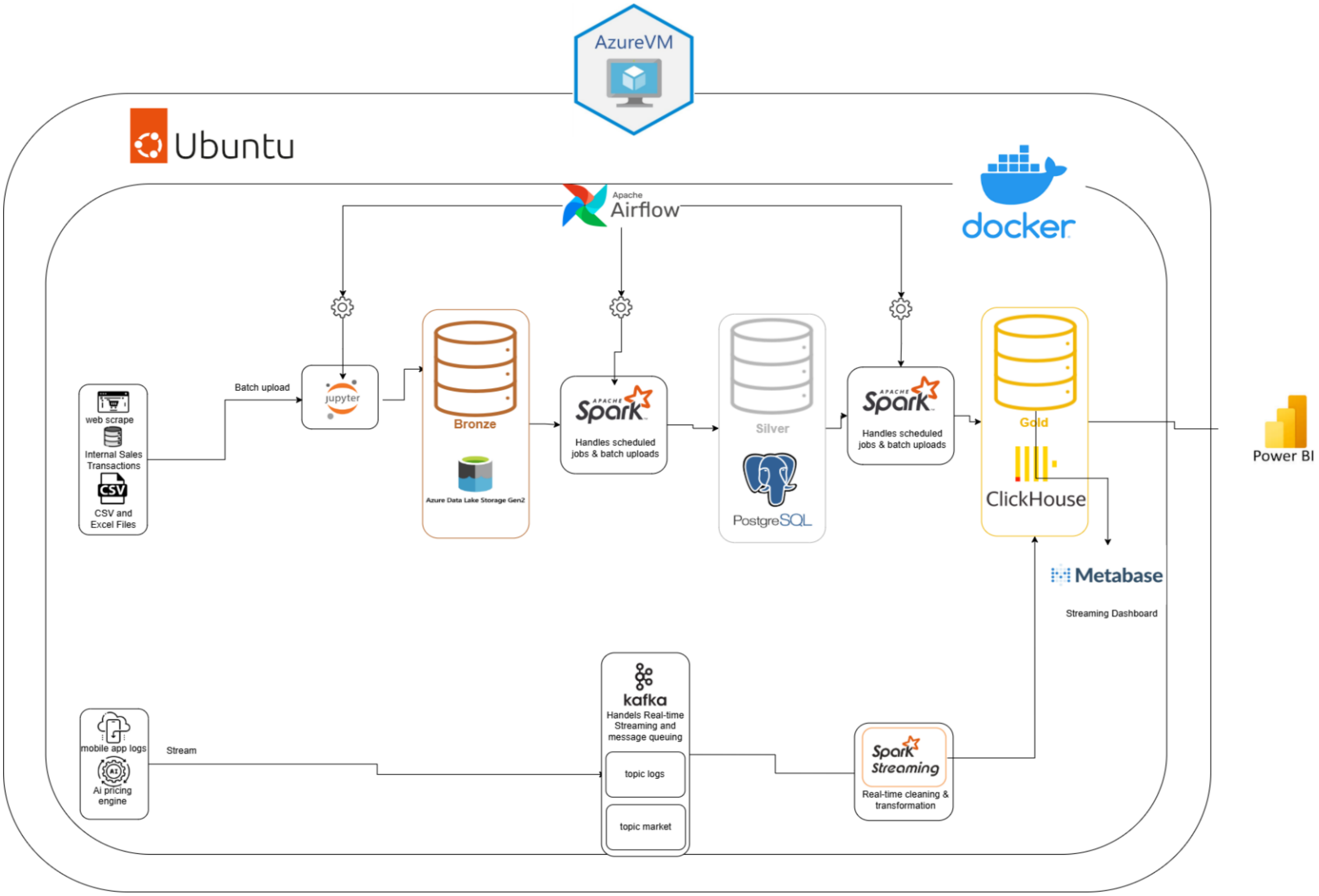
Jr.Data Engineer Assesment

Dayera Market

Data Pipeline (design)



1. Why Each Component Was Chosen

- **Azure VM & Docker:** To provide a flexible and cost-effective cloud environment. Docker is used for **containerization** of services (Kafka, Spark, DBs), ensuring isolation, easy deployment, and maintainability.
- **Kafka:** Serves as the **Real-Time Ingestion Layer**, guaranteeing high-throughput and reliable processing of data streams originating from mobile applications logs and the AI pricing engine.
- **Apache Spark:** The unified engine for all **Transformation and Cleaning (ETL/ELT)** operations, leveraging its capabilities for distributed processing across both batch and streaming tasks.
- **Apache Airflow:** Provides **Workflow Orchestration**, managing the scheduling, automation, and monitoring of all ETL jobs to ensure pipeline efficiency and continuity.
- **Storage Layers:**
  The bronze, silver, and gold model in pipeline design is a Medallion Architecture that organizes data into three layers of increasing quality
  - **Azure Data Lake (Bronze):** The raw, un-processed **Raw Storage** layer, designed to retain all source data for **Auditing** and potential **Reprocessing**.

  - **PostgreSQL (Silver):** Used as the intermediate layer for storing **Cleaned and Standardized** data, offering a structured environment before final aggregation.
  - **ClickHouse (Gold):** Selected as the **Analytical Data Warehouse** due to its exceptional performance in handling large-scale, real-time **OLAP** queries.
- **Metabase & Power BI:** The **Visualization Tools** connected directly to the Gold layer, used to deliver instant operational insights and historical business intelligence, respectively.

---

2. How real-time and batch data are handled

The architecture employs a hybrid design to effectively process batch data and real-time data:

- **Batch Processing:** Historical and less-frequent data (Web Scrapes, Sales Files) are ingested into the **Bronze** layer, processed by Spark for cleaning into the **Silver** layer (PostgreSQL), and then scheduled by **Airflow** for promotion to **Gold**.
- **Real-Time Streaming:** Log and AI engine data flow continuously through **Kafka**. **Spark Streaming** processes and transforms this data in **near real-time**, writing it directly to **ClickHouse (Gold)** to enable instant analytics.

---

3. Data Governance: Quality and Schema Evolution

We maintain a rigorous methodology to ensure high data quality and manage structural changes effectively:

- **Data Quality:** Strict validation rules (deduplication, null handling, type consistency) are applied in the **silver** layer using Spark. Failed records are isolated in a **Quarantine** zone, while Airflow monitors pipeline runs and triggers alerts on data anomalies.
- **Schema Evolution:** We adopt a **Schema-on-Read** approach for the flexible raw data in Bronze. In the structured silver and gold layers, structured formats (like Parquet) are used with **Schema Merging** enabled. A **Metadata** log in PostgreSQL tracks data lineage and column versions to ensure backward compatibility.

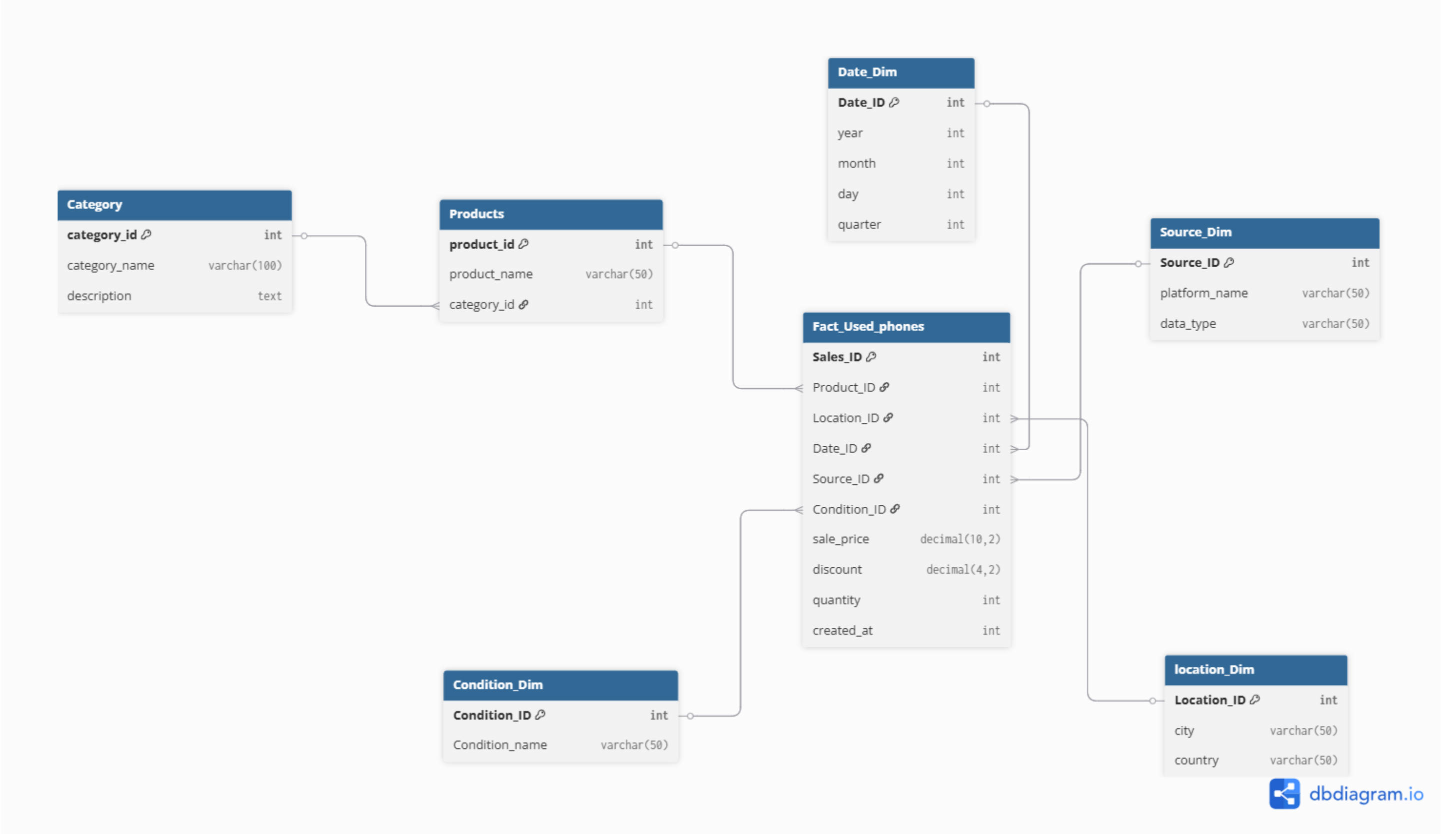
---

4. Data Warehouse Performance and Optimization

To achieve optimal query speed in the **Gold** layer, the focus is on architectural optimizations:

- **Partitioning and Indexing:** In ClickHouse, fact tables are partitioned by date **(YYYYMM)** and ordered by keys such as `product_id` and `sale_date` to greatly **optimize time-series query performance**.
- **OLAP Enhancements:** We leverage ClickHouse's **Columnar Storage** to achieve high **Data Compression** ratios and minimize query latency. **Materialized Views** and pre-aggregated tables are utilized to accelerate frequently executed dashboard queries.
- **Resource Management:** Performance is further secured by extensive **Parallelization** of Spark tasks and by an Airflow schedule that efficiently separates heavy batch and continuous streaming workloads.

Data warehouse Star Schema:



- The schema follows a **Star Schema** model centered around the `Fact_Used_Phones` table.
- The fact table stores key sales metrics such as **sale price**, **discount**, **quantity**, and **timestamps**.
- It connects to several dimension tables that provide descriptive context:
    - **Products** – details of each phone model.
    - **Category** – groups products into broader categories.
    - **Location_Dim** – geographic information (city, country).
    - **Source_Dim** – data origin (e.g., web scraping, marketplace, AI engine).
    - **Condition_Dim** – device condition (used, new, refurbished).
    - **Date_Dim** – supports time-based and trend analysis (year, month, day, quarter).

This design simplifies analytical queries and ensures **high performance** for reporting in Power BI or Metabase.