12-15-2022

Big Data Analytics of Medical Data

ASHWIN RAJASANKAR
Grand Valley State University

Big Data Analytics of Medical Data

ASHWIN RAJASANKAR

A Project Submitted to

GRAND VALLEY STATE UNIVERSITY

In

Partial Fulfillment of the Requirements

For the Degree of

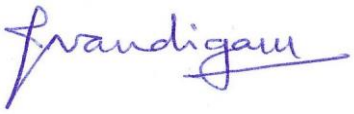
Master of Science in Applied Computer Science

School of Computing and Information Systems

December 2022

GRAND VALLEY STATE UNIVERSITY

The signatures of the individuals below indicate that they have read and approved the project of Ashwin Rajasankar in partial fulfillment of the requirements for the degree of Master of Science in Applied Computer Science.



_____ _____12/20/22_____

Dr. Jagadeesh Nandigam, Project Advisor Date

_____

<name of GPD>, Graduate Program Director Date

_____

<name of unit head>, Unit head Date

Abstract

Data has become a huge part of modern decision making. With the improvements in computing performance and storage in the past two decades, storing large amounts of data has become much easier. Analyzing large amounts of data and creating data models with them can help organizations obtain insights and information which helps their decision making. Big data analytics has become an integral part of many fields such as retail, real estate, education, and medicine. In the project, the goal is to understand the working of Apache Spark and its different storage methods and create a data warehouse to analyze data. The data used is obtained from the CMS (Centers for Medicare & Medicaid Services) website. The data consists of information such as prescriber name, drug generic name, drug brand name, drug price. Using Apache Spark, a data warehouse, the data can be queried and transformed to obtain valuable insights. Using Power BI, the data can be visualized which makes the data easier to understand and highlights the trends in the data.

Introduction

The field of big data analytics has grown at an exponential rate in the past decade. It is estimated that the big data analytics market is set to reach a value of $103 billion by 2023. Recent big data analytics has helped in many different fields such as real estate, advertising, and medicine. The impact of big data analytics on medicine has been huge, from manufacturing new drugs to analyzing side effects of different drugs it has accelerated the growth. It is estimated that big data in healthcare could be worth up to $71.6 billion by 2027, which is a 14.1% annual growth from 2022. There are many use cases of big data in the field of healthcare. Some examples are predictive analytics of illness, predictive medical imagining, drug development research etc. Common big data tools include Apache Hadoop, Apache Spark, MongoDB, Cassandra etc.

Apache Spark is a data processing framework, which is used to process large data workloads. It enables data parallelism when running large data stream to optimize queries. Spark has become widely popular in big data computing and replaced Hadoop in many use cases, primarily because of its speed when processing large data. The reason why Spark is faster than Hadoop is because Spark uses in-memory system while Hadoop uses local memory space to store data. Spark is also very versatile and provides many high-level APIs in Scala, Python, Java, and R.

Power BI is a business intelligence tool developed by Microsoft for creating data visualizations. It can be used to create different visualizations from pie charts to bar graphs. It also allows data to be loaded from different data sources such as CSV files, MySQL server, JSON etc. In the project Power BI will be used to create visualizations from the data queried from Spark.

In this project the working of Apache Spark is studied and is used to create a data warehouse of medical prescriber data and the information from the data warehouse is visualized using Power BI. The data in the warehouse is from the years 2013 to 2018. Using the available data, information such as most common drugs, most expensive drugs will be found which can help give insights and trends about the future of the medicine and increasing illness in specific parts of the US.

Project Management

When starting the project, the first priority was installing Spark and Power BI. Installing Spark was first attempted on a Windows machine, but because of memory limitations and other compatibility issues, it was installed on a Linux machine with better hardware. Power BI was also installed.

To get familiar with Spark and its configurations, I read the documentation and tried some tutorials. Spark can be run on cluster or client mode, in the cluster mode the program is run on a cluster and the workload is distributed, whereas in the client mode the workload is handled in the same machine that the program is run on. It was decided that the client mode was better to be used in this project.

After Spark was successfully set up in the system, the task of obtaining data for the data warehouse started. After some research, the CMS (Centers for Medicare & Medicaid Services) website was found, which had large scale prescriber data. The prescriber data from the years 2013 to 2018 were downloaded as CSV files from the website.

After all the raw data was downloaded, the next task was identifying the important parts of the data to use for the analysis and creating queries to identify insightful information from the data. Important values in the data were the drug cost, drug name, drug generic name, prescriber specialty etc. This information was used to query specific data from the data warehouse to get insights.

Organization

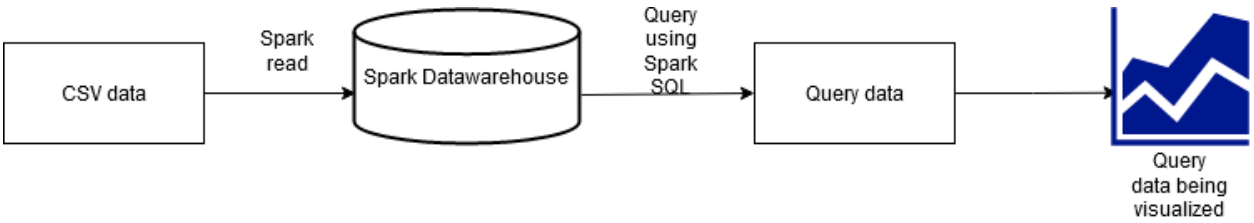


Figure 1. Flow of data

The above diagram shows the flow of data in the project. The main components of the project are the Spark warehouse and Power BI. First the CSV files are loaded into the data warehouse using Spark's read command. After all the CSV files are loaded, the data is partitioned by the year and stored as a Spark table. Partitioning is done to parallelize the computations done by Spark and it makes queries on large datasets run more efficiently and faster. The query results are stored into a CSV file and imported to Power BI. Using Power BI different plots can be generated, which makes the data easier to understand and makes identifying insights easier.

A data warehouse is considered as a repository of large amounts of information that can be analyzed to make informed decisions. A data warehouse is made up of multiple tiers. In the project the top tier is visual representations and plots created using Power BI, the middle tier is the Spark SQL and data frame which is used for loading and querying data and the bottom tier is the Spark table which contains the five years of data. Spark is ideal for creating a data warehouse since it can distribute workloads and handle large amounts of data. Since Spark has high level APIs in Python and Java it makes it easy for new users to quickly get the hang of it.

Another advantage of Spark is that it uses in-memory computation rather than disk memory. This makes it much faster than other big data tools such as Hadoop. Spark runs 100 times faster than Hadoop in memory and 10 times on disk. Because of the above reasons, Spark has become an important tool for big data analytics and is useful in creating data warehouses.

Reflection

During the course of the project, I learnt a great deal about Spark, its architecture and its use cases. I have always been interested in big data analytics tools and their uses. This project helped me learn more and improve my knowledge and skills. The first time I wrote a Spark SQL query, I was amazed at its speed compared to previous databases I have used such as MySQL and MongoDB.

While querying data about the most expensive drugs and number of prescriptions in states, I was astonished about the cost and how many people use them. This project also helped me grow my knowledge in the pharmaceutical domain, I learnt a lot about the big pharmaceutical companies and how they create and sell drugs. The data visualization part of the project of Power BI was also very interesting. I learnt a lot about how quicky data can be transformed and visualized. Comparing and trying out other BI tools such as Tableau or QlikView would have made the project better and given an insight of how each tool differs in visualization and data loading.

I also learnt that there are many Spark versions to use in the cloud. The best providers are Google cloud and Databricks. They provide Spark preinstalled in the environment, which saves the user time in installing them and configuring them. They also give options to create clusters as per the user's preference. This can be used in the future to analyze more data.

Conclusions

In the project, a data warehouse using Apache Spark was created to identify insights and query data. Power BI was used to visualize the queried data. The architecture of Spark and its advantages over other databases were studied.

For future enhancements the following can be done.

- The data can be used to check the increasing demand for specific drugs and can be used to check increase in diseases by season, which may help prevent spread.
- Most profitable and fast-growing pharmaceutical companies can be identified.
- Machine learning models can be created to identify prescription drugs for specific diseases and symptoms.

Appendix

Spark Code and Visualizations:

1. Creating a Spark session:

```
from pyspark.sql import SparkSession
spark = SparkSession \
    .builder \
    .appName("Python Spark warehouse") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

The spark session is responsible for creating an entry point to sparks functionality. It is an encapsulation of Spark context, hive context and SQL context.

2. Loading CSV file into Spark data frame:

```
df = spark.read.option("sep",
"\t").option("header","true").option("inferSchema","true").csv("
/bigdata/programs/spark-
programs/input/2018_data/PartD_Prescriber_PUF_NPI_Drug_18.txt")
```

The code shows the CSV file being loaded into the segments data frame.

3. Check the dimensions of the data frame:

```
print((df.count(), len(df.columns)))
```

It displays the number of rows and columns in the data frame.

4. Partitioning a data frame and saving it to a table:

```
(df
    .write
    .partitionBy("year")
    .saveAsTable("med_years"))
```

The above code partitions the data frame by year and saves it to a table called med_years

5. Most expensive prescribed drug:

```
spark.sql("select * from test_18 where total_drug_cost=(select
max(total_drug_cost) from test_18)").show()
```

The most expensive prescription drug was Ledipasvir and sofosbuvir combination is used with or without ribavirin **to treat chronic hepatitis C infection in adults and children 3 years of age and older.** It was prescribed by Kim Hinojosa a Nurse practitioner and costed $23 million.

6. Most prescribed drugs:

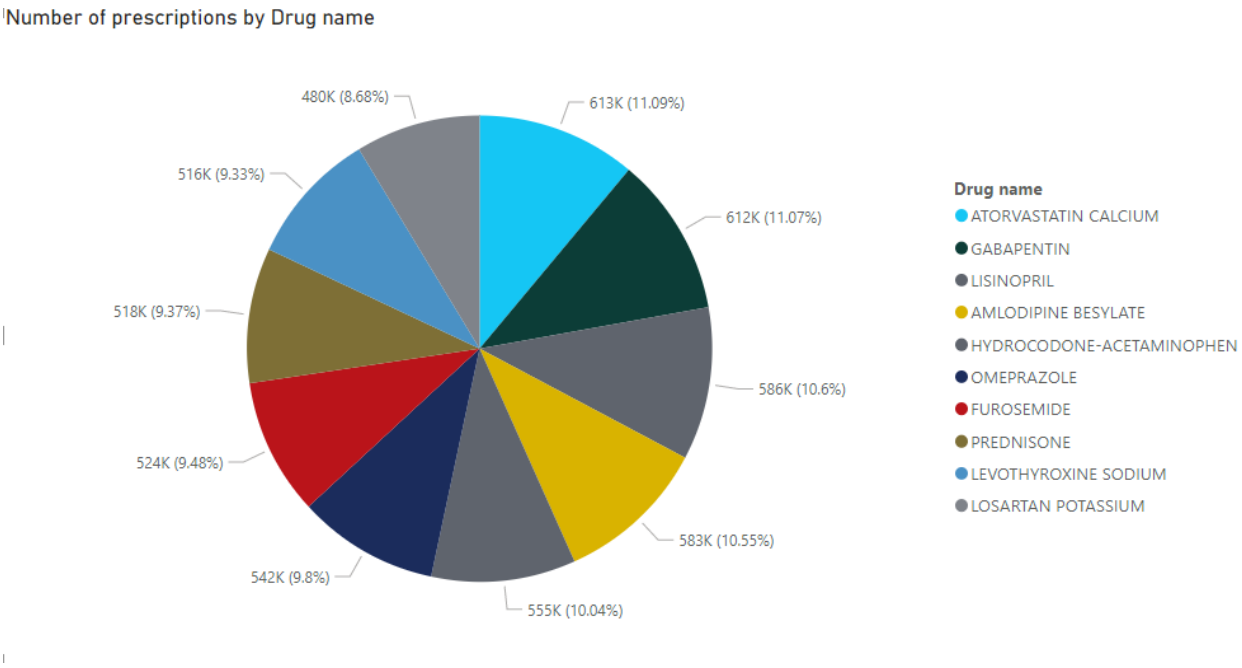
```
df = spark.sql("select drug_name, count(*) from test_18 group by
drug_name order by count(*) desc limit 10")
```

The query returns the top ten most prescribed drugs and is stored in the variable df.

```
df.write.options(header='True', delimiter=',').csv("/home
/ashwin/most_pres_drug.csv")
```

Using the above command, the query data is sent to a CSV file.

Using Power BI, the data is visualized.



As shown in the above visualization, the most commonly prescribed drug in Atorvastatin calcium, which is used **to lower the amount of cholesterol in the blood** and to prevent stroke, heart attack, and angina (chest pain).

7. Prescriptions by state:

```
df = spark.sql("select nppes_provider_state, count(*) from
test_18 group by nppes_provider_state")
```

The query returns the number of prescriptions grouped by state. The data from the query is stored to the variable df.

The data is stored in a CSV and visualized using Power BI.