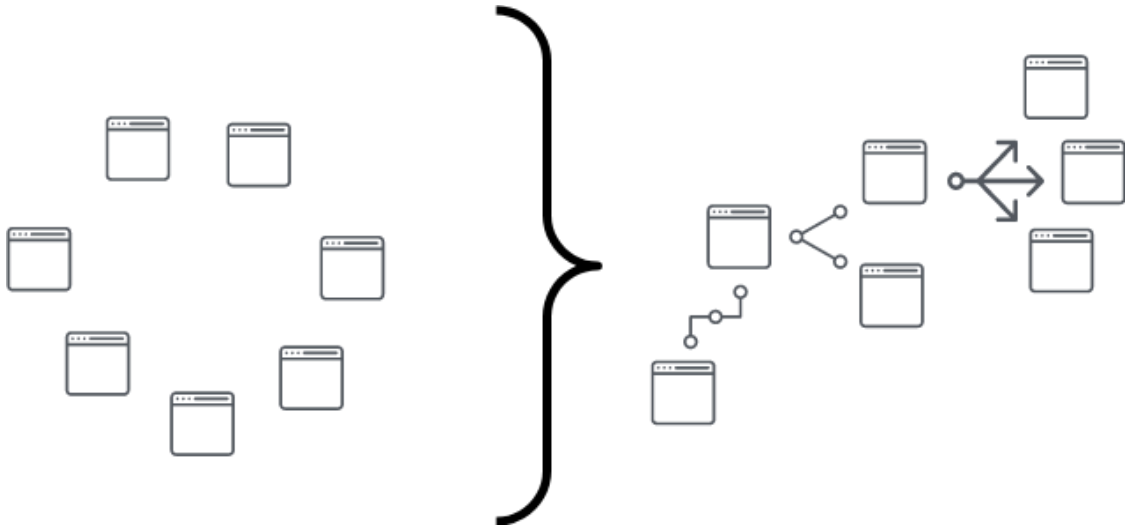
-Transformación de Datos

Este es el momento en que la información nueva será convertida a un formato compatible con base de datos. En este punto es crítico reducir el efecto distorsivo de la transformación.



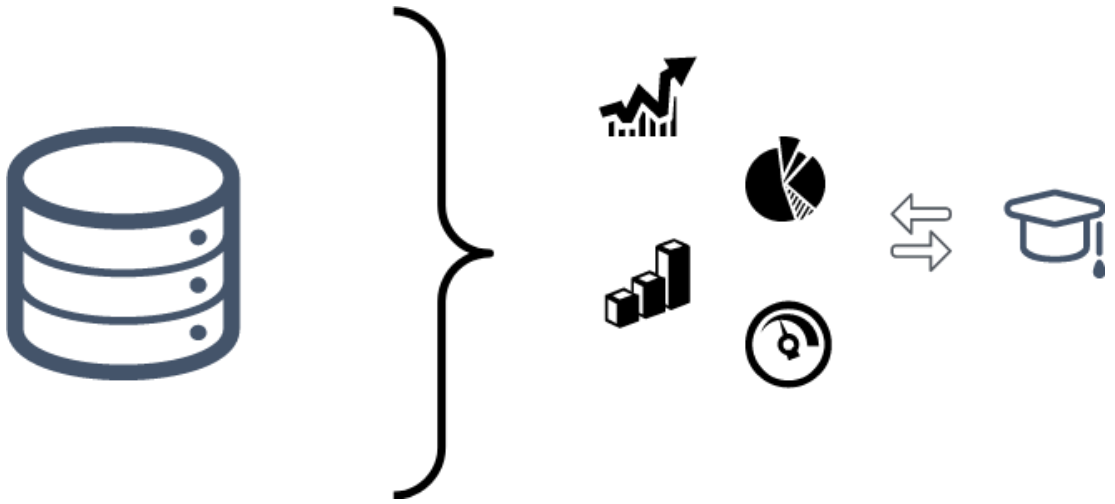
-Modelado de Datos

Se establecen las relaciones entre los datos, se determinan qué entidades contienen datos maestros y cuáles contienen hechos, dando lugar a las dimensiones y las métricas.



-Reportes y Visualización

Se genera Información a partir de los datos, esa información queda disponible para su análisis, que da lugar al Conocimiento.



Proyecto Integrador:

Parte 1

A lo largo de la materia Ustedes serán los analistas de datos de una compañía de venta de insumos tecnológicos al público. A lo largo de las prácticas se harán cargo de la información de la empresa y realizarán el proceso completo de captura, limpieza, análisis, diagnóstico, documentación, explotación y publicación de resultados.

La Dirección de Ventas ha solicitado las siguientes tablas a Marketing con el fin de que sean integradas:

* La tabla de puntos de venta propios, un Excel frecuentemente utilizado para contactar a cada sucursal, actualizada en 2021.
* La tabla de empleados, un Excel mantenido por el personal administrativo de RRHH.
* La tabla de proveedores, un Excel mantenido por un analista de otra dirección que ya no está en la empresa.
* La tabla de clientes, alojada en el CRM de la empresa.
* La tabla de productos, un Excel mantenido por otro analista.
* Las tablas de ventas, gastos y compras, tres archivos CSV generados a partir del sistema transaccional de la empresa.

Es necesario realizar la captura de esos archivos e ingestarlos dentro de nuestra base de datos.

Sugerencia:

Instalación MySQL y Wokrbench:
MySQL Server: https://dev.mysql.com/downloads/mysql/
MySQL Installer: https://dev.mysql.com/downloads/installer/
Workbench: https://dev.mysql.com/downloads/workbench/
MySQL: https://dev.mysql.com/doc/

Instrucción LOAD DATA

LOAD DATA sirve para tomar cualquier archivo CSV y cargarlo dentro de una tabla.
La sintaxis básica es:

```sql
LOAD DATA LOCAL INFILE '/importfile.csv'
INTO TABLE test_table
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
(field1, filed2, field3);
```

LOAD DATA indica que debe cargar un archivo csv (cualquier hoja de cálculo puede generar este tipo de archivos); si pasamos la opción LOCAL indicamos que el archivo está en nuestra máquina y que debe ser leído por el cliente y enviado al servidor; sino, la ruta (absoluta o relativa) es en el servidor.

Si dentro de la tabla hay registros, la violación de claves primarias podría causar la detención de la carga, entonces escribimos las opciones IGNORE (ignora las filas que violen el constraint y no las inserta) o REPLACE (agrega las filas reemplazando las existentes).

```sql
LOAD DATA LOCAL INFILE ‘ruta_archivo’
REPLACE INTO TABLE ‘nombre de la tabla’
```

Es necesario indicarle a MySQL cual es el separador de campos con "FIELDS TERMINATED BY ‘;’"
Por lo general se usa coma (,) aunque también pueden aparecer el punto y coma (;) ó el pipe (|)
También se puede indicar con que están encerradas las cadenas, si comillas simples, dobles, numerales (#), acentos viriguardilla (~) con "FIELDS ENCLOSED BY ‘»‘"
Se pueden agregar OPTIONALLY para indicar que algunos campos están encerrados con comillas, pero no todos con "FIELDS OPTIONALLY ENCLOSED BY ‘#’"
Para iniciar una línea, es posible usar "LINES STARTING BY ‘»’" (indica que las líneas empiezan en una cadena vacía)
Para terminar una línea, es posible usar "LINES TERMINATED BY ‘\n’", indicando que la línea termina con un salto de línea (\n).
Si en el archivo CSV hay una o más filas que representan los encabezados de los campos y desean «obviarlas» entonces es posible usar "IGNORE 1 LINES".

Para indicar su orden de guardado en la tabla; ejemplo, tenemos una tabla:
nombre, apellido, cédula, fecha_nacimiento
pero en el csv los campos vienen cedula, nombre, apellido, fecha_nacimiento, entonces colocamos entre paréntesis los campos asi:
(cedula, nombre, apellido, fecha_nacimiento)

Primera columna de mi csv corresponde al campo cedula, segunda al nombre y así sucesivamente.

Si acaso el archivo está guardado en ANSI, entonces se puede pasar opcionalmente el charset en el que está el archivo en la sentencia "CHARACTER SET latin1":
Si acaso el archivo está guardado en UTF8, entonces se puede pasar opcionalmente el charset en el que está el archivo en la sentencia "CHARACTER SET utf8":

```sql
LOAD DATA LOCAL INFILE ‘ruta_archivo’
CHARACTER SET latin1
INTO TABLE ‘nombre de la tabla’
```

* charset: Define el juego de caracteres con el que MySQL guardará los datos de forma interna.
* collation: Define la forma en el que MySQL buscará y ordenará los datos.

Un ejemplo de sintaxis completa queda:

```sql
LOAD DATA LOCAL INFILE ‘archivo’
CHARACTER SET
IGNORE
FIELDS TERMINATED BY ‘;’ ENCLOSED BY ‘»‘
LINES STARTING BY » TERMINATES BY »
IGNORE 1 LINES
(field1, field2, field3, @field4)
```sql

Parte 2

Con el objetivo de asegurarse de que la calidad de la información con la que se va a trabajar sea la óptima, es necesario realizar una lista de propuestas de mejora teniendo en cuenta los siguientes puntos:

1) ¿Qué tan actualizada está la información? ¿La forma en que se actualiza y mantiene esa información se puede mejorar?
2) ¿Los datos están completos en todas las tablas?
3) ¿Se conocen las fuentes de los datos?
4) Al integrar éstos datos, es prudente que haya una normalización respecto de nombrar las tablas y sus campos.
5) Es importante revisar la consistencia de los datos:
   - ¿Se pueden relacionar todas las tablas al modelo?
   - ¿Cuáles son las tablas de hechos y las tablas dimensionales o maestros?
   - ¿Podemos hacer esa separación en los datos que tenemos (tablas de hecho y dimensiones)?
   - ¿Hay claves duplicadas?

- ¿Cuáles son variables cualitativas y cuáles son cuantitativas?
- ¿Qué acciones podemos aplicar sobre las mismas?

Limpieza, Valores faltantes

6) Normalizar los nombres de los campos y colocar el tipo de dato adecuado para cada uno en cada una de las tablas. Descartar columnas que consideres que no tienen relevancia.
7) Buscar valores faltantes y campos inconsistentes en las tablas sucursal, proveedor, empleado y cliente. De encontrarlos, deberás corregirlos o desestimarlos. Propone y realiza una acción correctiva sobre ese problema.
8) Utilizar la funcion provista 'UC_Words' (Homework_Utiles.sql) para modificar a letra capital los campos que contengan descripciones para todas las tablas.
9) Chequear la consistencia de los campos precio y cantidad de la tabla de ventas.
10) Chequear que no haya claves duplicadas, y de encontrarla en alguna de las tablas, proponer una solución.

### Normalización

10) Generar dos nuevas tablas a partir de la tabla 'empleado' que contengan las entidades Cargo y Sector.
11) Generar una nueva tabla a partir de la tabla 'producto' que contenga la entidad Tipo de Producto.
12) Utilizar la función provista 'UC_Words' (Homework_Utiles.sql) para modificar a letra capital los campos que contengan descripciones para todas las tablas.
13) Utilizar el procedimiento provisto 'Llenar_Calendario' (Homework_Utiles.sql) para poblar la tabla de calendario.

Sugerencia:

Instrucción INSERT:

Es posible usarla a partir del resultado de otra consulta. Por ejemplo:

```SQL
INSERT INTO cargo (Cargo)
SELECT DISTINCT Cargo
FROM empleado
ORDER BY Cargo;
```

Instrucción UPDATE:

Es posible usarla a partir del resultado del resultado de una consulta de la tabla a modificar y otra/s tabla/s. Por ejemplo:

```SQL
UPDATE empleado e JOIN cargo c
```

```
    ON (c.Cargo = e.Cargo)
SET e.IdCargo = c.IdCargo;
```