End to End Big Data Engineering Project

[Ecommerce Data]

Pre - Requisites

1. Motivation to learn and complete project
2. A laptop and a notebook
3. Tech pre-requisite : Python and SQL
4. Azure free acount (\$200 free credit)
   - → account
   - → credit card

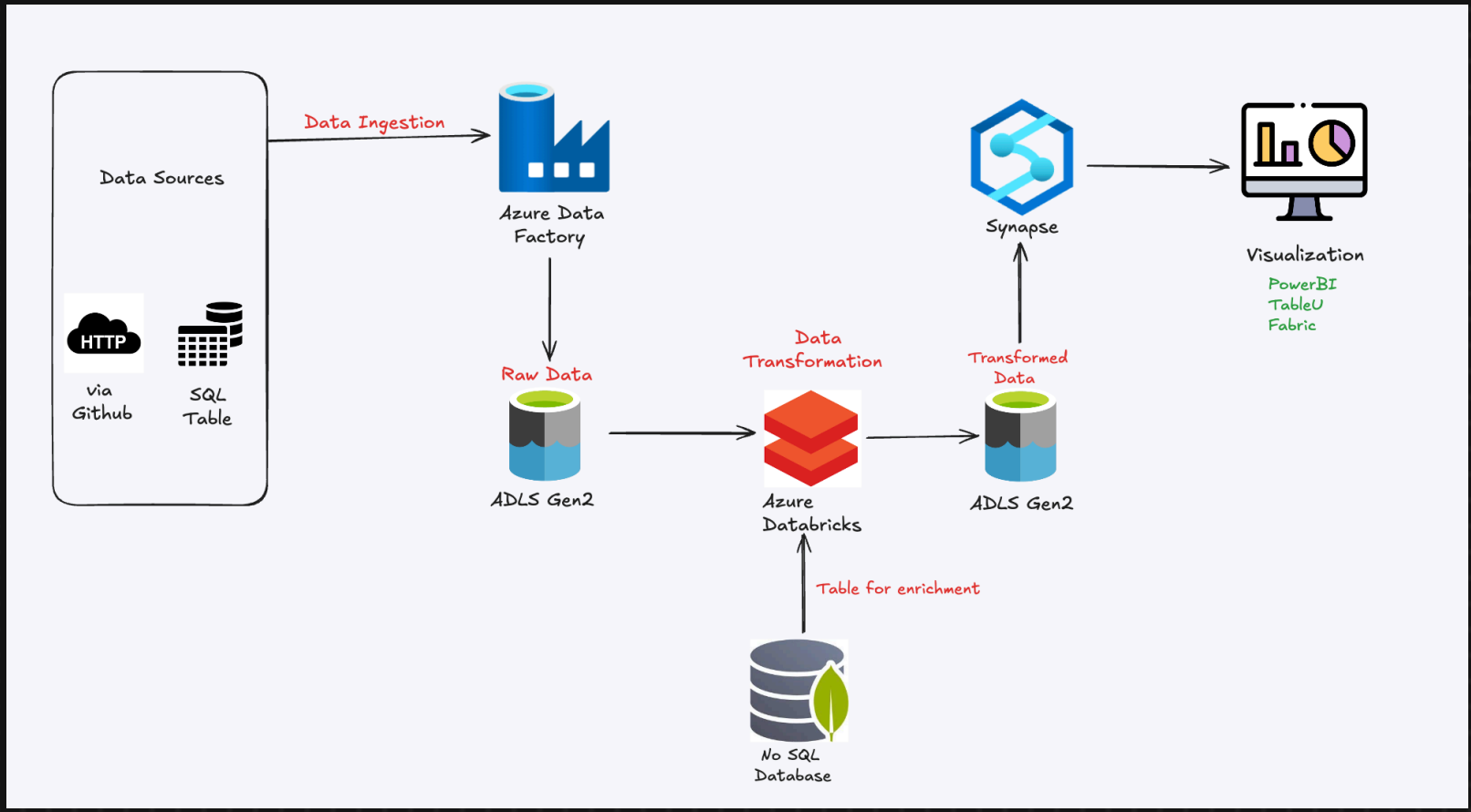
Ecommerce Dataset

Brazilian E-Commerce Public Dataset by Olist

100,000 Orders with product, customer and reviews info

- olist_customers_dataset.csv
- olist_geolocation_dataset.csv
- olist_order_items_dataset.csv
- olist_order_payments_dataset.csv
- olist_order_reviews_dataset.csv
- olist_orders_dataset.csv
- olist_products_dataset.csv
- olist_sellers_dataset.csv
- product_category_name_translation.csv

Project Architecture



Azure account

We will use a free account provided

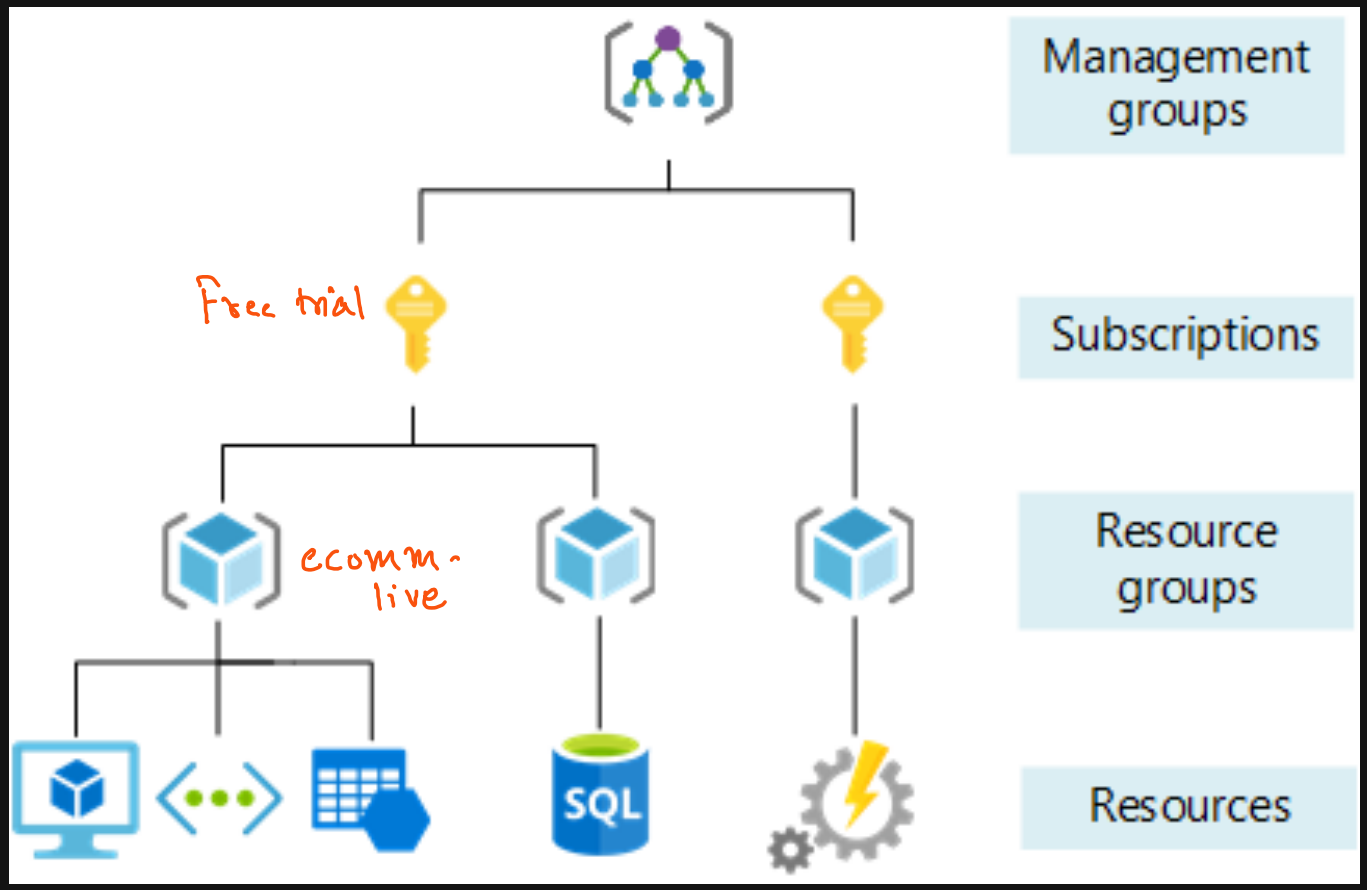
email-Id (gmail also work)

Credit card (only 2 rupees charged, no autopay)

Free account

Pay as you go

Azure account Structure



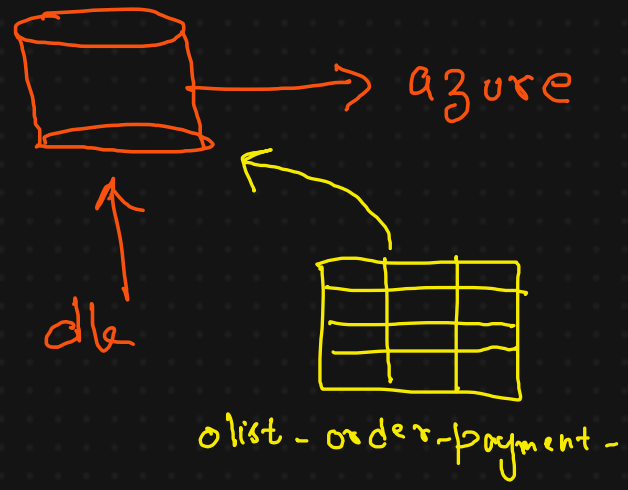
Databases

Sql

no sql

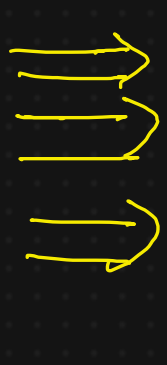
http

files.io



Multiple sources

ingestion

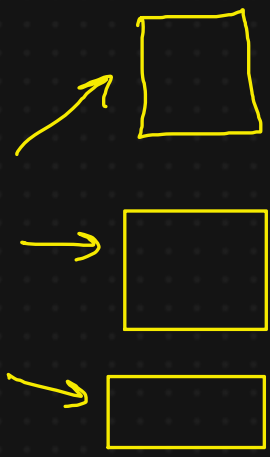


databricks

raw
Bronze

Semi-p
Silver

Serving

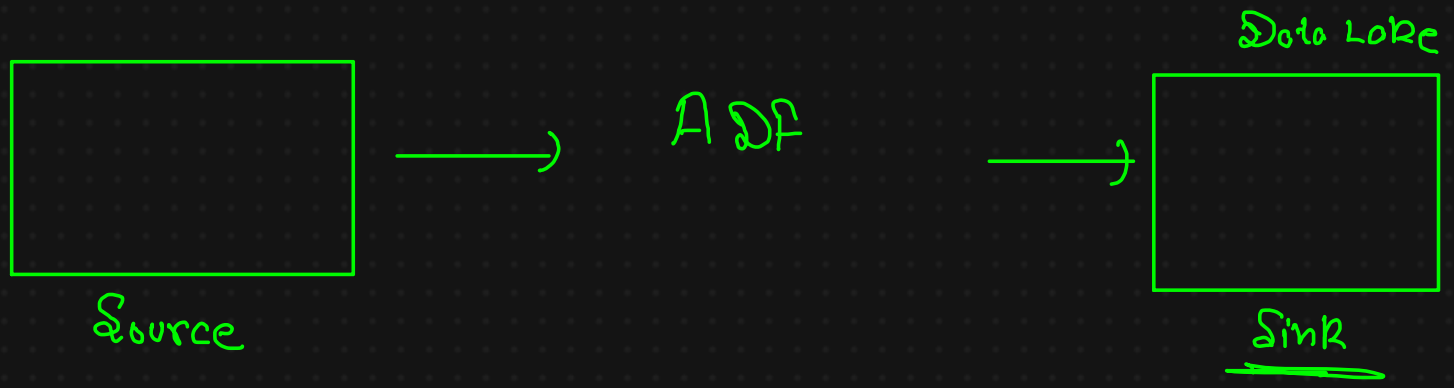


serving
Gold

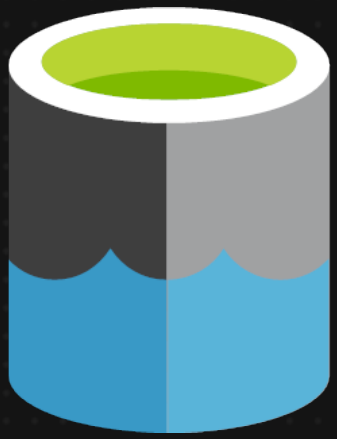
Azure Data Factory

Azure data factory is a cloud tool that collects data from different sources, organizes it and mores it where we need it.

→ Data collection (Ingestion)
→ Data Transformation (Cleaning and organizing)
→ Data movement



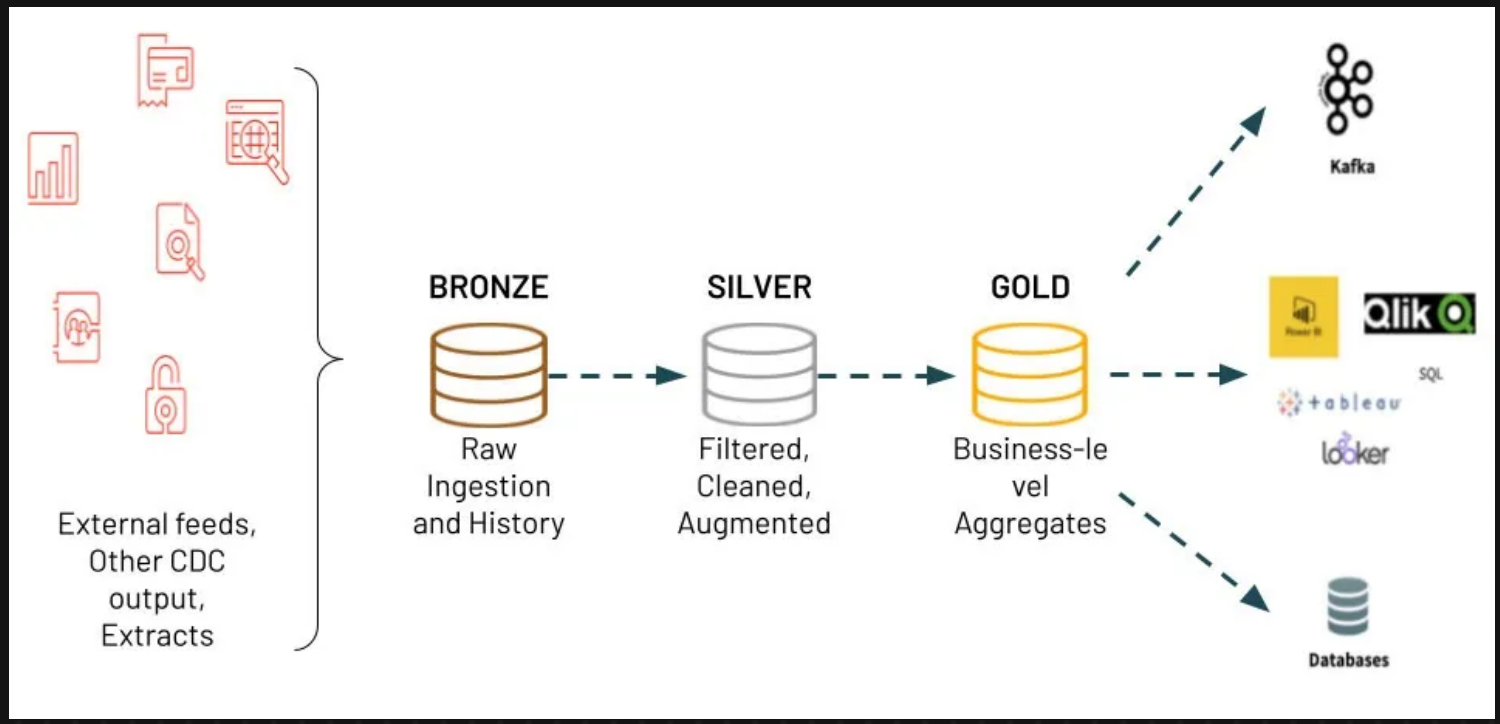
Azure Data Lake Storage (ADLS) Gen2



Azure data lake storage is a secure cloud platform that provides scalable, cost effective storage for big data analytics.

→ Store massive amount of data
→ Keeps data organized
→ Fast and efficient

Medallion Architecture



Data Ingestion

Used Azure Data factory with http & a SQL server.

→ parametrization

→ for each activity

→ lookup

Azure Databricks

→ Spark powered
→ Integrated with azure
→ Handles big Data easily
→ Great for machine learning

Azure Databricks Workflow

1. Read data from ADLS gen2
2. Perform basic transformation like cleaning, renaming & filtering
3. Use join operation to integrate multiple dataset.
4. Enrich the data via mongoDB.
5. Perform aggregation & derive insights
6. Write final data back to ADLS gen2.


Azure
Permission
Azure Data Lake Storage Gen2

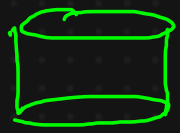
Azure Synapse

Azure Synapse is a cloud based data warehouse and analytics service that combines.

→ data integration
→ enterprise data warehousing
→ Big Data analytics

To access the data in azure Synapse, we need to give permission to Synapse via ADLS gen2.

Serverless



Azure Synapse

LAKE HOUSE Abstraction
(data lake + data warehouse)

DATA LAKE

SQL Pool :- Serverless vs Dedicated

Feature	Serverless SQL Pool (Pay-as-you-go)	Dedicated SQL Pool (Provisioned)
How it works?	You don't set up or manage servers. It is **on-demand**. Synapse spins up resources when you need them.	You reserve fixed resources (compute power) for your data, which are always available.
Payment Model	Pay only for the queries you run.	Pay for the reserved compute (fixed cost).
Use Case	Good for **exploring large datasets** or running queries occasionally.	Best for **big workloads** that need fast performance and predictable speed.
Data Location	Data stays in the **Data Lake** (e.g., ADLS Gen2).	Data is loaded into a **dedicated SQL database** in Synapse.
Setup Complexity	Very **easy** to set up—no servers to manage.	More setup needed, but provides **high performance**.

Bus

own Car

Azure Synapse Workflow

0. Medallin architecture and Lakehouse (Pre-requisite)
1. Create a serverless SQL pool. (for just reading/Querying)
2. Create Schema & then User / password
3. Use OPENROWSET to read data from silver layer.
   (here data is not stored, just referenced to silver layer)

4. Create a view to easily Query the data.

Data Migration to Gold Layer

Input

Format – parquet

Source – When we read the data

Output

Format → parquet

Source

CETAS ⇒ create external table as select