High-Level Design Document (HLD)

Project Name: Data Source Project : Airflow + Hive

Prepared By: Marwa Mansour

Date: 18-Dec-2025

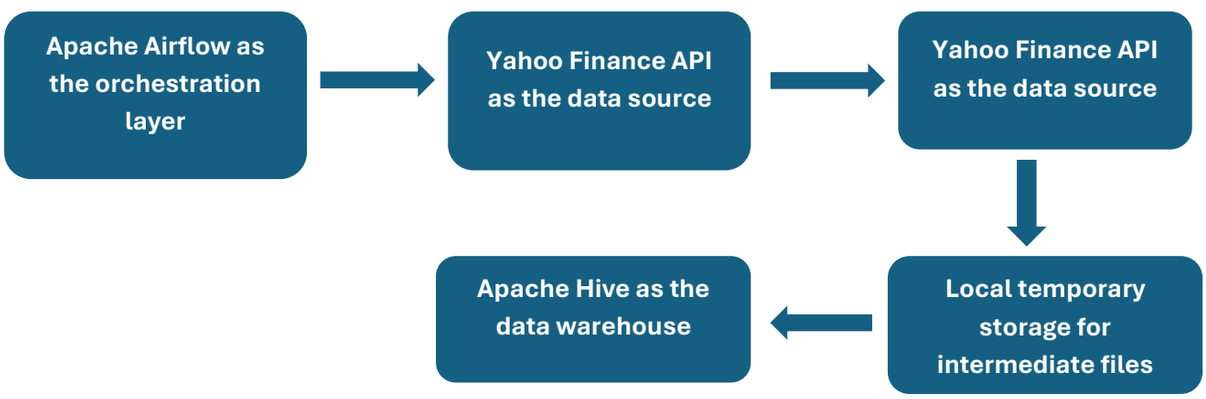
Business Logic (High Level)

- Airflow triggers the DAG on a daily basis.
- The pipeline fetches the latest EGX30 stock prices from the **Yahoo Finance**.
- The extracted dataset is parsed and converted into a structured format.
- Only the required attributes are retained:
- date, stock_symbol, price
- Data is validated to ensure:
    - Correct schema
    - No duplicate records for the same date and stock symbol.
- New daily records are appended to the existing dataset in HDFS ,Hive
- Execution status and logs are recorded in Airflow.

Tools Used

Tool	Purpose
Apache Airflow	Workflow orchestration & scheduling DAGs
Hive	Data warehouse for storing processed data

Data Flow Diagram



Expected Data Size

- Data size is not fixed and is determined by the number of stock records ingested per trading day.
  Storage grows linearly with the number of EGX30 constituents and trading days.