**CS5590 BigData Programming - Project Fall 2018**

**Project Title:**

1. Web Traffic anomalies and Web Traffic Prediction

2. Credit Loan risk analysis using Spark Machine learning.

**Objective**

The project has two part. In first part we will be using the server logs from renowned website defect and report web traffic anomalies r using Spark, Graph and Cassandra.

For second part we will be using the dataset from UCI data library and perform Classification. Using Spark and Machine learning library.

**Motivation:**

Edmund.com Forecast Model

**Significance:**

It helps in managing the 24/7 availability of website and serve customer better and faster way.

**Feature:**

Spark, graph analysis and real time dashboard for good and bad http request.

**Reference:**

http://blog.cloudera.com/blog/2015/03/how-edmunds-com-used-spark-streaming-to-build-a-near-real-time-dashboard/

**Datasets:**

**Web Logs:**

http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html

**Credit Loan:**

[https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)](https://archive.ics.uci.edu/ml/datasets/statlog (german credit data))

**Increment 1: Part 1 (Web Traffic Anomalies)**

**Dataset:**

**Web Log Analysis:**

http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html

**Detail design of Features:**

In the First we have downloaded the dataset from NASA web server for using in our project.

**Analysis:**

We have done the analysis of dataset and found out that the dataset is in below format.

The logs are an ASCII file with one line per request, with the following columns:

1. **host** making the request. A hostname when possible, otherwise the Internet address if the name could not be looked up.
2. **timestamp** in the format "DAY MON DD HH:MM:SS YYYY", where DAY is the day of the week, MON is the name of the month, DD is the day of the month, HH:MM:SS is the time of day using a 24-hour clock, and YYYY is the year. The time zone is -0400.
3. **request** given in quotes.
4. **HTTP** reply code.
5. **bytes** in the reply.

**Implementation:**

At this stage no implementation was performed for Web Log Analsysis

**Preliminary Results:**

Dataset need to be cleaned up properly for further analysis.

**Project Management**

**Implementation status report**

**Work completed:**

**Description:**

Web log Dataset has been loaded and format and feature has been identified/extraction.

**Responsibility (Task, Person)**

Downloading and analyzing Dataset: Raju Nekad**i**

**Contributions (members/percentage):**

Raju Nekadi 50%

Sushma Manne 50%

**Work to be completed:**

**Description**

Flume Setup

Reading Dataset from Flume and loading to Spark

Spark Analysis generating result

Loading dataset in Cassandra/Hive

Graph Generation

**Responsibility (Task, Person)**

Flume Setup/Raju Nekadi

Reading Dataset from Flume and loading to Spark/Raju Nekadi

Spark Analysis generating result /Raju Nekadi

Loading dataset in Cassandra/Hive Sushma Mane

Graph Generation/Sushma Mane

**Issues/Concerns:**

None

**References/Bibliography:**

https://blog.cloudera.com/blog/2016/06/how-to-detect-and-report-web-traffic-anomalies-in-near-real-time/

**Part 2: (Predicting loan credit risk using Apache Machine Learning)**

**Data Set:** data set has alphanumeric values in it, which has to be converted to numeric values.



**Design:**

Apply model

Extract Features

Load data

**Analysis:**

Below are the features of the data.

Features → {"balance", "duration", "history", "purpose", "amount", "savings", "employment", "instPercent", "sexMarried", "guarantors", "residenceDuration", "assets", "age", "concCredit", "apartment", "credits", "occupation", "dependents", "hasPhone", "foreign"}

**Implementation:**

This Project is implemented using Apache Spark Machine Learning

**Preliminary Results:**

Dataset need to be cleaned up properly for further analysis.

**Project Management**

**Implementation status report**

**Work completed:**

**Description:**

Web log Dataset has been loaded and format and feature has been identified/extraction.

**Responsibility (Task, Person)**

Downloading and analyzing Dataset: Chandra sekhar Pentakota

**Contributions (members/percentage):**

Chandra sekhar Pentakota50%

Bilal Mustafa 50%

**Work to be completed:**

**Description**

Apache Spark Setup

Reading Dataset and converting to Data Frame

Extract Features

Spark Analysis generating result

Training and Testing Model

**Responsibility (Task, Person)**

Apache Spark Setup / Chandra sekhar Pentakota

Reading Dataset and converting to Data Frame / Bilal Mustafa

Extract Features / Bilal Mustafa

Spark Analysis generating result / Chandra sekhar Pentakota

Training and Testing Model / Chandra sekhar Pentakota

**Issues/Concerns:**

None

**References/Bibliography:**

https://mapr.com/blog/predicting-loan-credit-risk-using-apache-spark-machine-learning-random-forests/