Big Data Project Proposal

Team: Nandish Bhagat (A20490179), Taranpreet Bhatia (A20460031), Kranti Mahadev Avhad(A20473187), Jasleen Kaur Bhatia(A20495939)

Topic: Customer Churn Prediction on Sparkify Dataset

Overview: The project mainly focuses on predicting customer churn(detect if the specific customer will cancel the service) on the Sparkify Dataset (mini dataset – 128 Mb & full dataset – 12 Gb). The dataset is composed of numerous user events in the audio streaming service provider like Spotify and Pandora. At first, we will analyze the mini dataset using spark Dataframe, Spark SQL & Spark ML API on local using PYspark and then deploy the spark cluster on AWS to run the model on full dataset. Also, we will use Tableau to visualize the insights from the dataset.

Tools:

We are using two tools to explore our data set and derive insights from and visualize it:

1. Apache spark
2. Tableau

Apache spark is a framework that is used to process data and perform numerous tasks on a large scale. It is also used to process data via multiple computers with the help of distributing tools. It is widely used among data analysts as it offers easy-to-use APIs that provide easy data pulling methods and it is capable of handling multi-petabytes of data as well.

Tableau is an end-to-end, powerful, secure, flexible, interactive visual analytics software that is used in Business Intelligence industries. Tableau connects data source to server such as Microsoft Excel, text files, JSON files, PDF files, special files, statistical file or web-based data, and creates result in the form of data visualizations, maps, dashboards etc. Generated results sheet can be shared with the client via social media. These generated files are downloaded in different formats such as tableau workbook or tableau package workbook.

References:

1)Shoro, Abdul & Soomro, Tariq. (2015). Big Data Analysis: Apache Spark Perspective. Global Journal of Computer Science and Technology. 15.

2) Khedikar, Kanchan A., Data Analytics for Business Using Tableau (April 27, 2021). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021, Available at SSRN: https://ssrn.com/abstract=3835030 or http://dx.doi.org/10.2139/ssrn.3835030