# Data Science Project Proposal

1. **Principal Investigator**

Akhil Reddy Buchala – [**buchalaa1@newpaltz.edu**](mailto:buchalaa1@newpaltz.edu)Rohan Reddy Yeruva – [**yeruvar1@newpaltz.edu**](mailto:yeruvar1@newpaltz.edu)Chandra Sekhar Pokuri – [**pokuric1@newpaltz.edu**](mailto:pokuric1@newpaltz.edu)

# Title of Project

Customer Churn Prediction with AWS EMR

# Mentoring

Professor Min Chen, Department of Computer Science, SUNY-New Paltz

[**chenm@newpaltz.edu**](mailto:chenm@newpaltz.edu)

# Objective of Project

Customer Churn prediction is one of the most applied Big Data and Data Science use cases in business. Its aim is to predict the customers that are likely to discontinue the use of any paid service/product. For a predicting churn, historical customer data is analyzed and various machine learning algorithms are created to find customer’s likelihood to churn. So with huge amount of available data to be applied to train various models will require a huge compute and storage capacity. So Hadoop-like technology, AWS EMR is used to analyze this use case. Hadoop provides huge storage, processing power and the ability to handle huge number of concurrent tasks or jobs.

# Background

Churn prediction is a method of analyzing the data to predict which customers are likely to discontinue with the usage of any paid service/product. Method predicts the probability that customer will churn.

For a company, churn can be due to various reasons. It depends of factors like customer profile, behavior and their needs. So understanding churn requires a segmentation of customers.

With help of churn prediction, business can try to engage customers before they churn. Various business strategies can be applied based on churn analysis.

# 6. Approach to the Study

The data used is a customer user logs data originated from customer interaction with an imaginary music streaming website. The data size is around 12GB, was created by Udacity and is publicly available.

The analysis will done in an AWS EMR notebook using PySpark.

The first step is to understand the data columns. The data will contain some null values and some un important columns. So data cleaning and transformations will done like clean the data by dealing with null values, dropping columns, etc.

After which we will try to implement various ML Supervised Classification algorithms like Logistic Regression, Random Forest Classifier, etc. and compare the accuracies. Because we have to classify, a Classification algorithm will be used.

# Initial timeline for implementation

|  |  |
| --- | --- |
| 1 Week | Literature review |
| 1 Week | Data Analysis, Data Cleaning, Data preparation |
| 1 Week | Implementing various ML algorithms |

**8. References**

1. Data reference: “s3n://udacity-dsnd/sparkify/sparkify\_event\_data.json”