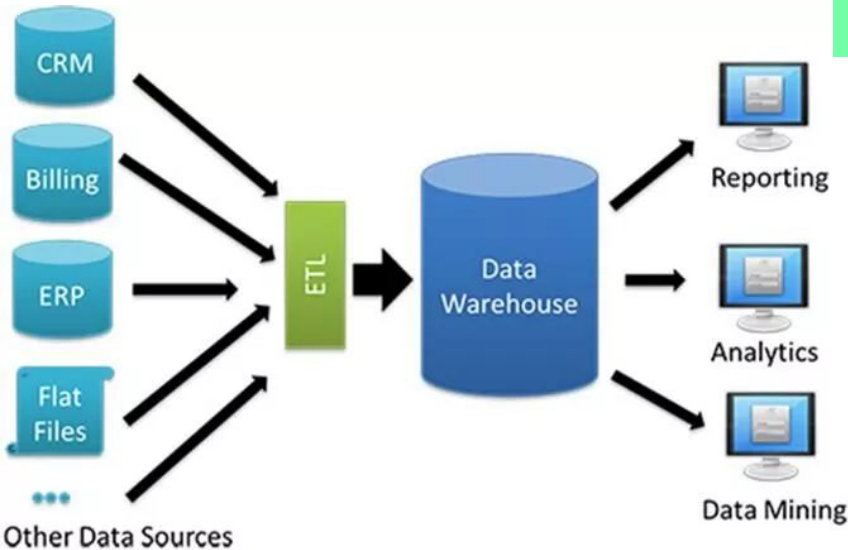





Data Warehouse

Data warehouse

Es un depósito central de información que se puede analizar para tomar decisiones más informadas. Los datos fluyen hacia un almacén de datos desde sistemas transaccionales, bases de datos relacionales y otras fuentes, normalmente con una cadencia regular.






¿Cómo funciona
el Data
Warehouse?

Data warehouse - ¿cómo funciona?



- El data warehouse puede contener múltiples bases de datos. Dentro de cada base de datos, la data es organizada en tablas y columnas. Dentro de cada columna, se puede definir una descripción de la data, como integer o string.
- Las tablas pueden ser organizadas dentro de schemas. Cuando la data es ingestada, es almacenada en varias tablas descritas por el esquema.




¿Qué beneficios
obtengo del
Data Warehouse?

Data warehouse - beneficios

Ofrecen el beneficio de permitir que se analicen grandes cantidades de datos y extraigan un valor significativo de ellos, así como mantener un registro histórico.

- **Orientado a un tema:** pueden analizar datos sobre un tema o área funcional en particular (como marketing).
- **Integrado:** los almacenes de datos crean consistencia entre diferentes tipos de datos de fuentes dispares.
- **No volátil:** una vez que los datos están en un almacén de datos, son estables y no cambian.
- **Variable en el tiempo:** El análisis del almacén de datos analiza los cambios a lo largo del tiempo.



¿Cómo se
almacena la
data
transformada?

Data warehouse - almacenar data



Se puede almacenar data desde:

- Procesos de información (pipelines)
- Archivos locales o desde un ruta (URI)
- Diferentes formatos: Avro, Json, parquet, ORC, CSV, etc.
- Se debe especificar DB, tabla, schema (opcional)

Data warehouse - almacenar data



Algunas herramientas de ETL ya permiten especificar directamente cuál será el Data warehouse de destino, por lo que casi automáticamente permite que la información sea almacenada en este.

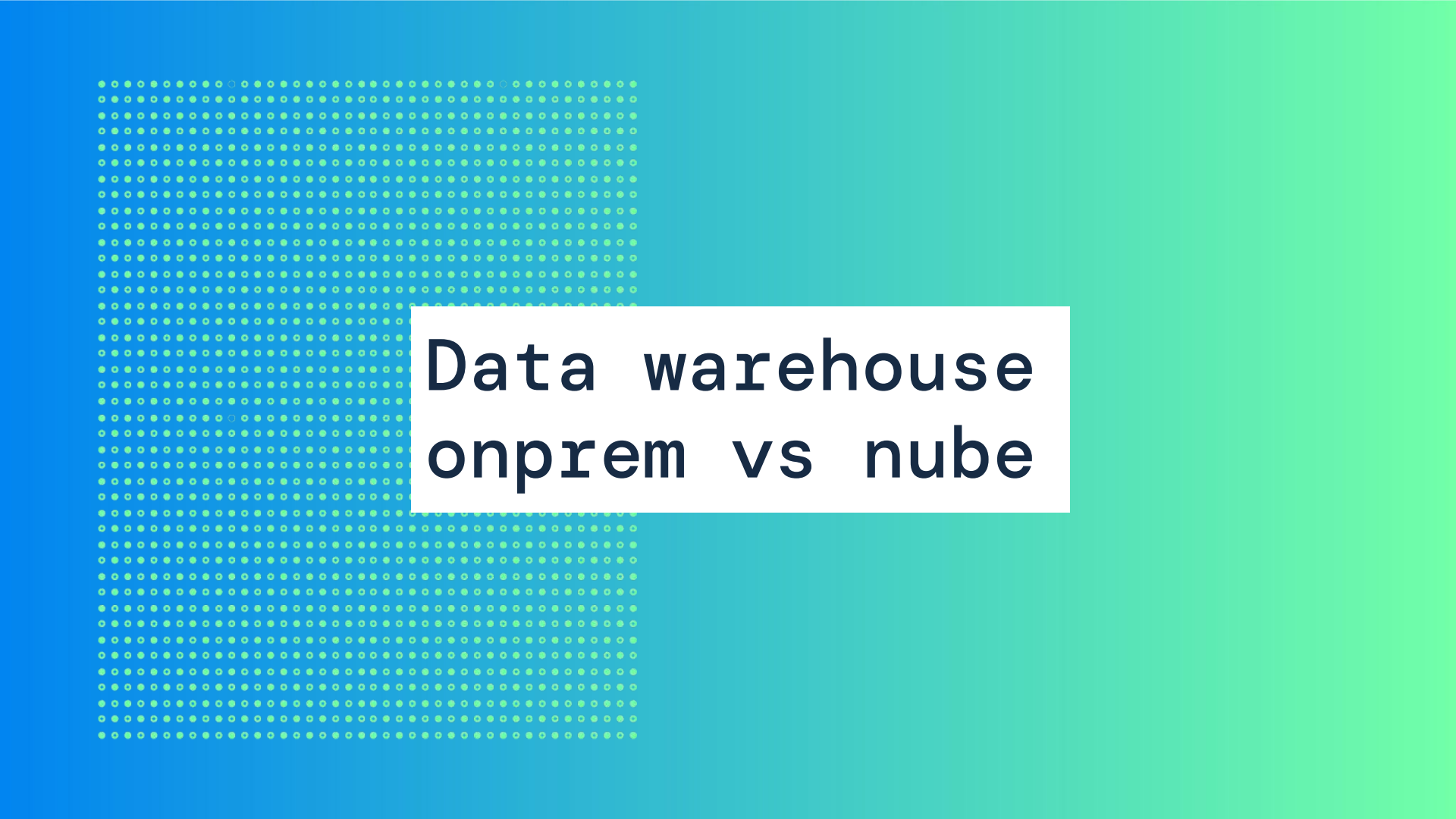
Data warehouse - data streaming



Con la llegada de las tecnologías de almacenamiento de bajo costo, la mayoría de las organizaciones hoy en día almacenan sus datos de streaming.

Hay varias opciones para almacenar streaming de datos, y sus ventajas y desventajas.

- Data warehouse: Es fácil analizar con SQL - Almacenamiento e ingest caros
- Broker de mensajes (kafka): fácil almacenar - storage 10 veces más caro que DL
- Data lake: barato, sin estructura de tablas - No SQL, alta latencia




Data warehouse onprem vs nube

Data warehouse - onprem vs nube

Debido a diferentes razones varios Data Warehouses están siendo migrados desde data centers on premise a la nube. Principales ventajas de migrar el DW:

- **Dimensionamiento:** más fácil consolidación y racionalización
- **Costos:** monetización más rápida de los datos en la nube
- **Seguridad:** la nube ofrece mejor protección
- **Licencias:** altos costos de renovación de licencias y tiempos de negociación
- **Hardware:** altos costos de renovación de hardware



¿Qué tipo de
almacenamiento
o existe?

Tipos de almacenamiento

Dependiendo del tipo de información a almacenar y la frecuencia, dependerá del data warehouse o DB a utilizar.



Apache HBase

Apache HBase

¿Qué es?

Apache HBase es una base de datos de Hadoop, un gran almacén de datos escalable y distribuido.

- Acceso de lectura/escritura aleatorio y en tiempo real
- Alojamiento de tablas muy grandes
- Base de datos no relacional, versionada, distribuida y de código abierto

Apache HBase

Opciones en Cloud

La ventaja que tienen estos servicios de DB en Big Data en la nube es que son auto escalables y sin gestión de servidores.

- Cloud Big Table - GCP
- Amazon DynamoDB - AWS
- Azure Cosmos DB - Azure



Apache Hive

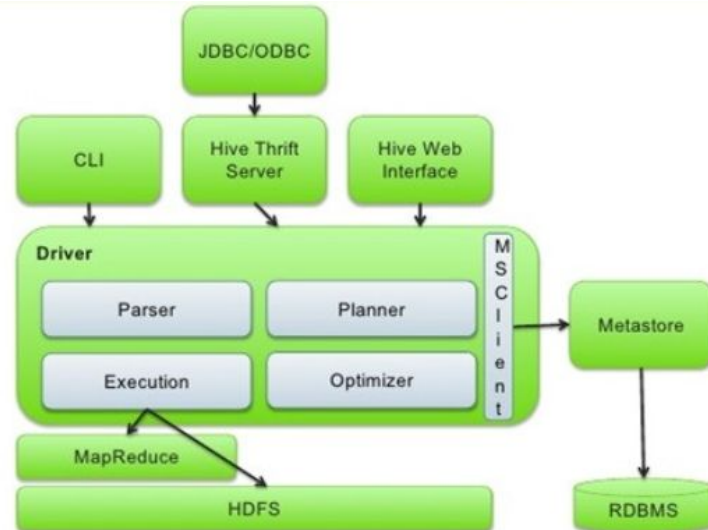
Apache Hive

¿Qué es?

Es un data warehouse que facilita la lectura, escritura y administración de grandes datasets que residen en almacenamiento distribuido mediante SQL.

Apache Hive

La estructura de datos se puede proyectar sobre los datos que ya están almacenados. Se proporciona una herramienta de línea de comandos y un controlador JDBC para conectar a los usuarios a Hive.



Apache Hive

Opciones en Cloud

La ventaja que tienen estos servicios de DB en Big Data en la nube es que son auto escalables y sin gestión de servidores.

- BigQuery - GCP
- Amazon Redshift - AWS
- Azure Synapse Analytics - Azure



DW exportando
data

Exportar data

- En algunas ocasiones necesitamos exportar la información de una o más tablas para ingestarlas en otro proyecto de Big Data (puede ser multi cloud) o para compartir información con personas externas a mi empresa (ej: consultores).
- La mayoría de los DWs nos permiten exportar información en algún formato que pueda ser leído fácilmente (ej: Csv, Txt)

Exportar data

Ejemplo

#Exportar a un directorio HDFS la tabla tripdata_table_km

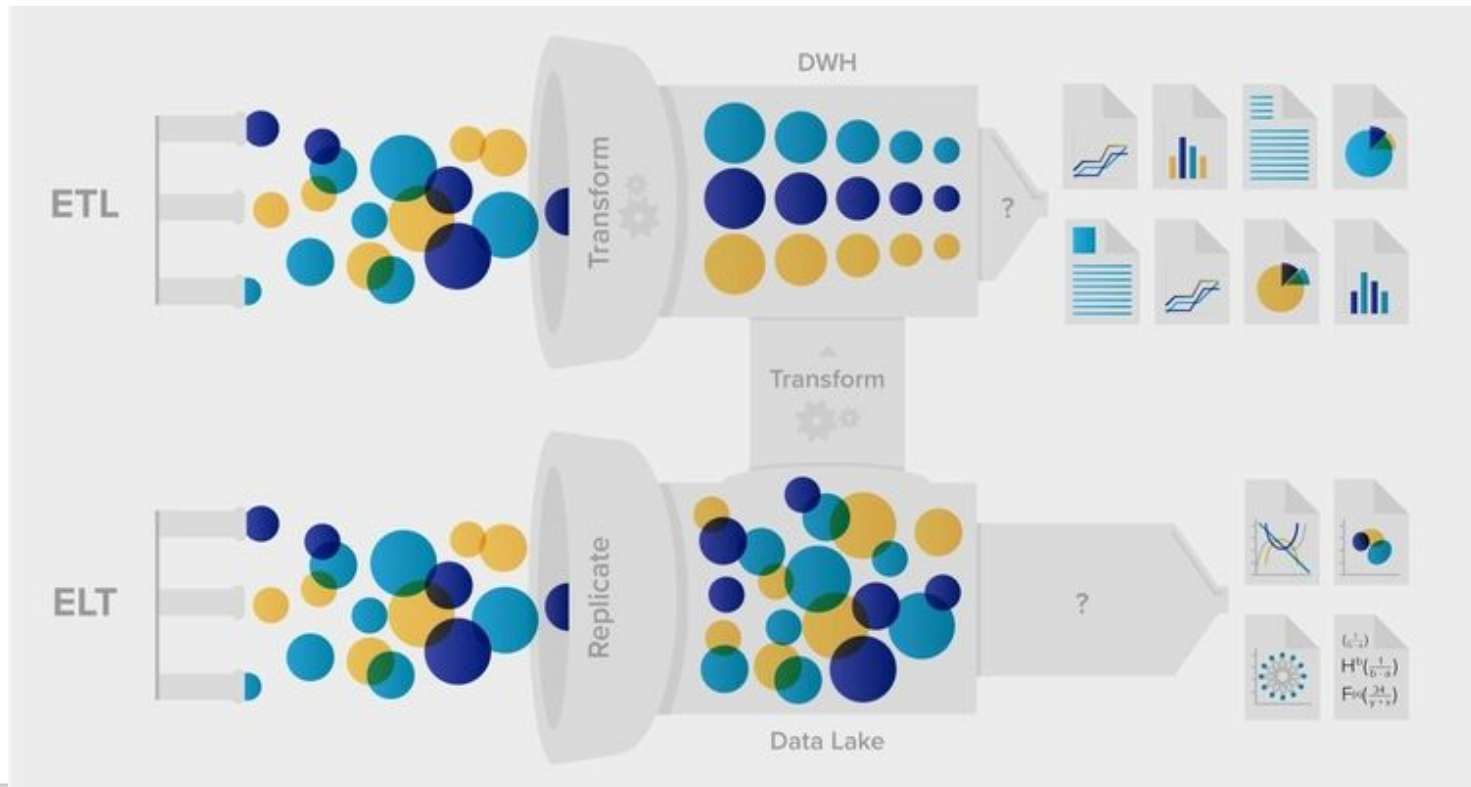
```
INSERT OVERWRITE DIRECTORY '/tmp/export' ROW FORMAT DELIMITED FIELDS  
TERMINATED BY ',' SELECT * FROM tripdata.tripdata_table_km;
```

Luego renombrar el file a csv




Data Lake

Data lake



Data lake

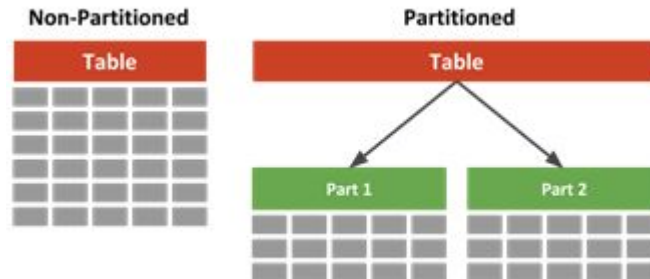
- Almacenan una gran cantidad de datos dispares y sin filtrar que se utilizarán más tarde para un propósito particular.
- Los datos de aplicaciones, aplicaciones móviles, redes sociales, dispositivos IoT, videos, sonido, etc. se capturan y se guardan como datos sin procesar.
- Cuando se necesita almacenamiento de bajo costo para datos no estructurados y sin formato de múltiples fuentes



Partitions,
clustering, y
sharding en
tablas

Partitions

El particionamiento soluciona problemas clave en el soporte de tablas e índices muy grandes al permitirle descomponerlos en piezas más pequeñas y más manejables llamadas particiones. Particionar una tabla puede hacer que las queries se ejecuten más rápido mientras gasta menos.



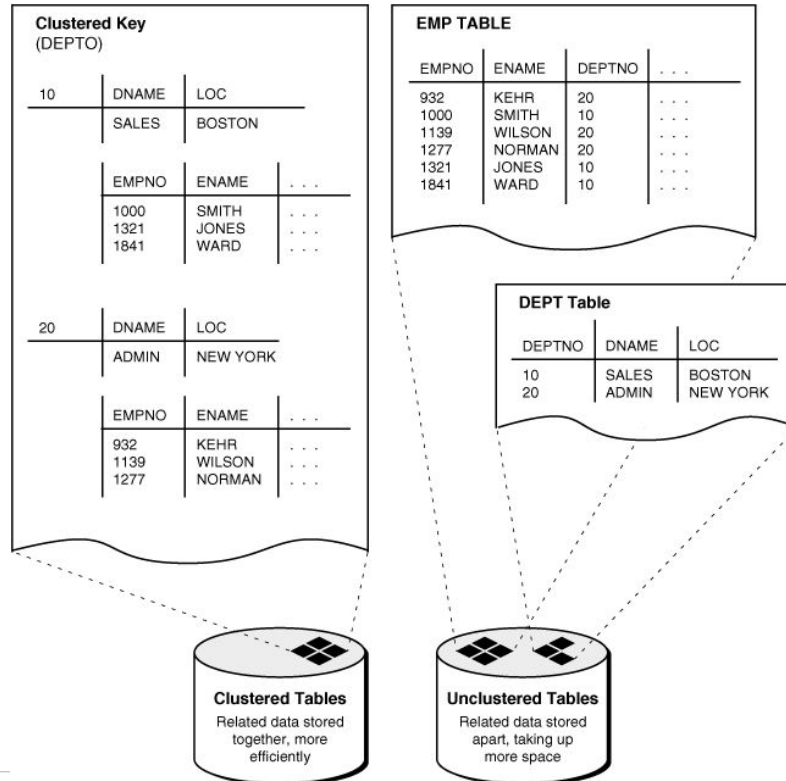
Clustering

Se compone de un grupo de tablas que comparten los mismos bloques de datos.

Las tablas se agrupan porque comparten columnas comunes y, a menudo, se usan juntas.

Clustering

REAL * IA PARA UN MUNDO



Sharding

El sharding es la práctica de optimizar las bases de datos al separar las filas o columnas de una tabla grande en varias tablas más pequeñas. Las nuevas tablas se denominan "shards".

El sharding y el particionamiento consisten en dividir un gran conjunto de datos en subconjuntos más pequeños. La diferencia es que el sharding implica que los datos se distribuyen entre varias computadoras, mientras que la partición no.

Sharding

Cada tabla nueva tiene el mismo esquema pero filas únicas (sharding horizontal) o tiene un esquema que es un subconjunto del esquema de la tabla original. (sharding vertical).

Original Table

CUSTOMER ID	FIRST NAME	LAST NAME	CITY
1	Alice	Anderson	Austin
2	Bob	Best	Boston
3	Carrie	Conway	Chicago
4	David	Doe	Denver

Vertical Shards

VS1

CUSTOMER ID	FIRST NAME	LAST NAME
1	Alice	Anderson
2	Bob	Best
3	Carrie	Conway
4	David	Doe

VS2

CUSTOMER ID	CITY
1	Austin
2	Boston
3	Chicago
4	Denver

Horizontal Shards

HS1

CUSTOMER ID	FIRST NAME	LAST NAME	CITY
1	Alice	Anderson	Austin
2	Bob	Best	Boston

HS2

CUSTOMER ID	FIRST NAME	LAST NAME	CITY
3	Carrie	Conway	Chicago
4	David	Doe	Denver



DataWarehouse