ELT Pipeline Architecture and Owner: Omar H. Attia

End-To-End Big-Data Engineering Project

Extraction, Loading and Transformation

E-Commerce – Retail Electronics Store

Python Data Generation 1

Utilizing python faker library to generate realistic e-commerce data for a retail store selling electronics. Transactions meta-data was picked accurately to simulate real-time transactional records. Then serialized data was sent as a JSON data to KafkaProducer for real-time processing and analysis.

Apache Kafka Data Streaming 2

The Kafka Producer connects to Kafka Broker with the default localhost address, serialized the data in JSON format. Applied gzip compression for efficient network utilization, and included logic to handle the transmission failure.

Logging sub-process was added to ensure handling of data. Also for Monitoring and Debugging

PySpark Streaming for Data Processing 3

Using PySpark's Structured Streaming, the setup codded via Python reads data from Kafka's Topic. Spark Session was also configured with 2GB of memory and 2 executors. Maximum offsets, processing 1k records per trigger. JSON data streamed from Kafka was parsed and formatted with a structured schema.

PySpark Data Loading (Write) 4

In PySpark streaming pipeline, processed data was written to Parquet files for efficient storage and faster querying. It scheduled updates every one minute and utilized checkingpoint feature to maintain fault tolerance. Ensuring continuous data processing.

Logging added.

PySpark Data Aggregation 5

Using PySpark batch processing. Parquet files were read to aggregate e-commerce data. A new spark session was configured like the previous one. Loaded these files and applied aggregation functions to derive insights for visualization matter.

Logging added.

Apache Airflow Orchestration 6

Utilizing Airflow scheduler DAGS. Default arguments were configured for scheduling and error handling, including daily execution and retries. Tasks included data from the E.L.T Pipeline. Task dependencies were managed to ensure sequential execution and monitoring.

PowerBi Analytics 7

Using PowerBi. Aggregated dataset was loaded on the PowerBi and used some interactive dashboard analytics to pinpoint the insights for the Retail Store key stakeholders.

Apache Zookeeper