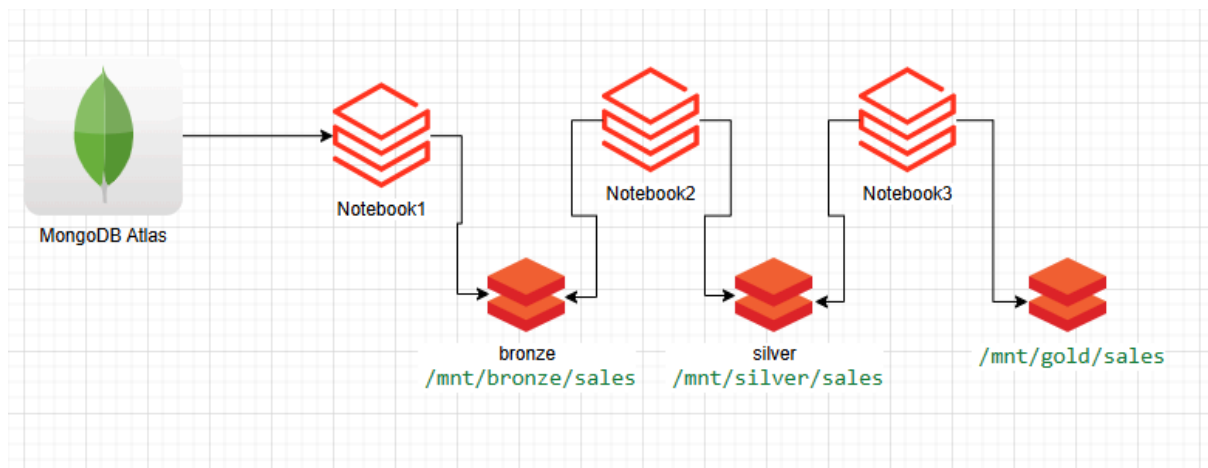


Replicación de Capas del Datalake (Bronze, Silver y Gold)

Crear 3 notebooks en Databricks para realizar un flujo que sea capaz de ingestar datos desde la base de datos de ejemplo **sample_supplies** de MongoDB Atlas (la cual cuenta con la colección: *sales*), los datos se procesarán y almacenarán en tres capas para replicar la estructura de un datalake por capas:

- **Capa Bronze(Notebook1):** Ingesta de datos en bruto (raw) desde MongoDB Atlas y almacenarlos en formato Parquet.
- **Capa Silver(Notebook2):** Lectura de los archivos Parquet de la Capa Bronze para limpieza y transformación de los datos y guardado de estos en la Capa Silver
- **Capa Gold(Notebook3):** Lectura de los archivos Parquet de la Capa Silver y ejecución de análisis y generación de KPIs y guardado de estos en Capa Gold.
- **Operación Extra:** Creación de Tabla Delta para la representación de los KPIs en un informe a través de los datos de la Capa Gold.



1- Capa Bronze: Ingesta de Datos(Notebook1)

Objetivo:

Ingerir y almacenar los datos en bruto desde MongoDB Atlas de la colección *sales*. Estos datos se almacenarán sin transformaciones, en un formato como Parquet, para su posterior procesamiento.

Tareas:

1. Conexión a MongoDB Atlas:

- Utilizar **Spark** para conectarse a la base de datos de MongoDB Atlas (En caso de no conseguirlo probar con el cliente de PyMongo).
- Leer y mostrar los datos de la colección *sales*.

2. Conversión a Tipos Nativos(Opcional):

- Si fuera necesario convertir campos especiales (por ejemplo tipos ObjectId) a tipos nativos (por ejemplo, `string`).

3. Almacenamiento en la Capa Bronze:

- Crear un DataFrame de PySpark con el esquema definido.
- Guardar el DataFrame en un directorio (por ejemplo, `/mnt/bronze/sales`) en formato **Parquet**.

2- Capa Silver: Transformación y Enriquecimiento de Datos

Objetivo:

Realizar transformaciones sobre los datos en bruto para limpiar y enriquecer la información, de modo que se facilite el análisis posterior.

Tareas:

1. Lectura de datos:

- Leer los datos desde (`/mnt/bronze/sales`) y trabajar con DataFrame..

2. Aplanar la Información del Cliente:

- Extraer los subcampos del objeto `customer` y crear columnas individuales:
 - `customer_gender`
 - `customer_age`
 - `customer_email`
 - `customer_satisfaction`
- Eliminar la columna anidada `customer` una vez aplanada.

3. Otras Transformaciones (opcional):

- Se pueden aplicar transformaciones adicionales, como conversión de fechas, filtrado de registros nulos, etc. Realiza las que creas necesarias.

4. Escribir los datos:

- Guarda el DataFrame resultante de las transformaciones en formato parquet en `/mnt/silver/sales`