CS 532 - Project Proposal

Danve, Riya
rdanve@umass.edu

Deshpande, Riya
riyarajeshde@umass.edu

Bodtman, Lucy
lbodtman@umass.edu

Elmali, Aylin
aelmali@umass.edu

2023-10-27

1 Project Overview

1.1 Project Title: Big Data Healthcare Analytics

Pregnancy is a challenging phase in a woman's life. Predicting the outcome of a pregnancy (succesful/unsuccesful/abnormalities/healthy) may also be aided by a number of social and economic factors. The outcome of the pregnancy affects the mother's health in addition to the life of the newborn. Due to inadequate care, many women, particularly in developing or underdeveloped nations, experience several problems during pregnancy, which can have long-term effects on their physical and mental health. In order to reduce risks, prevent unintended pregnancies, and maintain the health of both the mother and the newborn child, it is crucial to analyze these diverse aspects, their degree of correlation, and their significance.

The dataset (link included below) includes information on pregnancy outcomes, maternal health during pregnancy and the prenatal period, and different indices of household economic conditions from surveys conducted in nine Indian states. Analyzing these many variables will help identify the root causes of pregnancy problems as well as the effects of related physical, psychological, and socioeconomic variables on the course of the pregnancy.

Through this project and the associated analysis, we also aim to measure variances in various system performance factors and process with the changing system resources or other parameters. It includes gathering, cleaning, storing, analyzing, and, if desired, presenting data through an interface for visualization. We plan to use Hadoop with Spark for this performance analysis.

1.2 Dataset:

The dataset contains information on the outcomes of pregnancy (Live birth/still birth/abortion), maternal health history and habits, family planning methodologies, availability of primary healthcare, etc. surveyed for women from 9 Indian states. India being a developing country sees a lot of variation in the availability of resources, socio-economic conditions and overall health awareness across its various states and can prove to be a good starting point for running such analytics. These nine states, which account for about 48 percent of the total population, 59 percent of births, 70 percent of infant deaths, 75 percent of under 5 deaths and 62 percent of maternal deaths in the country, are the high focus states in view of their relatively higher fertility and mortality. The dataset can be accessed using the below URL:

Dataset Link: `https://www.kaggle.com/datasets/rajanand/ahs-woman-1`

1.3 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 Performance Analysis: Analyse system performance through parameters like runtime, response time and throughput.
- Visualization (Optional goal): Develop a visual interface using Tableau/Power Bi for gaining better insights.

2 Milestone Goals

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

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

3 Final Project Goals

- Perform any additional data analysis based on the earlier outcomes.
- Carry out system performance analysis by varying the parameters as follows:
    - Effect of increasing/decreasing the number of cores and partitions on the runtime.
    - Calculating throughput with respect to the increase in the number of records accessed.
    - Implementing query optimization to see its effect on the runtime/CPU time
- (Optional) Developing visualization interface for gaining better insights of the data and measuring the data retrieval speed by observing the dashboard rendering rate to analyse the system performance for varying values of cores, number of records, etc.