Team H: User Query Behavior Monitor - Finance Data Warehouse

- **Team members and roles**
    - Ji Wen (jw2537) - Data Visualization Engineer
    - Franklin Zhao (qz297) - Data/ML Engineer
    - Ankur Biswas (ab2249) - Data Scientist/Product Manager

- **Key goals**
    - Process the system logs and load them into Spark so that the data can be iterated over
    - Find reports that do not have the proper constraints to reduce the returning data volumes
    - Find reports run by various controller within the same teams with similar output; Try to determine if they should be combined wherever possible
    - Find users running multiple concurrent reporting sessions
    - Devise and implement a generic algorithm that will highlight odd behaviors in the plant in terms of usage pattern
    - Create a mechanism to perform the analysis on a slice of traffic as desired (e.g., business hour, business day, business week)
    - Build a dashboard so that the client can view results in real time and also compare usage patterns over specific time periods.

- **Deliverables**
    - Devise mechanisms and algorithms to identify unnatural and unwanted usage patterns in the data on a slice of traffic as desired
    - Determine a mechanism to persist the output of the result set
    - Represent the process and results using tableau - either using a report or a continually updating dashboard.

- **Major accomplishments**
    - Cleared initial legal considerations with client.
    - Narrowed scope of the project and final deliverables.
    - Finished getting up to speed with learning curve: learning tableau and getting ready for data visualization
    - Mocked up 10,000 rows of data (with only 7 featured fields suggested by the client)
    - Ran initial set of unsupervised ML algorithms on the data and visualized the results as shown in Figure 1-4. Got feedback from client on the same.

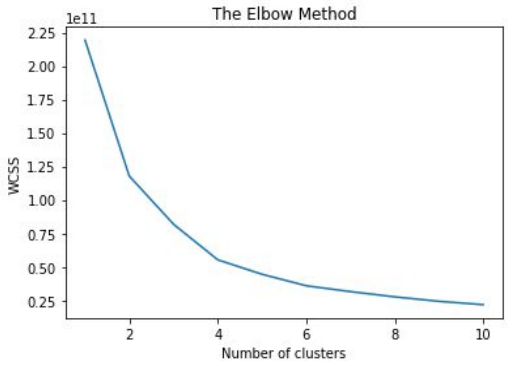

Fig.1. Result of the Elbow Method Analysis

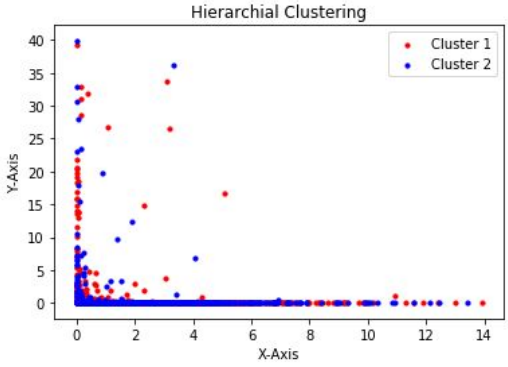

Fig.2. Result of the Hierarchical Clustering

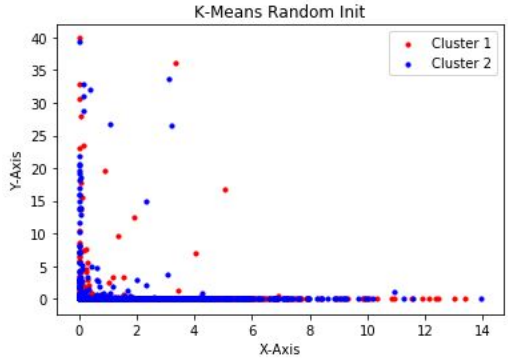

Fig.3. Result of the K-Means Clustering

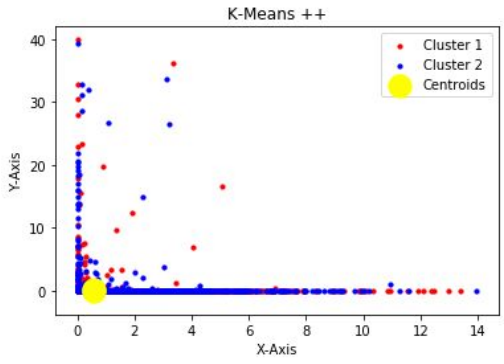

Fig.4. Result of the K-Means++ Clustering

- **Environment (Data, tech stack, tools)**
    - For data analysis: Python - Libraries including numpy and pandas; Spark.

- For machine learning models: Python - Libraries including numpy, sklearn and scipy. Will use more if required.
- For data visualization: Tableau.
- For version control: GitHub and Google Drive
- For Communication: Facebook messenger (intra team), Phone call, email and Zoom (with client).

- **Issues and Risks**
  - **Requirement may be late or incomplete**

    Because of NDA problems, we started to mock the data by ourselves since last week. It takes time to identify the important fields, to clarify how detailed they want the data analysis to be, and to learn about the type of distribution, mean, median and standard deviation in the real data. Still, we are not sure if our data is good enough for next stage, which may affect overall progress of work.

  - **Mock data may not accurately reflect the patterns in real data**

    We have basic idea of how real data looks like but the information we obtain is very limited. We are still working on improving our mock data, so it is a major problem we have right now.

  - **Client may not like the way we visualize the data**

    We have not showed any visualization to our client yet, so it is a potential risk we may encounter shortly.

- **Artifacts**
  - Project plan
  - Mock Data
  - Source Code (DataGenarator.py, clusteringAlgorithms.py, hierachial.py, kmeans.py, utils.py)