Project

- **Team number:** 6
- **Project title:** Real-Time Web Server Log Processing
- **Project description:**

  Building a data intensive pipeline to process web-server log data and provide feedback, take necessary infrastructure actions and provide visual analysis. This helps engineers to take necessary actions with their deployed applications on servers and also automate certain infrastructure horizontal scaling real-time.
- **Team members:**

Daxkumar Amin	dkamin
Khantil Choksi	khchoksi
Riken Shah	rshah9

Deliverables

1. **Data Generation Module**
- Consists of multiple producers which generate data at high velocity
- Logs data has attributes like IP, region, timestamp, API Endpoint, Response Status, Response Time
- Producers are designed based on configurable parameters to simulate velocity, volume and variety

2. **Data Streaming and Processing Module**
- Build data streaming pipeline using AWS Kinesis, which ingests data from multiple sources and distributes among different shards based on partition-key
- Configure DynamoDB to store data coming from multiple consumers, which persists data based on time-window
- Design Spark based processing module on EC2 instance

3. **Visualization and Monitoring Dashboard**
- Develop metrics like charts for real-time monitoring incoming log data. E.g. Response time for each end-point

Status

- Developed producers to simulate real-time high velocity web server log data with configurable attributes and insert it into Kinesis Stream using boto3 (python library)
- Create reproducible & scalable infrastructure with Kinesis, EC2, DynamoDB, IAM Roles, security groups using AWS CloudFormation
- Working on consumers to fetch and store data in DynamoDB for further processing
- Researching best possible implementations of Spark on EC2 and connect with DynamoDB
- Working on implementation of visualization and feedback module.

Issues

- Configuring Spark to spin on EC2 and successfully fetch data from DynamoDB
- Configuring optimum number of shards to make Kinesis stream scalable and auto-load-balancing
- Manage configuration parameters automatically based on the feedback loop
- Optimizing elastic storage of Kinesis stream to save on cost and improve efficiency.
- Generating live graphs based on incoming logs from the stream.