Anomaly Detection in Cloud Servers

Team 4 SLACKERS
Unnati Aggarwal (002741568)
Shantanu Sachdeva (002748942)

WHY?

The purpose of this project is to leverage telecom network logs to identify and flag anomalous activities, ultimately enhancing cybersecurity threat detection. By analyzing these logs, the project aims to develop algorithms and models that can automatically detect patterns indicative of potential threats or abnormalities within the network. This proactive approach can help network administrators and security teams to quickly identify and respond to potential cybersecurity incidents, thereby strengthening the overall security posture of the telecom network.

Use Case

- Network Administrator inputs stream of telecom server logs to the pub/sub topic to detect outliers.

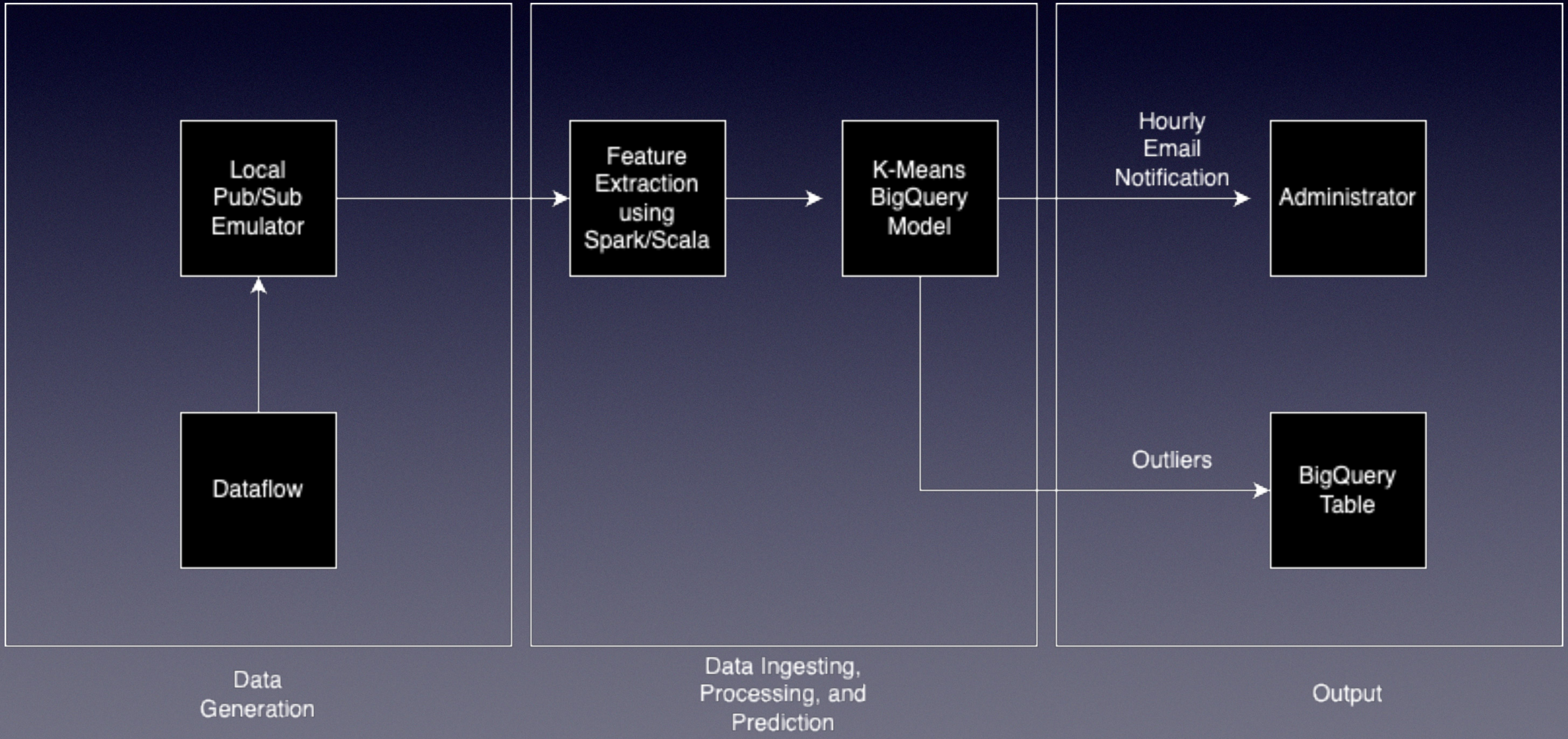
- Server Health Check software runs a batch of input logs hourly to detect a spike in outliers and emails the generated report to the administrator.

Methodology

Machine learning-based anomaly detection solution by highlighting the following key components:

1. Generating synthetic data to simulate production volume using Dataflow and Pub/Sub.
2. Extracting features and real time processing using Scala/Spark.
3. Training and normalizing data using BigQuery ML's built-in k-means clustering model.

Methodology Diagram



Data Sources

- Initially, we plan to use a pipeline of simulated data using DataFlow.
- Eventually, we aim to use live data streams.

Milestones

- 21st March: Establishing log schemas and mocking data to the Pub/Sub in real time. Establishing the feature extraction criteria on aggregated logs using Scala/Spark
- 28th March: Creating the K-Means model using BQ ML
- 4th April : Testing with outlier mock date to see if the model is able to detect it
- 11th April : Setting up output email notification
- 18th April : Final Presentation

The Big Question, Scala

- Data Ingestion
- Feature Extraction using Spark

Acceptance Criteria

- Data loss during ingestion from pub/sub topic should be less than 10%
- Ensure the prediction model has a minimum accuracy of 90% in identifying outliers

Goals

- Getting hands-on and understanding realtime or streaming analytics.
- Working with Google managed services like Dataflow, BigQuery ML
- Writing efficient and Scalable code in Scala