Sample Group Project
ELK (Elasticsearch, Logstash, Kibana)

Introduction

In this project, we will be working with NYC OpenData published by the city of New York pertaining to 311 service requests collected since 2010 with over 36 million rows with 41 columns. https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9.

Project Objectives:

1. To further expose you to using ELK stack as an analytic tool to analyze streaming realistic big data.
2. To demonstrated experience working with opened end problems, that are similar to problems that you we will face in our career as a Machine learning with big data professional.
3. At the end of this project we will:
    a. Gain sufficient confidence in creating Logstash configuration files and creating Elasticsearch indices, advanced queries, charts, maps and dashboards using Kibana i.e. fully using ELK stack in real big data scenarios.
    b. Gain an appetite for working with large streaming datasets.
    c. Be aware of the potential and benefits of analyzing large streaming datasets using big data tools.

Our Expectations

1. Demonstrate good understanding of the features that ELK stack provides, such as advanced queries, indexing, manage tables, aggregations, charts/graphs, maps and dashboards.
2. Demonstrate our ability to work with multiple large datasets and use the tools to gain valuable insights from the datasets. As well, you should be able to present these insights in a manner that is easily consumable by stakeholders and other interested parties.

Problem Background

We've been hired as Data Scientists by the city of New York to gain valuable insights from their huge data set for 311 service requests. Your task is to use the ELK stack in GCP platform. Successful completion of this task includes creating a ***Logstash configuration file*** as well as a ***geo-point template (for maps)***, creating a GCP instance and firing Logstash to ingest the NYC 311 service requests data into Elasticsearch and using Kibana to analyze and visualize the results as per the questions given.

Expected Results:

Code for your Logstash configuration file and ***geo-point template***, ***results for the analytical questions (tables, charts, tag clouds, maps and dashboard)*** in MS Word or PDF document. Where applicable, show the ***syntax/code*** or capture ***screenshots*** for all our analysis.

Analytical Questions

1. Create a table showing the top 10 cities with the highest calls alongside the count of top 10 complaint calls (by Descriptor) in each city.
2. Create a pie chart showing the top 5 cities with the highest calls alongside the top five calls (Descriptor) in each city
3. Create a tag cloud representing the top 20 call descriptors.
4. Create a coordinated map of all the major call descriptors in each city
5. Create a dashboard for all visualizations of 1to 4 above.