Analytical Platform for eCommerce Store using AWS

Business Overview

Ecommerce analytics is the process of collecting data from all of the sources that affect a certain shop. Analysts can then utilize this information to deduce changes in customer behavior and online shopping patterns. Ecommerce analytics spans the whole customer journey, from discovery through acquisition, conversion, and eventually retention and support.

In this project, we will use an eCommerce dataset to simulate the logs of user purchases, product views, cart history, and the user's journey on the online platform to create two analytical pipelines, Batch and Real-time. The Batch processing will involve data ingestion, Lake House architecture, processing, visualization using Amazon Kinesis, Glue, S3, and QuickSight to draw insights regarding the following:

- Unique visitors per day
- During a certain time, the users add products to their carts but don't buy them
- Top categories per hour or weekday (i.e. to promote discounts based on trends)
- To know which brands need more marketing

The Real-time channel involves detecting Distributed denial of service (DDoS) and Bot attacks using AWS Lambda, DynamoDB, CloudWatch, and AWS 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 dataset contains user behavioral information from a large multi-category online store along with fields such as event_time, event_type, product_id, price, user_id. Each row in the file represents one of the following event types:

- View
- Cart
- Removed from Cart
- Purchase

Tech Stack:

➔ Languages-

- SQL, Python3

➔ Services -

- AWS S3, AWS Glue, AWS Athena, AWS Cloud9, Apache Flink, Amazon Kinesis, Amazon SNS, AWS Lambda, Amazon CloudWatch, QuickSight, Apache Zepplin, Amazon DynamoDB, AWS Glue DataBrew

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.

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.

QuickSight

Amazon QuickSight is a scalable, serverless, embeddable, machine learning-powered business intelligence (BI) service built for the cloud. It is the first BI service to offer pay-per-session pricing, where you only pay when your users access their dashboards or reports, making it cost-effective for large-scale deployments. It can connect to various sources like Redshift, S3, Dynamo, RDS, files like JSON, text, CSV, TSV, Jira, Salesforce, and on-premises oracle SQL-server.

AWS Glue DataBrew

AWS Glue DataBrew is a visual data preparation tool that allows users to clean and normalize data without writing any code. When compared to custom-created data preparation, DataBrew helps minimize the time it takes to prepare data for analytics and machine learning. Many pre-built transformations are available to automate data preparation activities such as screening anomalies, transforming data to standard formats, and rectifying erroneous values.

Amazon DynamoDB

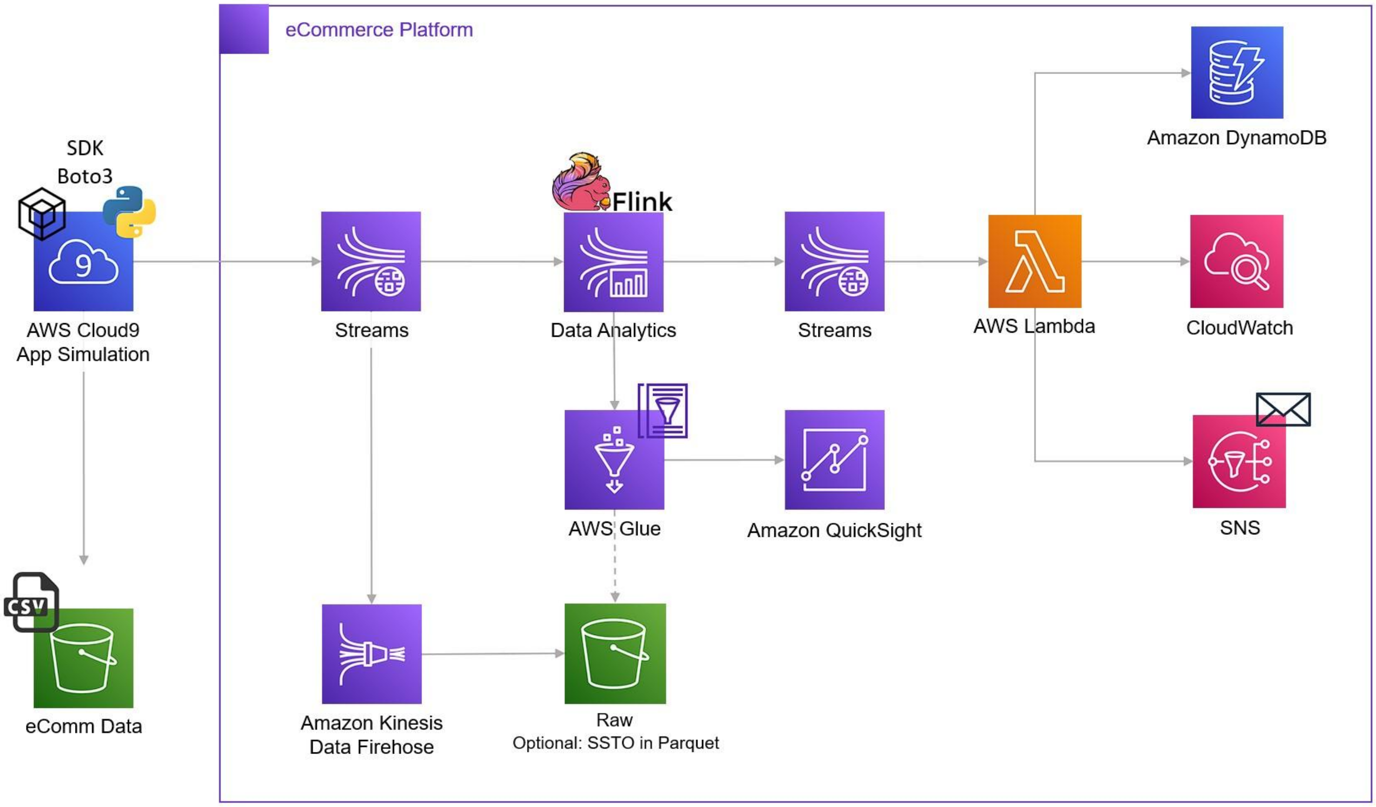
Amazon DynamoDB is a fully managed key-value NoSQL database service that delivers quick and predictable performance while also allowing for seamless scaling. DynamoDB relieves developers of the administrative responsibilities associated with running and growing a distributed database. DynamoDB allows you to design database tables that can store and retrieve any quantity of data while also serving any degree of request volume. You may increase or decrease the throughput capacity of your tables without experiencing downtime or performance reduction. Amazon DynamoDB supports PartiQL, an open-source SQL-compatible query language that enables efficient data querying independent of where or how it is stored.

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
- Using Glue and Athena to define Partition Key
- Understanding Lambda Functions
- Creating Lambda function for DynamoDB and SNS
- Understanding DynamoDB Data Modelling
- Integrating Lambda and Kinesis
- Performing ETL for Parquet format using Glue DataBrew and Spark
- Creating QuickSight Dashboards

Architecture


eCommerce Platform
SDK
Boto3
AWS Cloud9
App Simulation
Streams
Flink
Data Analytics
Streams
AWS Lambda
Amazon DynamoDB
CloudWatch
SNS
AWS Glue
Amazon QuickSight
CSV
eComm Data
Amazon Kinesis
Data Firehose
Raw
Optional: SSTO in Parquet