**Social media pipeline documentation**

**Project Overview :**

This project aims to create an end to end pipeline to transform batch raw social media data into meaningful insights to support business intelligence, reporting, and advanced analytics. The pip0eline will facilitate the ingestion, processing, storage, and analysis of data from multiple sources, ensuring data integrity, scalability, and accessibility.

**The objectives:**

1- Develop a robust data ingestion pipeline that can handle batch data from sources .

2- Implement data processing workflows to transform raw data into structured formats, making it suitable for analysis and reporting.

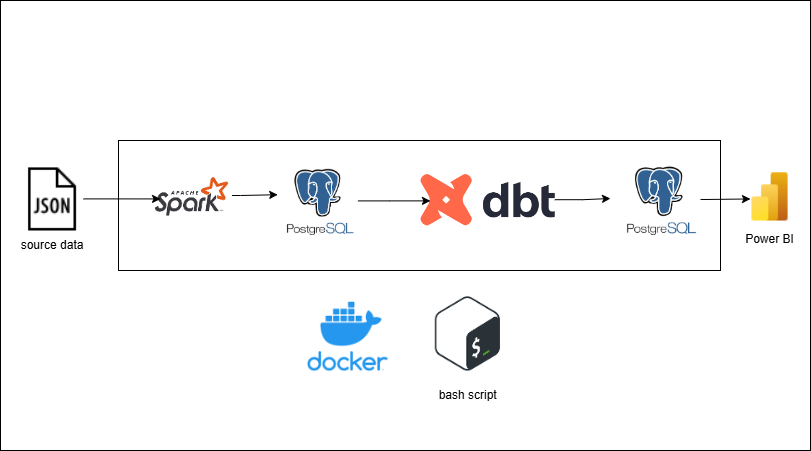
3- Establish a data warehouse using PostgreSQL to serve as a centralized repository for business intelligence and reporting.

4- Provide actionable insights and reports to business teams,

**Tools and Technologies**

* **Apache Spark:** Data processing, transformation, and analysis
* **PostgreSQL:** Data warehousing and SQL-based querying
* **DBT:** creating data marts and models by applying transformations and aggregations
* **Power BI :** creating a dashboard to serve business needs
* **Docker:** Containerization for consistent and isolated environments

**Architecture and Design :**



-So in this document I will clarify each phase of the pipeline that consists of 5 phases :

**1 – Ingesting from data source :**

-The data comes from its source in a json format (semi – structured )

And it has nested json objects , so each main object represents user and inside this object we can find another objects for posts related to that user and inside each post object we can find another objects for each comment related to that post

**-**I used pyspark to ingest the data from this json file in its raw format and saving it in an initial structured data frame

**2- data validations and cleansing :**

-extracting a data frame for each entity from the initial data frame

-applying validation on user email data to make sure that there is no data entry issue

-filling null values with another values based on the field data type and meaning

-extracting a field for each reaction type from one reactions field which has list of values

- adding total reactions field to calculate the total sum of reactions

-extracting dates from timestampe fields to be easier to use in analytics

**3 – data warehouse design :**

The chosen data warehouse design here is a star schema with one fact table and three dimensions

* Each record of interaction fact table represents each comment on each post
* If the post does not have any comment so it will be represented by only one record in the fact table
* If the post has multiple comments so it will have multiple records ( record for each comment )

-users dimension for additional users data

-posts dimension for additional posts data

-comments dimension for additional comments data

A black screen with white text

AI-generated content may be incorrect.

**4- data marts for analytics :**

-creating 6 data marts and models using DBT for analytical purposes :

1-top users getting reactions on posts

2- top users getting comments on posts

3- tags in posts getting top total reactions

4- tags in posts getting top love reactions

5- tags in posts getting top angry reactions

6-top locations by total posts count

**5 – dashboard ( final layer ) :**

-Creating a dashboard to serve business needs using Power BI

**A screenshot of a social media analytics report

AI-generated content may be incorrect.**