CS 532 - Project Milestones

Danve, Riya
rdanve@umass.edu

Deshpande, Riya
riyarajeshde@umass.edu

Bodtman, Lucy
lbodtman@umass.edu

Elmali, Aylin
aelmali@umass.edu

2023-11-10

1 Project Overview

1.1 Project Title: Big Data Healthcare Analytics

1.2 Key Objectives

- Data Integration: Collect and ingest diverse healthcare data sources. Use Hadoop HDFS to store the dataset
- Data Quality: Clean, transform, and normalize data.
- Analytics: Utilize Spark for performing the required analysis on the stored dataset.
- System Perfomance Analysis: Analyze system performance through parameters like runtime, response time, and throughput.
- Visualization: Develop a visual interface using Tableau/PowerBI for gaining better insights.

2 Milestone Goals and Progress

- Preprocessing the dataset to clean, transforma and normalize the data using Python. **COMPLETED**
- Ingest data and store into Apache Hadoop HDFS. **COMPLETED**

- Utilize Spark to run primary analytics on the dataset to find out the correlation between various factors for predicting preganancy.
  INCOMPLETE

  We were not able to use Spark to run primary analytics on the dataset so far. We have figured out what data we want to analyze, and have stored the data on Hadoop. We have also successfully installed Spark and are working on integrating Hadoop and Spark. However, we ran into some technical difficulties while setting up Hadoop which delayed the process. Because of this, we have not started with the analytics yet, however now that the other steps are completed we plan to begin analysing the data with Spark.