Healthcare Analytics

# Naveenkumar Nekkalapu

Data Science CU Boulder

Boulder, CO, USA [nane9419@colorado.edu](mailto:nane9419@colorado.edu)

Vidit Vivek Sharma

Data Science

CU Boulder

Boulder, CO,USA [vish4271@colorado.edu](mailto:%20vish4271@colorado.edu)

Mahin Kadabi

Data Science

CU Boulder

Boulder, CO, USA [maka1640@colorado.edu](mailto:maka1640@colorado.edu)

## ABSTRACT

Hospitals, clinics, and healthcare companies are facing challenges in understanding the distribution of the customers and the different types of diseases they treat. The vast amount of patient healthcare data which add to a large amount of cloud storage requires a huge number of resources to process them. A single CT/MRI image takes up-to terabytes of storage data. Taking the help of the Big Data Ecosystem to analyze the patients will help the hospitals understand the vast number of patients.

The project's objective is to build machine learning data pipelines for hospitals so that they may analyze patients and the type of diseases they are facing or might be facing in the future. This will help them to treat the patients before time and will also help them in utilizing a robust model.

## INTRODUCTION

The first question is why we need big data to analyze customer behavior. The question is valid: why is there a need for Big Data? The answer is simple: Organizations have the opportunity to harness their data and use big data analytics to find new opportunities. This results in wiser company decisions, more effective operations, greater profitability, and happier clients. Using Spark's simple-to-use APIs, there is a huge ability to work with huge datasets. Using tools like Hadoop and Spark offer significant cost advantages for storing, processing and analyzing large volumes of data. It consists of a set of more than 100 data transformation operators as well-known data frame APIs(Spark APIs) for working with semi-structured data. Various techniques can be implemented to understand how healthcare companies utilize this information to increase revenues, cut costs, improve customer relationships, and reduce risks and more.

**2 Techniques Used**

Google Cloud Platform Services(GCP), Apache Hadoop,Cloud SQL, Apache Spark, Tableau, PySpark, SQL Query

**3 Related Work**

Healthcare data analytics, hospitals and clinics are looking for new data mining techniques that will suite evolution of information technology and help analyze a large amount of complex data. Certain types of data visualization techniques are being studied. One of the recent advancements in this field is advanced data visualization.

Advanced data visualization is different from other standard bars and line chart, since it can scale its visualizations for millions of data points and handle different data types.

Another related work is the Hadoop distributed file system. HDFS enhances healthcare data analytics system by dividing large amount of data into smaller one and distributed it across different systems.

**4 Components Utilized**

**4.1 Apache Hadoop**

Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. It offers a software framework for the MapReduce programming model-based distributed storage and processing of massive data. Using this technique, the parallel processing of enormous data sets is achieved and is sustainable.

Graphical user interface, application

Description automatically generated

Fig 1: Hadoop Architecture

**4.2 Google Cloud Platform Services**

Google cloud platform provides a web-based, graphical user interface that can be used to manage Google Cloud projects and resources. Google Cloud consists of a set of physical assets, such as computers and hard disk drives, and virtual resources, such as virtual machines (VMs), that are contained in Google's data centers around the globe.

**4.3 Apache Spark**

Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters. Spark is a unified analytics engine for large-scale data processing. It provides high-level APIs in Scala, Java, Python, and R, and an optimized engine that supports general computation graphs for data analysis. It also supports a rich set of higher-level tools including Spark SQL for SQL and DataFrames, pandas API on Spark for pandas workloads, MLlib for machine learning, GraphX for graph processing, and Structured Streaming for stream processing.

Diagram

Description automatically generated

Fig 2: Main components of Apache Spark

**4.4 Cloud SQL**

Fully managed relational database service for MySQL, PostgreSQL, and SQL Server with rich extension collections, configuration flags, and developer ecosystems.

**4.5 Tableau**

Tableau is a visual analytics platform transforming the way businesses use data to solve problems hence empowering people and organizations to make the most of data.

**5 Methods**

There are several different methods which are used in this project. The first step in this project is to gather and collect the dataset. The dataset taken for the project is called Healthcare stroke data and it was downloaded from kaggle. The dataset has 5000 rows and many columns including id, gender, age, hypertension, average blood sugar levels, etc. There are 12 columns in all.

Diagram

Description automatically generated

Fig 3 Architecture implemented

The above picture depicts the architecture used in our project. The data is taken from an external source. Then we use GCP to create the dataproc cluster and the buckets. The status of the dataproc cluster is that is running. There is a unique UUID associated with a cluster.

Graphical user interface, text, application

Description automatically generated

Fig 4 Dataproc cluster created

The buckets are used to upload the dataset to the GCP. The bucket is named as bigdataproject123.

Graphical user interface, text, application, email

Description automatically generated

Fig 5 Bucket

The next step is that we create a SQL instance. The SQL instance created is MySQL instance with database version “MYSQL\_5\_7”. The zone and activation policy is specified when creating the instance. This is implemented in the GCP console while connected to the instance of the project. This MySQL database created will be further used in our project.

Table

Description automatically generated

Fig 6 Database Instance created

There are different users which are authorized to access the MySQL database.

Graphical user interface, text, application

Description automatically generated

Fig 7 Database Users

The bucket details created for importing the dataset is shown in the below picture.

Graphical user interface, text, application, email

Description automatically generated

Fig 8 Dataset imported in Bucket

To connect to the Cloud SQL from a particular source a public IP is used which requires an authorized network or the cloud proxy instance to connect to this instance.

Graphical user interface, text, application

Description automatically generated with medium confidence

Fig 9 Authorized Network

To connect to the database a public IP address is used with a specific connection name.

Graphical user interface, text, application, email

Description automatically generated

Fig 10 Connected to SQL Instance

After these tasks are completed, the JupyterLab notebook is loaded from the Datapro cluster. The dataset is imported by using Gsutil URL API. Once the dataset is loaded, we perform the data cleaning and preprocessing. There were few columns that had null values and it was replaced it with the mean value of the particular column. In the next steps data visualization and EDA were performed.

Chart, bar chart

Description automatically generated

Fig 11 Distribution of Gender

Data visualization was performed so that there was a good understanding about the distribution of data. Another plot consisted of the smoking status of the patients.

Chart, bar chart

Description automatically generated

Fig 12 Plot of Smoking Status

In the next steps the ML pipeline was built on PySpark. One hot encoding and vector assembler were used for building the ML pipeline. The main task was to make a prediction if a patient would have a heart stroke or not.

**6 Evaluation and Result**

The classification algorithm used in our project was the Decision Tree classifier. An accuracy of 95.14% was achieved for this model.

A screenshot of a computer

Description automatically generated with medium confidence

Fig 13 Prediction Values

The following picture depicts the predicted values achieved for the decision tree classifier model. The predicted values are then converted from spark dataframe to pandas dataframe. Further they are then converted to a csv file and then sent as an output file to the bucket. Hence the bucket contains the output file of the predicted values as a CSV file.

The next tasks include utilizing Tableau for our project. As described in the architecture Tableau is used for the final step. In this task we use Tableau to connect to the database and perform few data visualizations and build a dashboard for our story. The Tableau server is connected to the cloud sql instance database to perform the EDA.

Chart, bar chart

Description automatically generated

Fig 14 EDA on Tableau

Then the EDA for prediction for the stroke dataset is then calculated in the next task.

Chart, bar chart

Description automatically generated

Fig 15 Prediction Output on Tableau

The output of the stroke data and predicted values is shown as a story on Tableau.

**7 Conclusion**

This project made an attempt to predict the stroke disease of a particular individual by taking in different factors related to the disease. This project was also a combination of integrating different types of components involved in a big data project. Utilizing the Google cloud platform to build the dataproc cluster, create the bucket, SQL instance, HDFS, PySpark and Tableau has enabled this project to be applied for solving big data problems.

**8 Future Scope**

This project can have great applications across the healthcare domain. With large amounts of data arriving at hospitals and clinics, preprocessing and processing is of the utmost importance. Utilizing a cloud like architecture can help solve these modern-day problems.

## REFERENCES

1. <https://data-flair.training/blogs/hadoop-tutorial/>
2. <https://www.ijsrp.org/research-paper-0319/ijsrp-p87104.pdf>
3. [https://www.researchgate.net/publication/333201013\_Healthcare\_Analysis\_U sing\_Hadoop\_Framework](https://www.researchgate.net/publication/333201013_Healthcare_Analysis_U)
4. [https://wiki.matbao.net/apache-spark-la-gi-loi-ich-khi-su-dung-apache-spark](https://www.researchgate.net/publication/333201013_Healthcare_Analysis_U)