Build a Real-Time Dashboard with Spark, Grafana, and InfluxDB

Business Overview

A time-series metric includes a collection of data points that appear in a specific order over a particular timeframe. In this context, a "metric" refers to the data recorded at each time interval. A time-series metric has two primary characteristics:

- It may be assigned a numerical value.
- It refers to the fact that the measure evolves with time.

Each numeric data point relates to a timestamp as well as one or more identified dimensions. Time series data is commonly used to analyze website traffic, pricing fluctuations, demographic data, user clicks, and IT operations since it captures changes over time. As one might guess, because this time series data is frequently gathered in short periods, the data builds quickly. As a result, having a database optimized for time series data is critical, which is why they've been so popular in recent years.

In this project, we build a real-time e-commerce users analytics Dashboard. By consuming different events such as user clicks, orders, demographics, create a dashboard that gives a holistic view of insights such as how a campaign is performing country level, gender basis orders count, real-time purchase insights.

Data Pipeline

A data pipeline is a mechanism that allows data to be transferred from one system to another. The data may be modified or not, and it could be processed in real-time (or streaming) rather than in batches. The data pipeline includes everything from harvesting or capturing data using various tools to storing raw data, cleaning, verifying, and transforming data into a query-worthy format to visualizing KPIs and orchestrating the above process.

Dataset Description

The batch data consists of 100,000 auto-generated user demographic data points, including the following features:

- Id

- Age
- Gender
- State
- Country

While the stream data is based on user purchase events and is produced every 1 second along with a timestamp when joined with batch data. This data includes the following features:

- Id
- campaignID
- orderID
- total_amount
- units
- tags- click/purchased

Tech Stack

➔ Language: Java8, SQL

➔ Services: Kafka, Spark Streaming, MySQL, influxDB, Grafana, Docker, Maven

Apache Kafka

Apache Kafka is a distributed data storage designed for real-time data intake and processing. Streaming data is created continuously by hundreds of data sources, which generally transmit data records simultaneously. A streaming platform must cope with the continual input of data and process it sequentially and gradually.

Kafka is most used to create real-time streaming data pipelines and applications that react to data streams. It mixes communications, storage, and stream processing to allow historical and real-time data to be stored and analysed.

Spark Streaming

Spark Streaming is a Spark API service that enables data engineers and scientists to handle real-time data from various sources, including Kafka, Flume, and Amazon

Kinesis, etc. Data may be delivered to file systems, databases, and live dashboards once it has been analysed.

MySQL

MySQL is a SQL (Structured Query Language) based relational database management system. Data warehousing, e-commerce, and logging applications are just a few of the uses of the platform.

InfluxDB

InfluxData developed InfluxDB, an open-source time-series database. It's developed in Go and designed to store and retrieve time series data quickly and reliably in domains including operations monitoring, application metrics, Internet of Things sensor data, and real-time analytics.

Grafana

Grafana is a web application for interactive visualisation and analytics that is open source and cross-platform. When linked to supported data sources, it displays charts, graphs, and alerts on the web for mainly time series data.

Approach

- User purchase events in Avro format is produced via Kafka.
- Spark Streaming Framework does join operations on batch and real-time events of user purchase and demographic type.
- MySql Holds the demographic data such as age, gender, country, etc.
- Spark Streaming Framework consumes these events and generates a variety of points suitable for time series and dashboarding.
- Kafka connect pushes the events from the Kafka streams to influxDB.
- Grafana connects to different sources like influxDB, MySQL and populates the graphs.

Key Takeaways

- Create a full-fledged real-time low latency Spark-Streaming jobs.
- Understanding basics of Spark Streaming
- Understanding basics of Kafka and producing/consuming from topics.
- Introduction to influxDB and its use cases
- Integrate influxDB and Grafana for a dashboard.
- Core concepts of real-time streaming, time-series databases.
- Introduction to Docker.
- Using docker-compose and starting all tools.
- Troubleshooting issues related to software versions, software setup in local, choosing correct libraries.
- Code Walkthrough.
- Exploring Tools via UI or CLI.
- Finetuning frameworks using configuration parameters.

High-Level Design