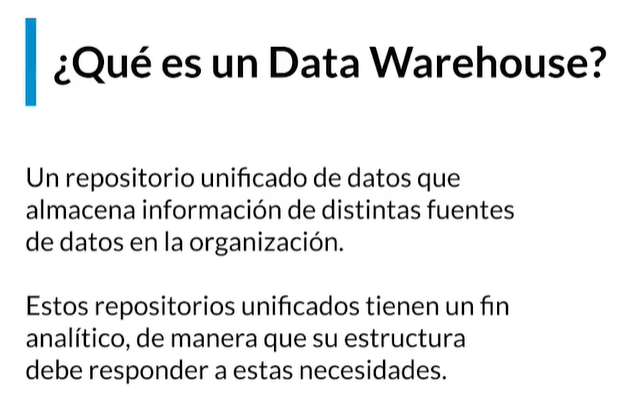
**DIAPOSITIVA 1**



**Fuentes**: sistemas transaccionales (como sap), aplicaciones empresariales (como crm), tablas de Excel, información de bases de datos, archivos planos, etc.

**Tipos de datos**: estructurados, no estructurados y metadatos (datos que describen otros datos).

**Fin**: No almacenar datos (para eso s3)

No es consultar datos, hacer queries (para eso una bd o si están en s3 athena)

Sí es la analítica de datos, es decir, extraer información valiosa a partir de datos crudos, revelando patrones y tendencias que ayudan en la toma de decisiones basada en datos.

En síntesis, un data warehouse es una estructura analítica, que no mantiene operaciones constantes del negocio, sino que está orientado a la analítica de datos.

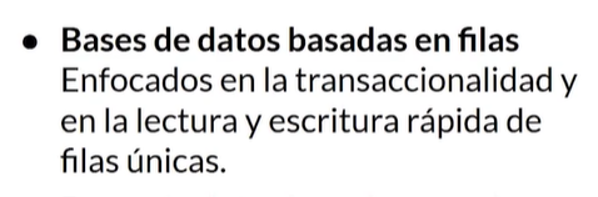
**¿Cómo me llevo los datos al datawarehouse?** Extracción, transformación y carga, entonces tenemos múltiples procesos etl alimentando mi dw.

**DIAPOSITIVA 2**

**BD basadas en filas**: Orientadas a OLTP: Transacciones on line - Procesamiento de transacciones en línea.

Soportan toda la operación transaccional del negocio.

Como mysql, Oracle, postgres: Muchisimas transacciones de manera rápida e integra en la bd.

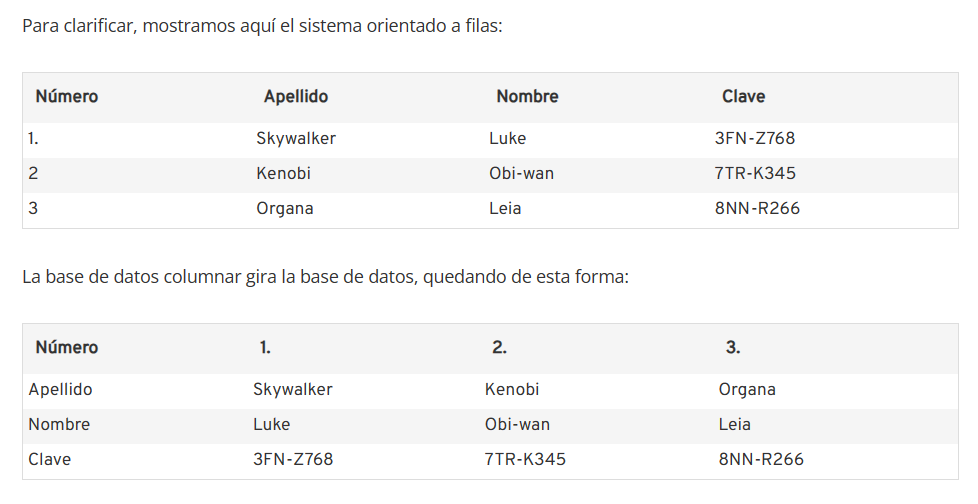


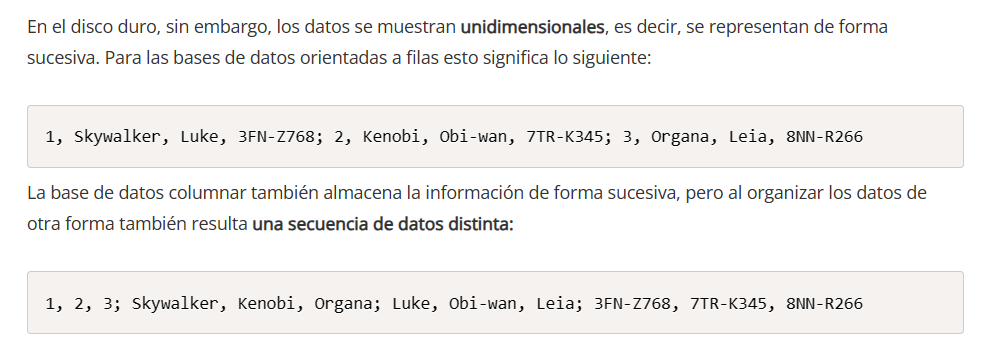
**BD columnares**: Consultas complejas de grandes lotes de datos de la manera más eficiente.

Procesar queries complejos de grandes cantidades de datos con un fin analítico

Procesan grandes cantidades de información, se encuentran optimizadas para procesos de analítica.







**DIAPOSITIVA 3**

Amazon Redshift es un servicio de almacenamiento de datos administrado a escala de petabytes en la nube. Se basa en la arquitectura MPP (procesamiento masivo paralelo), que consta de varios nodos de Elastic Compute Cloud (EC2) y permite una ejecución rápida de consultas complejas en grandes cantidades de datos, mediante la distribución del procesamiento de la consulta a múltiples nodos y unidades de procesamiento virtuales dentro de cada nodo de su almacén de datos.

Redshift utiliza SQL para analizar datos estructurados y semiestructurados en almacenamientos de datos, bases de datos operativas y lagos de datos, con hardware y machine learning diseñado por AWS para ofrecer rendimiento al mejor precio a cualquier escala.

En resumen, es un repositorio unificado de datos con un propósito analítico.



**DIAPOSITIVA 4**

El secreto de redshift para procesar grandes cantidades de datos es REPARTIR EL TRABAJO.

Redshift no se instala en un único servidor sino que se instala en un clúster (arreglos de varios servidores conectados {NODOS})

Nos conectamos a un **nodo líder** el cual organiza y asigna las tareas a los otros nodos, los nodos seguidores, y cada **nodo seguidor** reparte su tarea entre sus Slices. Para trabajar la tarea asignada por el nodo lider en paralelo.

SLICES es una tabla interna en Redshift que se utiliza para mapear **sectores** a nodos. Cada nodo en el clúster tiene varios slices, y cada slice es responsable de procesar una parte específica de los datos

Es decir, el nodo líder me orquesta los procesos.

Redshift viene de postgresql, la arquitectura de postgresql sirve como base para la arquitectura de redshift:



**DIAPOSITIVA 5**

1. Entender la naturaleza de los costos, por qué me van a cobrar por este servicio: Cómputo, memoria y almacenamiento

https://aws.amazon.com/es/redshift/pricing/

2. Determinar el tipo y el tamaño más adecuado a la carga de trabajo, hay 2 familias: dc2 era la familia anterior y ra3 es la familia nueva

Determinar número, tipo y tamaño de las instancias

RA3: desacoplado el almacenamiento del cómputo

Con la última generación de tipos de nodos RA3, puede escalar la computación y el almacenamiento de forma independiente según sus cargas de trabajo.

El almacenamiento para este tipo de nodo es Amazon Redshift Managed Storage (RMS), que es respaldado por Amazon S3.

Las puedo adquirir el servicio a demanda o con reserva

3. Naturaleza de la solución: Voy a hacer queries atemporales, poco frecuentes, no voy a estar constantemente teniendo estructuras ni relaciones complejas me puedo llevar los datos a s3 y usar athena.

Cargas a doc: redshift serverless, solo quiero obtener insights temporales