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**
   Data generation module will use the base data set of web logs and with some random sampling continuously keep on generating data at high speeds (based all configurable parameters). This will be used for the ingestion of the data by the data processing module.
2. **Data Processing Module**
   The main component of the data processing module is the multi-level fault-tolerant distributed queue system (kafka/kinesis) which can scale easily. There is also an analysis module associated which will carry out the processing like Apache Flink / Spark.
3. **Monitoring Dashboard**
   The dashboard monitoring module is where all the aggregated information as well as notifications can be visualized.

Dependencies

1. Data - Web log data set, 26k rows
2. Data Storage - S3 bucket
3. Data streaming platform - Apache Kafka / AWS Kinesis
4. Data processing framework - Apache Spark / Flink
5. Deployment environment on Cloud - AWS EC2 / VCL
6. Programming resources - Python3, standard py libraries
7. Pipeline trigger (for feedback) - AWS lambda

Note: These dependencies are subject to change.

Issues

1. Generating the data for high velocity
2. Setting up the environment with required technology and framework on VCL and/or AWS
3. If we decide to build data streaming pipeline with AWS Kinesis rather than Apache Kafka, then configuring Apache Spark / Flink processor over it can require more effort.
4. As of now we have decided to use Python 3.6 for most of our work. We will have to use to some third-party libraries / interfaces for Apache framework, as they support APIs mainly in Java and C.