**Equipo 30**

**A01793999 - Nathalia Prada**

**A01793933 - Christian Suárez Gil**

1. Fundamentos de bases de datos y para ciencia de datos.

Las bases de datos son el medio a través del cual se almacenan, y se da disponibilidad para su consulta, conjuntos de datos estructurados. Normalmente, cuentan con información relacionada de manera lógica y con una administración central y estandarizada de cada una de sus tablas, para poder conectarlas en un ámbito de negocio. Las bases de datos son de gran importancia para la ciencia de datos, dado que suelen ser la fuente más usada de consulta de información en la mayoría de las organizaciones.

Una base de datos está definida un conjunto estructurado de datos guardados en la memoria del computador o en la nube que es accesible de varias maneras.

Algunas de las bases de datos más comunes son las bases de datos relacionales, Las bases de datos relacionales se basan en el modelo relacional, una forma intuitiva y directa de representar datos en tablas. En una base de datos relacional, cada fila en una tabla es un registro con una ID única, llamada clave. Las columnas de la tabla contienen los atributos de los datos y cada registro suele tener un valor para cada atributo, lo que simplifica la creación de relaciones entre los puntos de datos.

También existen las bases de datos no relacionales de sus principales características es que no trabajan con estructuras definidas. Es decir, los datos no se almacenan en tablas, y la información tampoco se organiza en registros o campos. Tienen una gran escalabilidad y están pensadas para la gestión de grandes volúmenes de datos.

ETL vs ELT: tradicionalmente las bases de datos se gestión a través de la extracción transformación y cargue de información, sin embargo, con las nuevas capacidades de cómputo de las bases de datos y la nube grandes cantidades de datos se están cargando para luego transformadas lo que cambia el paradigma tradicional de ETL a ELT (donde el cargue ocurre antes de la transformación).

1. Fundamentos de almacenes de datos (Data Warehouse) para ciencia de datos.

Los Data Warehouse o almacenes de datos son grandes repositorios que consolidan e integran información de las diferentes fuentes y bases de datos de una compañía, con el objetivo particular de facilitar su extracción para procesos de Visualización y Ciencia de datos. Estos se encuentran optimizados para hacer más eficientes los procesos de uso de la información y no solo su almacenamiento. Cobran relevancia para la Ciencia de datos, ya que es el tipo de infraestructura al cuál están migrando las grandes empresas, al notar que las bases de datos tradicionales no son suficientes a medida que crecen los volúmenes y fuentes de información.

El esquema de estrella es común mente usado en los almacenes de datos, en esta técnica de modelamiento dimensional los datos se dividen en dimensiones y hechos.

Un registro en una tabla de “hechos” (facts) contiene “metadata” de una entidad, así como “medidas” (valores numéricos) que se quieren rastrear y luego resumir. Una dimensión es una propiedad de la entidad por la que puede agrupar o "cortar y seleccionar" los registros de hechos, y una tabla de dimensiones contendrá más información de esa propiedad.

Ejemplo:

Un registro transaccional de una tienda es un “hecho” (fact) que contiene los datos característicos de la compra: fecha de compra, numero de orden, cantidad pagada, y número de cliente. Y la tabla de dimensión asociada de la tienda contendría información adicional de la tienda, como su nombre. Con la tabla de dimensión tienda se puede se puede obtener un resumen de compras por tienda, por ejemplo.

El esquema de estrella es muy usado en la visualización de datos como por ejemplo en Power BI.