Real-time Streaming Platform using Apache Flink and Kinesis

Business Overview

Perishable data is information whose value can significantly decline over a period of time. When the operating conditions change to the point that the knowledge is no longer helpful, the information loses a large amount of its value. This information is frequently produced in an IoT computing context. Managing perishable data provides a number of benefits. Real-time data provides you with the information you need almost instantly and in the context which helps make smarter business decisions. The use of data from edge computing devices speeds up reaction time and minimizes time to action. Such data does not need to be kept in data centers for a long time and maybe deleted after usage, saving storage space and money. Systems are usually designed on constant patterns and can't scale with peaks in the load resulting in latency and server failures.

This Project will simulate real-time accidents data and architect a pipeline that will help us analyze and take quick actions using AWS Kinesis, Apache Flink, Grafana, and Amazon SNS.

Data Pipeline

A data pipeline is a technique for transferring data from one system to another. The data may or may not be updated, and it may be handled in real-time (or streaming) rather than in batches. The data pipeline encompasses everything from harvesting or acquiring data using various methods to storing raw data, cleaning, validating, and transforming data into a query-worthy format, displaying KPIs, and managing the above process.

Dataset Description

This Project uses the US car accidents dataset which includes a few of the following fields:

- Severity
- Start_Time
- End_Time
- Location
- Description
- City
- State

Tech Stack:

➔ Languages-

- SQL, Python3

➔ Services -

- AWS S3, AWS Glue, AWS Athena, AWS Cloud9, Apache Flink, Amazon Kinesis, Amazon SNS, AWS Lambda, Amazon CloudWatch, Grafana, Apache Zepplin

Amazon S3

Amazon S3 is an object storage service that provides manufacturing scalability, data availability, security, and performance. Users may save and retrieve any quantity of data using Amazon S3 at any time and from any location.

AWS Glue

A serverless data integration service makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development. It runs Spark/Python code without managing Infrastructure at a nominal cost. You pay only during the run time of the job. Also, you pay storage costs for Data Catalog objects. Tables may be added to the AWS Glue Data Catalog using a crawler. The majority of AWS Glue users employ this strategy. In a single run, a crawler can crawl numerous data repositories. The crawler adds or modifies one or more tables in your Data Catalog after it's finished.

AWS Athena

Athena is an interactive query service for S3 in which there is no need to load data, and it stays in S3. It is serverless and supports many data formats, e.g., CSV, JSON, ORC, Parquet, AVRO.

Grafana

Grafana is a web application for interactive visualization 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.

Apache Flink

Flink is a scalable data analytics platform and distributed processing engine. Flink may be used to handle massive data streams and give real-time analytical insights about the processed data to your streaming application. Flink is built to work in a variety of cluster setups, with in-memory calculations of any size. For distributed computations over data streams, Flink also offers communication, fault tolerance, and data distribution. Flink applications use unbounded or bounded data sets to process streams of events. Unbounded streams have no fixed termination and are handled indefinitely. Bounded streams have a defined beginning and endpoint and may be handled in batches.

Amazon Kinesis

Amazon Kinesis Data Streams is a real-time data collection and processing service from Amazon. Kinesis Data Streams apps are data-processing applications that may be created. Kinesis Data Firehose is part of the Kinesis streaming data platform, which also includes Kinesis Data Streams, Kinesis Video Streams, and Amazon Kinesis Data Analytics. When using Kinesis Data Firehose, the user does not need to develop apps or manage resources. Configure the data producers to send data to Kinesis Data Firehose, and the data will be automatically transferred to the specified destination. Kinesis Data Firehose may also be used to transform data before delivering it.

Key Takeaways

- Understanding the project Overview and Architecture
- Understanding ETL on Big Data
- Introduction to Staging and Data Lake
- Creating IAM Roles and Policies
- Understanding the Dataset
- Setting up AWS CLI
- Understanding Data Streams and Amazon Kinesis
- Understanding Apache Flink
- Creating a Kinesis Data Analytics Application
- Working with Apache Zepplin Notebooks
- Create Athen Tables using Glue Data Catalog
- Creating Lambda Function to send SNS Notifications
- Using CloudWacth and Grafana for Visualization

Architecture