

Study and Analysis of Various Big Data Analytics Tools for Data Processing

S.S. Aravinth¹, P.Vijay Anand², A.Arul Prasath³ and Dasari Anantha Reddy⁴

¹Associate Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Guntur-522502
Email: aravinthkrithick@gmail.com

²Assistant Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Guntur-522502

³Assistant Professor, Veltech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, TN, India

⁴Assistant Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Guntur-522502

Abstract—Huge amount of data is generated in this digital world with the size of terabytes and petabytes. The traditional database management system is failed to process these exaggerated data. Big data is the combination of high volume, high velocity and different veracity of data. These data are extracted from multiple sources such as machine data, social media data, medical data, banking data, transaction data. E commerce data and business data. In order to process the huge amount of data, the specially designed software, tools, framework and database file systems are required. Since the structure of big data is considered as unstructured and semi structured, the highly scalable and fault tolerant processing environment is needed to cop up with these big data. The big data analytics life cycle is divided into few steps, such as business understanding, data understanding, data preparation, modelling, evaluation and deployment. The employed software or tools are designed and developed to fit into these life cycle of big data analytics.

In this study, the predominantly used Hadoop framework, and mongo DB database are studied and applied on sample dataset to process the larger amount of big data. The results section is producing performance analysis report of these above mentioned data processing tools. Each and every tools are performing well and producing the expected efficient results on those huge amount of data. 2021 tokyo Olympics dataset is downloaded from the kaggle website and processed in this research study. This dataset contains 5 fields of records like details of athletes, coaches, gender entries, medals and names of the teams worldwide.

Index Terms— BIG DATA ANALYTICS, HADOOP, HBASE AND MONGODB.

I. INTRODUCTION

Big data analytics is a process of collecting, processing and extracting the hidden insights from the huge size of database. This extracted information is used to decision making process of the organization or industries. Unlike the traditional database, these big data systems are required high speed data processing tools to process all the globally available data. Many available sources are indicated in the given below figure 1. The main origin of big data growth is social media data.

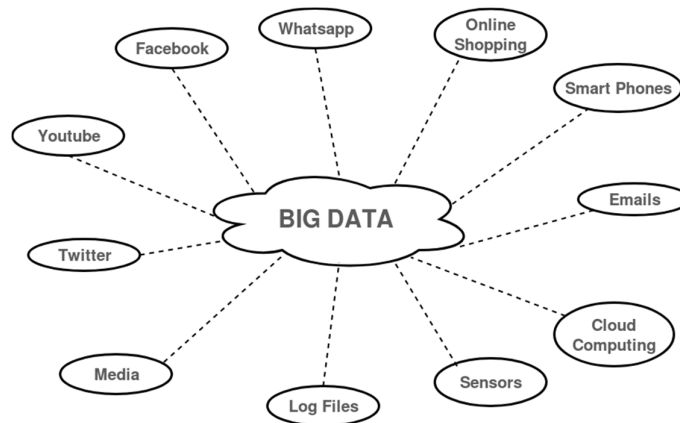


Figure 1. Source of Big Data

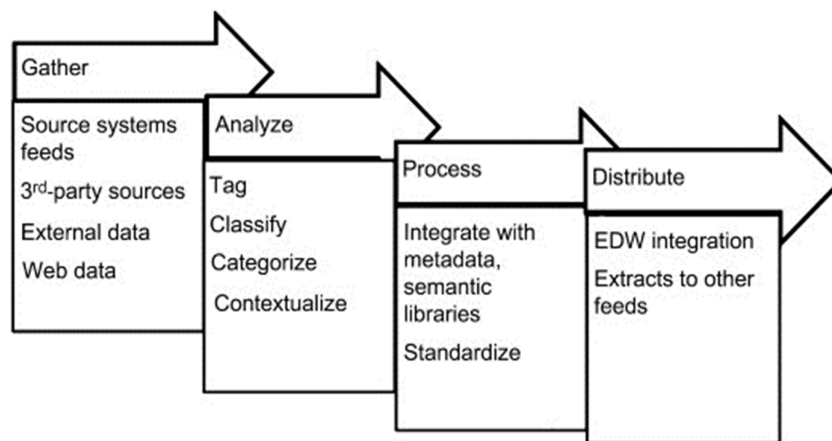


Figure 2. Steps in Big Data Analytics

As per the above figure 2., the four steps are involved in big data analytics process such as, gathering the informative data from external web sources, analysing the data in the stipulated time, processing the integrated data to extract the insights and distributing that information to operational format for end users. These big data analytics techniques have some challenges also such as data quality, discovery, storage, analytics and security. In other hand, the technical specialists are required to extract, load and transformation of data. Some warehousing tools are also applied in this data analytics process.

II. DATA PROCESSING WITH HADOOP

Hadoop is abbreviated as highly available object oriented platform for storing and processing the larger set of data. This is a specially designed file system as well as database. The major components of Hadoop are Hadoop Distributed File System(HDFS) and Map Reduce(MR). Here HDFS is for storing the data and files where as MR is a program to communicate with the Hadoop cluster. The secondary component of Hadoop is Yet Another Resource Navigator (YARN). This YARN manages the resource allocation of Hadoop Cluster. The given below figure 3. is explaining about the components of Hadoop inter process communication. As per this figure 3. the name node and data node associated with HDFS are configured. In other side, the job tracker and task tracker associated with MR are connected. Only one name node for the entire cluster is permitted. But there is no restriction on data node.

This Hadoop has number of commands to process the data stored in HDFS. The input format of data is key, value pairs. Only one job tracker is allowed per cluster to pass the map reduce programs. But here are no restrictions on task tracker TT. If the name node fails to initiate the process, then the secondary node will take charge and it will do the process.

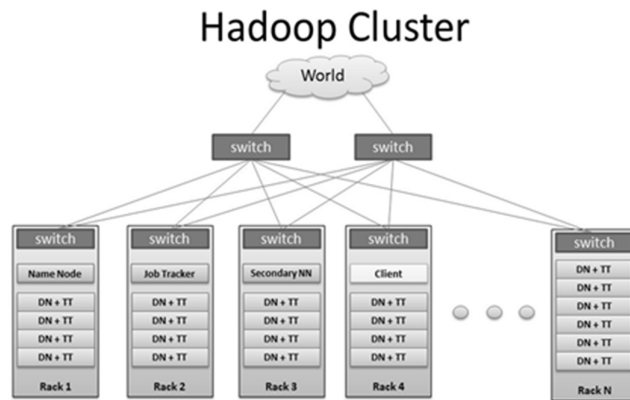


Figure 3. Hadoop Cluster Set up with HDFS and MR

The basic replication factor of this Hadoop is 3. Each and every time, the data gets loaded for three times. The entire data is divided into various sub blocks to store the files. After the allocation of blocks the data will be given with Object Identifiers(ID's). These object ID's are making the speed process of data extraction, storing and processing with in the Hadoop cluster.

III. DATA PROCESSING WITH MONGODB

The Mongo database is a type of No SQL database to process the structured data which are stored in a rows and columns of schema. Mongo database supports for document specific database as illustrated in the following query.

Query 1. Example MongoDB Table

First name = 'Aravinth'

Address = "London"

Spouse= [{Name: "Terasa"}]

First name = 'Aravinth'

Address = "Florida"

In this above query two separate documents are processed with help of Mongo DB database. So this combining feature will increase the runtime and execution time of huge databases. Above all, the Mongo DB has the following features

1. Replication
2. Indexing
3. Reporting
4. Duplication of data
5. Backup and recovery mechanism
6. Load balancing

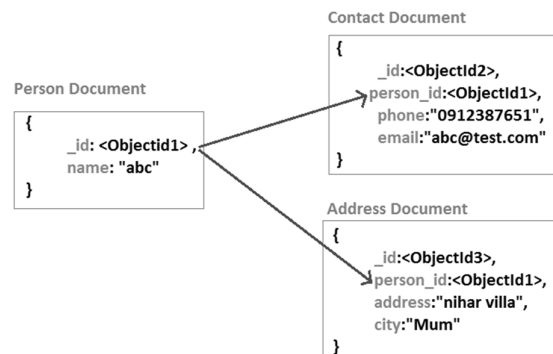


Figure 4. Parsing the Web Document using Mongo DB

This web document scraping is the process extracting the web pages of user information. This document contains personal document, contact document and address document. Id and name are the fields of person document. Id, person id, phone number, and email id are the fields of contact document. Similar to this, id, person id, address and city are the fields of address document.

IV. IMPLEMENTATION

This section focuses on the workflow and implementation details. Initially the Olympics website has been studied and some fields are identified.

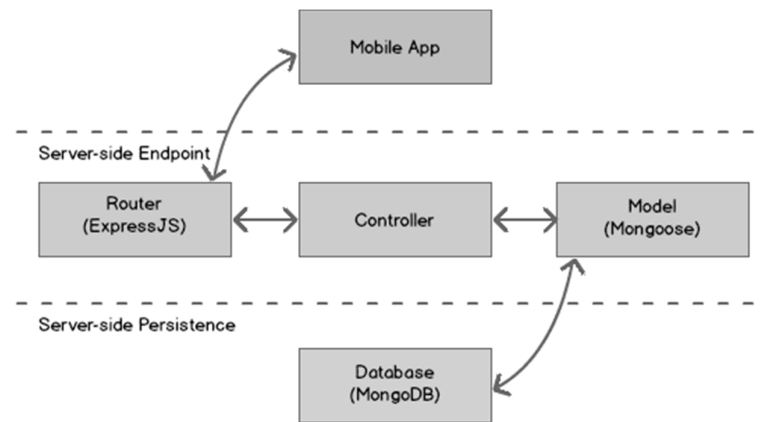


Figure 5. Mongo DB Working Environment Set up

The mongo db database set up is divided into two core server side points. Those are server side end point and persistence point. Solving and directing the route to data is taken care by controller. The given below figure 6. depicts the steps involved in Hadoop installation with the enterprise edition. Hadoop cluster set up establishment process involves the master and slave communication that will be employed in the data centre Hadoop deployment. The version of installed Hadoop is 3.X which is very faster and latest version in the market.



Figure 6. Hadoop installation and environment setup

V. RESULTS & DISCUSSION

As discussed in the abstract, the Olympics 2021 dataset is considered here for processing with respect to Hadoop and Mongo DB. This data contains 5 fields of the Olympics game. The following table 1. Clearly gives the processing time of these sample dataset of 5 fields.

TABLE I. PROCESSING TIME OF DATASET

S.No	Data Field	Name	Processing Time by Oracle SQL	Processing Time by Hadoop	Processing Time by Mongo DB
1.	33258 Rows	Athletes	12.36 milliseconds	1.3 milliseconds	1.323 milliseconds
2.	180 Rows	Coaches	1.3 milliseconds	0.63 milliseconds	0.639 milliseconds
3.	188 Rows	Entries Gender	2.69 milliseconds	2.1 milliseconds	1.369 milliseconds
4.	658 Rows	Medals	5.6 milliseconds	3.6 milliseconds	3.5 milliseconds
5.	2976 Rows	Teams	9.3 milliseconds	3.69 milliseconds	3.56 milliseconds

As per the above table, the processing time of the data fields are compared with traditional database system SQL, Hadoop and MongoDB database.

- Equality (=)
- Inequality (<, !=, or >)
- Less than or equal (<=)
- Greater than or equal (>=)
- IS NULL
- IS NOT NULL

These above SQL operators are being used in Oracle Sql, Hadoop and Mongoddb database respectively. The following lines of queries are processed to extract the insight from those hidden data. As per the table 2., the number of queries are processed by comparative databases are shown.

TABLE II. NUMBER OF QUERIES PROCESSED BY SQL, HADOOP AND MONGODB

S.No	Data Field	Name	Queries Processed by Oracle SQL	Queries Processed by Hadoop	Queries Processed by Mongo DB
1.	33258 Rows	Athletes	369	36	89
2.	180 Rows	Coaches	89	56	83
3.	188 Rows	Entries Gender	56	22	36
4.	658 Rows	Medals	89	46	59
5.	2976 Rows	Teams	12	3	9

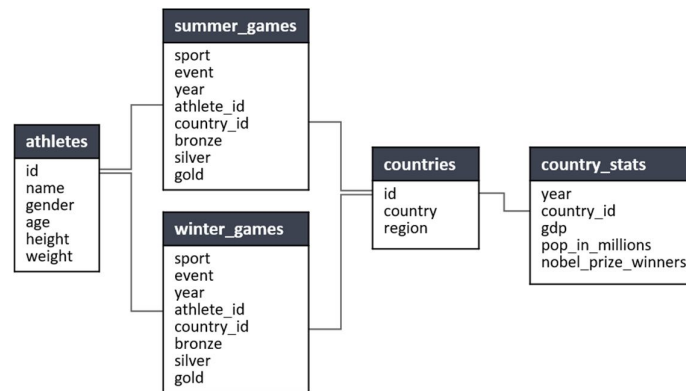


Figure 7. Mapping of ER Diagram with the Dataset

As a pre-processing step, initially the fields of dataset are mapped with ER diagram to find the relationship with the fields name and countries. The field of athletes is compared with winter-games and also country is compared with summer games.

```

In [154]: q = """
SELECT
    Year, Sex, COUNT(Sex) as count
FROM
    Athletes
Where Year > 1900 AND Sex = 'F'
GROUP BY
    Year
ORDER BY
    Year DESC
"""
pd.read_sql_query(q,conn).head(10)

Out[154]:

```

	Year	Sex	count
0	2016	F	6223
1	2014	F	2023
2	2012	F	5815
3	2010	F	1847
4	2008	F	5816
5	2006	F	1757
6	2004	F	5546
7	2002	F	1582
8	2000	F	5431
9	1998	F	1384

Figure 9. Applying queries on Dataset and Processing

The functions and aggregate operations are applied to compare and differentiate the various columns. In other hand, the python package such as pandas, numpy and matplotlib are used to do the data processing with these datasets.

VI. CONCLUSION

After the advent of social media, the big data are generated every day. Since the big data processing is considered as a challenging one, the computationally intensive performing algorithms are being deployed in the industry. As the traditional database management system are considered as inefficient to process the data, the special database called Hadoop and mongo database are playing a major role in processing the big data set. In this paper, the big data set has been extracted and processed as well in quicker time to produce the prompt output extraction.

REFERENCES

- [1] Thanekar, S.A., Subrahmanyam, K., Bagwan, A.B., "Big data and MapReduce challenges, opportunities and trends", International Journal of Electrical and Computer Engineering., Volume 6, Issue 6, 2016, Pages 2911-2919.
- [2] Londhe, A., Prasada Rao, P.V.R.D., "Platforms for big data analytics: Trend towards hybrid era", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017., 19 June 2018, Pages 3235-3238.

- [3] Banchhor, C., Srinivasu, N., "Integrating Cuckoo Search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification", Data and Knowledge Engineering, Volume 127, May 2020, Article number 101788.
- [4] Balaraju, J., Prasada Rao, P.V.R.D., "Designing authentication for hadoop cluster using DNA algorithm", International Journal of Recent Technology and Engineering, Volume 8, Issue 3, September 2019, Pages 6930-6935
- [5] Sirisha, N., Kiran, K.V.D., "Authorization of data in Hadoop using Apache Sentry", International Journal of Engineering and Technology(UAE), Volume 7, Issue 3.6 Special Issue 6, 2018, Pages 234-236.
- [6] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara and K. Sadatdiynov, "A survey of data partitioning and sampling methods to support big data analysis," in Big Data Mining and Analytics, vol. 3, no. 2, pp. 85-101, June 2020, doi: 10.26599/BDMA.2019.9020015.
- [7] S. Salloum, J. Z. Huang and Y. He, "Random Sample Partition: A Distributed Data Model for Big Data Analysis," in IEEE Transactions on Industrial Informatics, vol. 15, no. 11, pp. 5846-5854, Nov. 2019, doi: 10.1109/TII.2019.2912723.
- [8] S. Kumar and M. Singh, "A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem," in Big Data Mining and Analytics, vol. 2, no. 4, pp. 240-247, Dec. 2019, doi: 10.26599/BDMA.2018.9020037.
- [9] K. Rattanaopas and S. Kaewkeeree, "Improving Hadoop MapReduce performance with data compression: A study using wordcount job," 2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2017, pp. 564-567, doi: 10.1109/ECTICon.2017.8096300.
- [10] C. Li and W. Yang, "The distributed storage strategy research of remote sensing image based on Mongo DB," 2014 Third International Workshop on Earth Observation and Remote Sensing Applications (EORSA), 2014, pp. 101-104, doi: 10.1109/EORSA.2014.6927858.
- [11] Shinde, S.A., Raja Rajeswari, P., "A novel hybrid framework for cuff-less blood pressure estimation based on vital bio signals processing using machine learning", International Journal of Advanced Trends in Computer Science and Engineering, Volume 9, Issue 2, 2020, Pages 1556-1561.
- [12] Tumuluru, P., Lakshmi, C.P., Sahaja, T., Prazna, R., "A Review of Machine Learning Techniques for Breast Cancer Diagnosis in Medical Applications", Proceedings of the 3rd International Conference on I-SMAC IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2019, December 2019, Article number 9032427, Pages 618-623.
- [13] Saba, S.S., Sreelakshmi, D., Sampath Kumar, P., Sai Kumar, K., Saba, S.R., "Logistic regression machine learning algorithm on MRI brain image for fast and accurate diagnosis", International Journal of Scientific and Technology Research, Volume 9, Issue 3, March 2020, Pages 7076-7081.
- [14] Dharmapuri, S.L., Dandamudi, P.K., Botcha, V.M., Kolla, B.P., "Detecting central nervous system disorder using machine learning technique (XGB classifier)", International Journal of Emerging Trends in Engineering Research, Volume 8, Issue 4, 2020, Pages 1142-1147.
- [15] Rizwana, S., Challa, K., Rafi, S., Imambi, S.S., "Enhanced biomedical data modelling using unsupervised probabilistic machine learning technique", International Journal of Recent Technology and Engineering, Volume 7, Issue 6, March 2019, Pages 579-582.
- [16] Srinivas, P.V.V.S., Pavan Sai Sujith, L.V.N., Sarvani, P.M., Kumar, D.S., Parasa, D., "Prediction of hospital admission using machine learning", International Journal of Scientific and Technology Research, Volume 8, Issue 12, December 2019, Pages 2764-2770.