User Generated Events Processing with Analytics Pipeline

About Data

E-commerce user events with features like timestamp, category\_id, event\_type. Dataset is in CSV format with total size of 15 GB.

Staging

Data Extraction

* Reading the dataset from CSV files into Pyspark DataFrames.

Data Cleaning and Transformation

* Applying schema and transformations using Apache Spark.
* Generating tables like ‘brand details’ using grouping and aggregation on Pyspark DataFrames.

Data Loading

* Cleaned DataFrames are loaded into PostgreSQL using Pyspark and JD4BC connector.

Visualisation

* Tableau dashboard (Postgres Connector and Custom SQL connector) with KPIs such as CVR, Cart Abandonment Rate, Repeat Customer Rate.

Real-Time Pipeline

Publishers

* Simulated event stream generation script (Python).

Subscribers

* MongoDB sink – inserts Kafka messages into MongoDB Cloud database with milliseconds delay.
* AWS S3 sink – writes backlog of Kafka messages after every 1000 records to S3 in JSON format partitioned by the timestamp.

Visualisation

* Real-time analytics dashboard with MongoDB Charts.

Batch Pipeline

Data Extraction

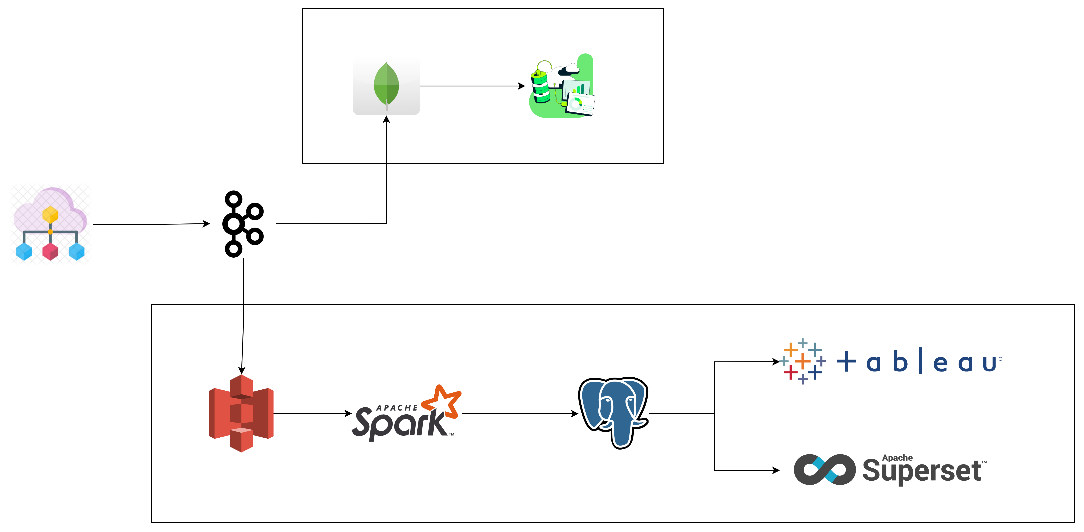
* Data of selected time frame is read from S3 into Pyspark DataFrame.

Data Cleaning and Transformation

* Apply schema using Spark.

Data Loading

* Cleaned DataFrames are loaded into PostgreSQL using Pyspark and JDBC connector.



Technologies Used

* Apache Spark (batch processing), Kafka (streaming pipeline), Superset (Dashboard)
* MongoDB (Real-time Inserts/ Charts), AWS S3 (Lake)
* PostgreSQL (RDBMS), Tableau (Dashboard)

Insights Gained

* Use of SQL queries in Superset allows room for optimization. Techniques like indexing (especially on frequently used columns) can be used to significantly improve query performance.
* Tableau has functions like SPLIT() which are complex to implement in ANSI SQL alone. Combined with the ability to add Custom SQL Queries as sources, Tableau provides comprehensive approach to visualizing relational data.
* Kafka is well suited for high volume and low latency operations required for user events streaming. Along with confluent connectors Kafka works well for this use case.