ISSN No:-2456-2165

Collating SQL Databases, No-SQL Databases and Machine Learning Algorithms for Data Analysis

Dylan Coelho
IT dept. SFIT SFIT
Mumbai, India

Cliff Machado
IT dept. SFIT SFIT
Mumbai, India

Leon Correia
IT dept. SFIT SFIT
Mumbai, India

Shree Jaswal
IT dept. SFIT SFIT
Mumbai, India

Neil Fernando
IT dept. SFIT SFIT
Mumbai , India

Abstract:- Big Data Tools and Machine learning algorithms have been applied to data analytics and prediction frequently. This paper evaluates and illustrates the differences between SQL and NoSQL for storage of Big Data and processing and compares various algorithms used for analysis and predictions. The paper shows our basic understanding of Hadoop and Spark cloud and compares the two platforms on various parameters such as the time taken for input data and the time taken for the output data and the total memory used by the databases. The system has implementing the Databases in Hadoop and Spark.In Hadoop, the Hive database will be used for implementingthe SQL part and Cassandra for NOSQL. In Spark the SQLpart will be implemented using Post GreSQL and NOSQL uses MongoDB. We get the end results by comparing various parameters like the input, output data and the total memory used will be represented graphically after which a user will be in a position to choose the appropriate database accordingto their requirements. Additionally, we will also be studyingand comparing various Machine Learning algorithms by implementing them on the selected dataset. To compare the algorithms, we will be considering parameters of Accuracy, Root Mean Square Error and Mean Absolute Value. Choosing the right machine learning algorithm can be difficult, but doing so is essential to answering the given question with great speed and accuracy. In order for the user to yield the required insights, algorithms must be carefully analysed and studied upon considering parameters like these. The final research results will be illustrated with the help of graph on a UI which will help to better understand the results obtained on our selected datasetfor this particular paper.

Keywords:- *Hadoop, NoSQL, Spark, SQL.*

I. INTRODUCTION

Machine learning (ML) algorithms and Databases have beenfrequently applied for data analytics and prediction. While choosing a database, user has to either choose a relational (SQL) or non-relational (NoSQL) data structure.

Even it the two databases are suitable, there are some keydifferences between the them that one must keep in mind when making a decision. It is also important to develop Machine Learning algorithms to analyse the data-sets efficiently and accurately to predict the desired. Machine learning algorithms usually fall into one of three categories - supervised learning, supervised learning, and reinforcement learning. When dealing with different types of business challenges, analysts should carefully consider data factors, speed and accuracy requirements and other parameters to produce the information you want. The choice of the ML algorithm is dependent ona combination of factors like the problem statement and the type of output you want, the type and size of the data, the available calculation time, the number of features and the viewof the data, to name a few. Machine learning algorithms canbe divided into supervised and supervised learning.

Supervised learning algorithms are used when training data which have a variety of output that is consistent with input variables. The algorithm analyzes input data and reads the function to show the relationship between input variables and output. . Unsupervised learning algorithms are used when training data have no response variability. Such algorithms attempt to detect internal patterns and structures hidden in data. Clustering algorithms are types of unsupervised learning algorithms. By using the right algorithms, organizations can expect to benefit from dynamic data , which reflects the uniquecircumstances that guide each business. As their algorithms continue to learn and develop, so do data-driven decisions.The comparison between these databases and Ml algorithms can help make such decisions. The two platforms chosen for the comparison are Apache Spark and Apache Hadoop- HDFS.

Apache spark is used as it is a multi-language engine for executing data engineering, data science, and machine learningon single-node machines or clusters[1]. PostgreSQL, MySQL, Oracle and Microsoft SQL Server are all SQL databases. The Hadoop Distributed File System (HDFS) provided by Apache Hadoop is a distributed file system designed to run on commodity hardware. It is similar to existing distributed file systems[2].

However, the differences from other distributed file systems are quite significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications that have large data sets. HDFS is part of the Apache Hadoop Core project. NoSQL database examples in- clude Redis , Neo4j, BigTable, Cassandra, CouchDB, BigTable,HBase and MongoDB.

II. RELATED WORK

The research work of Mahmudul Hassan, Srividya K. Bansal [3] , demonstrates methods for distributed RDF data storage and querying schemes for HBase and Cassandra clusters. Results show that HBase outperforms for queries involving one subject. Cassandra performs better on queries with multiple subjects. The research work of Ana Flores, StalinRamirez, et.al.[4] elaborates on the research of response times on relational and non-relational data base models in a database provided by the Funcion Judicial del Ecuador. MongoDB with respect to SQL Server took less time to resolve the first queries, maintaining a small variable response time that tends to decrease. The research work of Chao-Hsien Lee and Zhe- Wei Shih , Based on experimental results over two different cloud platforms, the NoSQL database can always provide better performance than the SQL database while executingthe ML algorithm[5]. They use Random Forest and K means algorithms. The research work of Christine Niyizamwiyitira and Lars Lundberg , the paper evaluates the performance of SQL and NoSQL database management systems using cluster computing to run the database systems, with external load generators[6]. They compare the efficiency and use case of different database management systems which includes setup and configuration complexity. The research work of S. Ravikumar and P. Saraf proposed that the system works intwo methods: Regression and Classification. In regression, the system predicts the closing price of stock of a company.[7]In classification, the system predicts whether the closing price of stock will increase or decrease the next day. The research work of A. Moses and R. Parvathi , proposes stage by stage machine learning processes to build an efficient model capable of predicting traffic volume based on features which brings out hidden insights in vehicular movements.[8] This research has resulted in identifying an optimal model to the publicly avail- able dataset. The research work of Sunil Kaushik, Akashdeep Bhardwaj and Luxmi Sapra, The study in this paper attempts to solve rainfall prediction problems using machine learning techniques.[9] It evaluates machine learning algorithms using the rainfall data and other parameters – humidity, wind speed, max temperature and min temperature.

III. METHODOLOGY

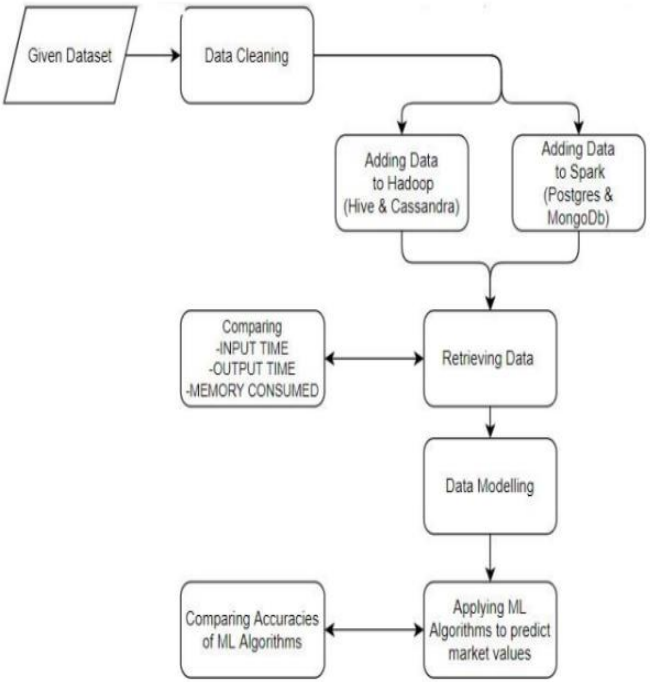

Fig. 1. Methodology

The project is implemented in several phases which aredata gathering , installation of big data tools, running andconnecting big data tools , data cleaning, coding the functions, adding the data to databases , retrieval of data , applying ml algorithms and comparing the results at the end

Phase 1: Data Gathering
We begin with finding the appropriate dataset

Phase 2: Installation of Big Data tools
We install the necessary big data tools- Hadoop and Sparkand set up the environment

Phase 3: Running and Connecting the Big Data tools
We run the big data tools via the CMD and make the required connections

Phase 4: Data cleaning
To begin the project in Jupyter notebook, we begin with data cleaning. We eliminate noisy data, null values and make it suitable for implementation

Phase 5: Coding the Functions
All the necessary functions are implemented using python

Phase 6: Adding the data to the databases
Once the data is ready, we load it into respective SQL and NOSQL databases of Hadoop and Spark Respectively

Phase 7: Retrieval of Data

We retrieve the data again in the Jupyter notebook and note the parameters of Input time, Output time and memoryutilized.

Phase 8:Applying ML Algorithms

Various machine learning algorithms are implemented and theAccuracy and RMSE of each algorithm is noted respectively.

Phase 9:Comparison of results

To display the parameters, we use inbuilt python libraries to generate the graphs for our parameters and compare the data.

IV. IMPLEMENTATION

The dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. It contains Major League Baseball Data from previous seasons. A data frame with 322 observations of major league players on the variables such as At Bat: Number of times at bat, Hits: Number of hits, Runs: Number of runs, Salary: annual salary on opening dayin thousands of dollars, etc.

Unnamed: 0	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	...	CRuns	CRBI	CWalks	PutOuts	Assists	Errors	League_N	Division_W
0	0	507	48	19	114	87	34	17	3540	388	...	554	317	239	1198	386	19	1	0
1	1	646	81	29	23	93	1	6	6632	827	...	213	276	500	1129	48	28	1	1
2	2	593	161	8	11	78	13	10	2746	859	...	276	249	93	404	70	21	0	1
3	3	641	80	19	61	29	6	8	2592	407	...	128	234	1074	1135	359	30	0	0
4	4	599	51	19	119	108	1	15	6109	463	...	661	129	92	330	337	25	0	0

Fig. 2. Dataset

A. Spark Platform

➢ *SQL storage:*

We begin by importing and storing the data in the Spark Platform for SQL using PostGreSQL. PostgreSQL[10] is a powerful, open source object-relational database system that uses and extends the SQL language combined with many features that safely store and scale the most complicated data workloads.

```
t1_start = getTime()

#Check once BaseBallSet number with above cell and then run this cell
dfPostGres = postgres.read \
    .format("jdbc") \
    .option("url", "jdbc:postgresql://localhost:5432/postgres") \
    .option("dbtable", "BaseBallSetSeven") \
    .option("user", "postgres") \
    .option("password", "user") \
    .option("driver", "org.postgresql.Driver") \
    .load()

t1_stop = getTime()

print("Elapsed time:", t1_stop, t1_start)

PostGresToSpark = t1_stop-t1_start
print("Elapsed time during Retreiving Data From PostGres Database in seconds:",
                                        t1_stop-t1_start)

dfPostGres.printSchema()
```

```
Elapsed time: 2021-10-08 13:24:32.038932 2021-10-08 13:24:31.956897
Elapsed time during Retreiving Data From PostGres Database in seconds: 0:00:00.082035
root
```

Fig. 3. SQL Storage Spark

➢ *NO-SQL storage:*

The task of Importing and storing data in Spark Platform for NOSQL is done using MongoDB. retrieve data from hive sql is noted respectively. 11]It is a document oriented, cross platform database that en- sures high performance, high availability, and easy scalability. MongoDB works on concept of collection and document.

```
t1_start = getTime()

dfMongo = my_spark.read.format('com.mongodb.spark.sql.DefaultSource').load()

t1_stop = getTime()

print("Elapsed time:", t1_stop, t1_start)

MongoToSpark = t1_stop-t1_start

print("Elapsed time during Retreiving Data From MongoDb Database in seconds:",
                                        MongoToSpark)
```

```
Elapsed time: 2021-10-08 13:17:43.467340 2021-10-08 13:17:42.980536
Elapsed time during Retreiving Data From MongoDb Database in seconds: 0:00:00.486804
```

Fig. 4. NoSQL Storage Spark

B. HADOOP PLATFORM

➢ *SQL storage:*

Storing and importing of data in Hadoop for SQL is done using a Hadoop tool called as ApacheHive. It is a data warehouse software which facilitates reading, writing, and managing large dataset residing in distributedstorage using SQL. Structure can be projected onto data already in storage[12]. A command line tool and JDBC driver are provided to connect users to Hive. The time taken to retrieve data from hive sql is noted respectively.

```
startTime = getTime()
dataframe = pd.read_sql("select * from BaseBallSet", conn)
endTime = getTime()
HiveToPython = endTime-startTime
print('Time taken to Retrive Data From Hive Sql ',HiveToPython)
dataframe.head()
```

```
Time taken to Retrive Data From Hive Sql  0:00:00.775994
```

Fig. 5. SQL Storage Hadoop

➢ *NO-SQL storage:*

NO-SQL storage in Hadoop platform is done using Apache Cassandra. Cassandra[13] isa NoSQL distributed database. By design, NoSQL databases are lightweight, open-source, non-relational, and largelydistributed. with the advent of Big Data and the need to rapidly scale databases in the cloud. Cassandra is among the NoSQL databases that have addressed the constraints of previous data management technologies, such as SQL databases.

ISSN No:-2456-2165

Fig. 6. NoSQL Storage Hadoop

C. Application of ML Algorithms

Fig. 7. Modelling

The data is split into 80:20,before applying the ML algorithms. The following algorithms are applied on the data for prediction: 1. Linear Regression 2. Ridge Regression 3. Lasso Regression 4. Elastic Net 5.KNN 6. SVR 7. Gradient Boost

V. RESULTS

For ease of use and better presentation of results a simple user interface has been designed which contains a selection panel and displays the results in graphical formats. The results are fetched from the jupyter notebook. The user can choose between the various comparisons they have to study. The Database Analysis compares the databases with the helpof three graphs i.e. Input time , Output time and Memory consumed. The Machine learning Analysis compares the Algorithms on basis on Accuracy and RMSE.

Fig. 8. UI

A. Databases

Given below are the graphs for comparison for input, output and Memory.

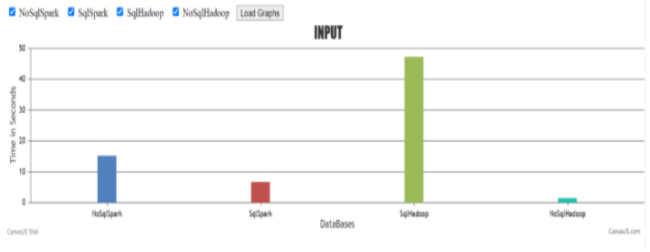

Fig. 9. Input data to Databases

In the above graphical comparison, SQL Hadoop took the most time while taking input. No SQL hadoop was the fastest amongst all in taking input. No-SQL thus gave a better performance here among all the databases.

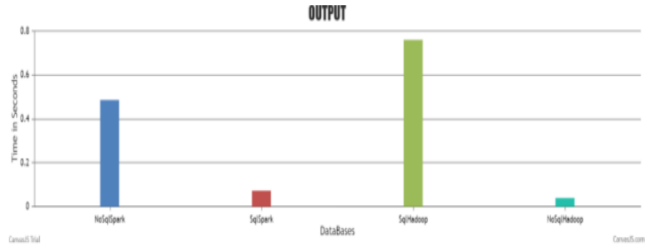

Fig. 10. Output data to Databases

In the comparison above ,just as input time ,SQL Hadoop has the highest output time taken. No SQL hadoop was the fastest amongst all which is similar to the results obtained in the input comparisons. No-SQL Hadoop thus gave a better performance here amongst all the databases.

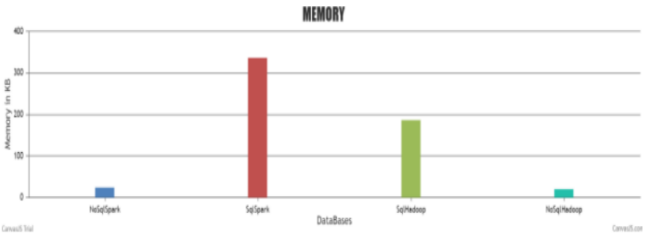

Fig. 11. Memory consumed by Databases

In the above comparison of memory consumed , SQL Spark consumed the most amount of memory meanwhile, No- SQL was the most efficient and consumed the least memory amongst all the other big data platforms. Based on results, the NoSQL database provided better performance than the SQL database

B. Machine Learning Algorithms

Given below are the graphs for comparison for input, output and Memory. Given below are the graphs for comparison for Accuracyand RMSE for various machine learning algorithms.

ISSN No:-2456-2165

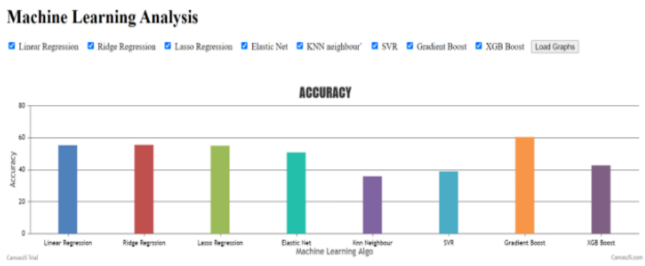

Fig. 12. Accuracies of ML algorithms.

The comparison of accuracy of eight machine learning algorithms is represented in the above graph. Majority of the ML algorithms had an accuracy between 40-60. Gradient boostin this performed the best and had the highest accuracy of 60. KNN neighbour had the lowest accuracy among all the algorithms which were compared.

Fig. 13. RMSE of ML algorithms

RMSE values of algorithms were obtained and are illustrated in the above graph.Gradient boost in this performed the best and had the least error . KNN neighbour had the highest error among all the algorithms which were compared. Based on experimental results over eight ML algorithms, Gradient boost performs the best for our particular dataset.

Fig. 14. Mean Absolute Value of Algorithms

Mean Absolute values of algorithms were obtained andare illustrated in the above graph. It refers to the mean of the absolute values of each prediction error over all the instances of the test data-set. Gradient boost in this performed the best and had the least error . XGB boost had the highest error among all the algorithms.

Based on experimental results over eight ML algorithms, Gradient boost performs the best for our particular dataset. KNN neighbour had the least accuracy and the highest RMSE value. XGB boost algorithm was outperformed by the remaining algorithms upon obtaining the Mean Absolute Value.

VI. CONCLUSION AND FUTURE WORK

A comparison between SQL and NoSQL databases and various machine learning algorithms was presented in this project. With the help of comparison, one can identify which database should be appropriate for a particular dataset. One can also decide which algorithms to use based on its accuracy or RMSE. The Hadoop NOSQL database i.e MongoDB performed the best in terms of input time, output time and memory consumed. Hadoop and Spark are the main platforms used for the storage of large amounts of data and using such platforms makes it easy to store and retrieve such large amounts of data easily. For Machine learning algorithms,Gradient boost was the best performer, providing the highest accuracy as well as the least RMSE where as KNN neighbour demonstrated poor accuracy as well as the most RMSE value.

REFERENCES

[1]. “Apache spark™ - unified engine for large-scale data analytics,” ApacheSpark™ - Unified Engine for large-scale data analytics. [Online]. Avail- able: https://spark.apache.org/. [Accessed: 11-Mar-2022].
[2]. “HDFS architecture,” Apache Hadoop 3.3.2 – HDFS Architec- ture. [Online]. Available: https://hadoop.apache.org/docs/stable/hadoop- project-dist/hadoop-hdfs/HdfsDesign.htmlIntroduction. [Accessed: 11-Mar-2022].
[3]. Mahmudul Hassan, Srividya K. Bansal, “Semantic Data Querying Over NoSQL Databases with Apache Spark”, IEEE International Conference on Information Reuse and Integration for Data Science, 2018.
[4]. Ana Flores, Stalin Ramirez, Javier Vargas, Renato Urvina, Jose Lavin, Renato Toasa, “Performance Evaluation of NoSQL and SQL Queriesin Response Time for the E-government”, ICEDEG-18 Proceedings - Quito, Ecuador, 2018.
[5]. Chao-Hsien Lee and Zhe-Wei Shih, “A Comparison of NoSQL and SQL Databases over the Hadoop and Spark Cloud Platforms using Machine Learning Algorithms”, IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), 2018.
[6]. Christine Niyizamwiyitira and Lars Lundberg, “Performance Evaluation of SQL And NOSQL Database Management Systems in a Cluster”, International Journal of Database Management Systems (IJDMS) Vol.9,No.6, December 2017.
[7]. S. Ravikumar and P. Saraf, ”Prediction of Stock Prices using Machine Learning (Regression, Classification) Algorithms,” 2020 International Conference for Emerging Technology (INCET), 2020.
[8]. A. Moses and R. Parvathi, ”Vehicular Traffic analysis and prediction using Machine learning algorithms,” 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic- ETITE), 2020.

[9]. Sunil Kaushik, Akashdeep Bhardwaj and Luxmi Sapra, "Predicting Annual Rainfall for the Indian State of Punjab Using Machine Learning Techniques", 2020 2nd International Conference on Advances in Com- puting, Communication Control and Networking (ICACCCN), 2020.
[10]. Group, P. S. Q. L. G. D. (2022, April 22). PostgreSQL. Retrieved April 22, 2022, from https://www.postgresql.org/
[11]. The Application Data Platform. MongoDB. (n.d.). Retrieved April 22, 2022, from https://www.mongodb.com/
[12]. General. Apache Hive. (n.d.). Retrieved April 22, 2022, from https://hive.apache.org/
[13]. Welcome to Apache Cassandra's documentation! Apache Cassandra. (n.d.). Retrieved April 22, 2022, from https://cassandra.apache.org/doc/latest/