Managing Big Data with Hadoop Map Reduce For Solving the Problems of Traditional RDBMS

HTU RA
Lecturer, Computer University (Myitkyina), Kachin State, Myanmar

Abstract- ***According to the usage of Internet is extremely increasing, the amount of data being generated by everywhere makes traditional database technologies unable to store and process efficiently and effectively. Moreover, Relational Database Management System (RDBMS) is hardly possible to manage extremely large amount of data called "Big Data" in structured, semi-structured and unstructured forms from diverse data sources. In this paper, Hadoop, distributed big data processing platform, is applied for managing big data to overcome the storage and processing issues of traditional RDBMS. For experimentation, "Bag of Words" dataset from UCI Machine Learning Repository is utilized as unstructured big text data tested on Apache Hadoop MapReduce Multi Node Cluster. According to the outcomes of experimentation, applying Hadoop offers not only faster execution time but also better data scalability performance for managing and processing big text data.***

Indexed Terms- Bag of Words, Big Data, Hadoop, MapReduce, and RDBMS.

I. INTRODUCTION

Relational Database Management System (RDBMS) typically uses a relational database and schema for storing, retrieving and managing structured data applying Structured Query Language (SQL). Although RDBMS have been influencing the technology by managing the structured data, scaling of extremely large data is not hardly possible according to ACID properties such as Atomicity, Consistency, Isolation and Durability. Although RDBMs have shown remarkable ability in database technologies, simple data structures and a limited set of data types especially in SQL92 type RDBMS became an issue when implementing the new kinds of applications using complex data and data structures. In addition, RDBMS is absolutely weaken in handling semi-structured and unstructured data. Modern database applications actually need to store and manage large volume of multimedia data objects such as music, videos, images and maps, etc. Besides, specific business applications require the ability to define data types by users intended for the representation of complex relationships, including composition and aggregation. In current time, the volume of data in our world has been increasing even in day by day. Datasets are typically within the terabytes range and sometimes they may reach to petabytes or even exabytes. Among them, machine generated data volume is much more than the traditional data [4].

According to various surveys and studies, companies such as Facebook, eBay, Amazon, and Google generates petabytes or even exabytes of data every day. Facebook handles 600,000 photographs for a user in one second. Google also handles billions of URLs with related internet content, crawl metadata, geographic objects and hundreds of terabytes of satellite image data, with hundreds of millions of users and thousands of queries per second. Handling extremely large amount of data with the traditional storage techniques has come out of the term "Big Data". Big data can be categorized into three main types such as structured data, semi structured data and unstructured data. Relational databases and spreadsheets are good examples of structured data. For unstructured data, data can be in any form from diverse data sources such as social media, email, photos, multimedia etc. It cannot be evaluated with standard statistical algorithms and methods because it does not follow any particular rule. In semi structured data, similar entities are congregated together. According to the estimation of researchers in data analytics research, structured data typically composed of only about 5% of the total volume of generated data. The rest portions of total volume of generated data constitute semi-structured or unstructured data which

makes it more difficult to handle and process in extracting meaningful information. Nowadays, the primary focus of big data and big data analytics is to discover meaningful insight using traditional data mining techniques such as rule-based systems, decision trees, and pattern mining and other techniques upon the extremely large data sets efficiently [9], [10], [12].

To solve the issues and problems of existing RDBMS for managing and processing big data, the NoSQL movement creates open source non-relational solutions. NoSQL databases are a new type of databases for non-relational data storage and the have many good typical features such as simplified data model, little or no support for OLTP, no standard query language and no support of integrity constraints, etc. Among them, the weakness of ACID properties is the most suitable matter in the management and processing of big data. Moreover, MapReduce programming model which is developed by Google and Hadoop including HDFS (Hadoop Distributed File System) offers parallel computation and processing platform upon extremely large datasets [14]. Hadoop platform is actually intended to provide highly parallelizable and executable a program on a large cluster of commodity tens of thousands of machines managing data partitioning, task scheduling, and also recovery from machine failures. In this paper, hadoop, distributed big data processing platform, is applied for the managing and processing of big data to facilitate the storage and processing issues of traditional RDBMS. In addition, the paper focuses to overcome the issues and challenges of using traditional tools and techniques in managing big data. The paper is organized as follows: the related works which are motivated for the paper are described in section 2. The background theory associated with big data, the primary focus of the paper, is presented in section 3. In section 4, hadoop, distributed big data processing platform including its components for storage and processing big data is discussed. The experimentation portion of the paper is also expressed in section 5. Finally, the discussions and future works are described in section 6.

II. RELATED WORKS

George Feuerlicht [4] presented that the evolution of database management systems by reviewing and analyzing the developments of research and implementations by modern database management applications. The author made some observations about applications which are clearly data-intensive where there is no database and schema and there is no also support for database queries. These applications are truly difficult to regard as database applications demanding new systems and technologies. Big data is typically a collection of different types of data which is extremely large and complex to process using traditional database management tools or data processing applications. S. Vikram Phaneendra and E. Madhusudhana Reddy [13] discussed about up-to-date big data systems and technologies including Apache Hadoop MapReduce architecture and its future use-cases. Jaroslav Pokorný [10] presented that many possibilities concerned with how to store and process big data. The author also explained managing of big data depends on applications in which how much data volume will access, the complexity of mining algorithms will be used and so on. A new feature is, that some of these systems have more different components that enable access and process data stored in various ways. In addition, the current trends and approaches which are related with big data management and processing intended for possible research directions. In an era of data analytics, the analysis of big data absolutely requires faster improvements by addressing many technical challenges concerning with it. Raveena Pandya, Vinaya Sawant, Neha Mendjoge and Mitchell D'silva [3] discussed about the challenges for big data not only the obvious problems of data scalability but also heterogeneity of data from data collection to result visualization. In addition, unstructured big text data is generated from many sources. Zaheeruddin Ahmed [15] emphasized about the efficient way of storing unstructured big data and then applying suitable approaches of analyzing it. The author also discussed some of the challenges and issues related to big data analytics concerning with text analytics.

III. TRADITIONAL DATA VS BIG DATA

A. Traditional Data

In earlier days, not only the type of data available was limited to apply but also a limited set of technologies to manage and process data in most of the applications. Traditional database technologies are enable to store and process the amount of data available efficiently in updating, storing and querying and so on. RDBMS has actually influenced the database technologies to manage and query in tabular data, we called structured data. The structured data can be organized using a pre-defined data model which can be significantly seen in relational databases and Excel applications.

B. Big Data

For unstructured data, it cannot be organized using pre-defined model. Video, text, audio and others are good examples of unstructured data. And, the semi-structured data can be assumed between the categories of structured and unstructured data, for example, Extensible Markup Language (XML). The storage and processing of extremely large amount of structured or semi-structured as well as unstructured data anytime we desire is a big issue to solve. The emergence of "Big Data" has become according to some difficulties and limitations to manage and process extremely large amount of data with the traditional database technologies like RDBMS and others. Generally, big data can be defined when the size or amount of data is beyond the capabilities of traditional database management technologies and tools to collect, process, retrieve, manage, store and analyze. Research on analyzing big data has proposed many scalable algorithms and efficient data structures, and some optimizations to analyze it [2], [6]. Big data can be characterized as follows:

Volume: It refers to the amount of data which is being collected approximately ranging from terabytes to exabytes.

Velocity: It refers to the rate of data generation. Typically, traditional data analysis is based on periodic updates, for example, daily, weekly or monthly. However, big data should be in real or near real-time processing indeed.

Variety: It refers to diverse types of data that are being collected such as structured, semi-structured and unstructured data.

Veracity: It refers to examine the reliable data from uncertain and imprecise raw data to obtain valuable insights.

Variability: It refers to data which are produced from diverse data sources may be an increasing complexity levels in managing data.

IV. HADOOP: DISTRIBUTED BIG DATA PROCESSING PLATFORM

For storage and processing big data, hadoop provides parallel and distributed processing of very huge datasets using Map Reduce together with HDFS. It is a good solution for storing and managing big data efficiently.

A. Hadoop Distributed File System (HDFS)

One primary component of the Hadoop is HDFS which break data in the Hadoop Cluster into many pieces and distributed across the different severs in this cluster. HDFS maintains the replication of the data for preventing the failures of hardware and data loss. Moreover, it has the architecture of master/slave for storing data. A HDFS cluster typically consists of a single Name Node and many Data Nodes, one per node in the cluster, which handles the storage of data connected to the nodes that they run on this cluster [5].

B. MapReduce

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel and distributed algorithm on a cluster. It is intended for distributed and parallel processing of big data offering reliable and fault-tolerant services. It is still very simple technique in compared with the area of distributed databases [8], [11]. There are two distinct phases in MapReduce:

1) Map Phase: The input data or job or workload is partitioned into smaller ones. Besides, the tasks are assigned to Mapper, which processes each unit block of data to produce a sorted list of (key, value) pairs. This list, which is the output of mapper, is

passed to the next phase. This process is known as shuffling.

2) Reduce: The input data is analyzed and merged to produce the final output which is written to the HDFS in the cluster.

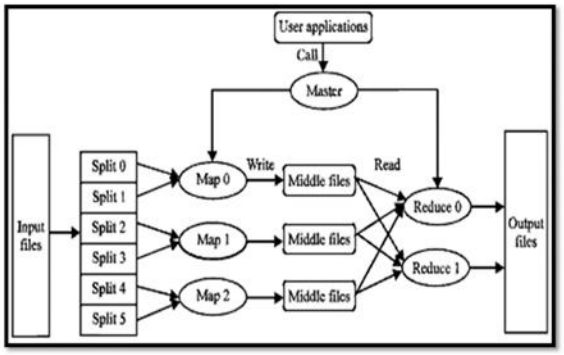

Fig. 1 Process Flow of MapReduce

V. EXPERIMENTATION

A. Experimental Setup

In this paper, the experiments were performed on Apache Hadoop MapReduce Multi Node Cluster which consists of three machines or processing nodes. One for master (server node) and two slave nodes in this cluster. On both the machines, a DataNode (Storing of Data) will be running. A Multi Node Cluster in Hadoop typically contains two or more DataNodes in a distributed Hadoop environment.

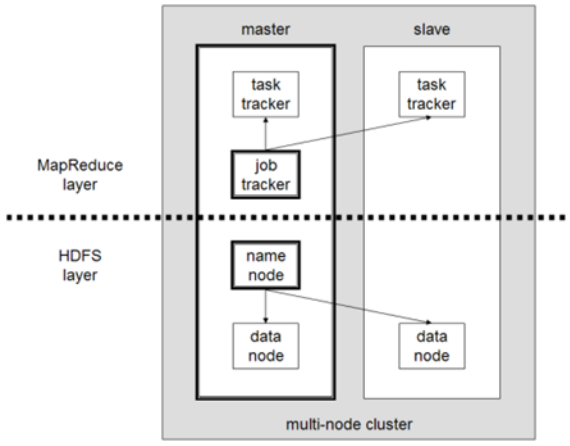

Fig.2 Multi Node Hadoop Cluster

B. Applied Dataset

Nowadays, with increasing volume of unstructured text documents, huge collection of documents in digital libraries, repositories, and etc. are piling up quickly and effectively managing on these documents is urgently required to solve. Moreover, text data analytics involves getting insights from text content in documents, books, social media, and various other sources [1], [7]. The raw, unstructured text data via "Bag of Words" dataset from UCI Machine Learning Repository is applied as unstructured big dataset in this paper. This dataset contains totally five text collections in the form of bags-of-words. For each text data collection, there are extremely large number of documents are organized from more than 300,000 web pages. It composed of 8000,000 number of row or data records or instances and 100,000 number of columns or attributes in the form of bags-of-words.

C. Experimental Results

According to the data structure and size of "Bag of Words" dataset utilized in this paper, we encountered a big problem or issue when we load the whole dataset including all text documents data using traditional RDBMS. There's a time constraint on how much time we have to store these massive data volume. Moreover, by applying SQL, there are some queries which are going to be attempting to retrieve from it. In this situation, we must also consider the processing time in retrieving or querying some outcomes from these extremely large amount of text data. Therefore, we made the experiments on distributed data processing platform, Apache Hadoop MapReduce Multi Node Cluster. To verify the performance of applying this cluster for unstructured big text document dataset, the experiments were made by varying the number of data records or size of data volume as shown in figure 3. According to the figure, it can be clearly seen that the execution time resulted from Hadoop platform is very faster than in compared with traditional platform with single machine. In addition, data processing on traditional standalone machine, it presents a major effort with single processor's calculation which will prevent data scalability. According to the experiments, after processing more than 7000,000 data records, it produces "Out of Memory" results. Therefore, it limits its usage only on small and medium datasets. In order to overcome these limitations, Multi Node Hadoop Cluster is the best option to choose for managing big datasets. Although the execution time generated from cluster of machines should be faster than single machine, the important matter is that scalability of data

volume which is essential in managing big data especially unstructured big text data in this paper.

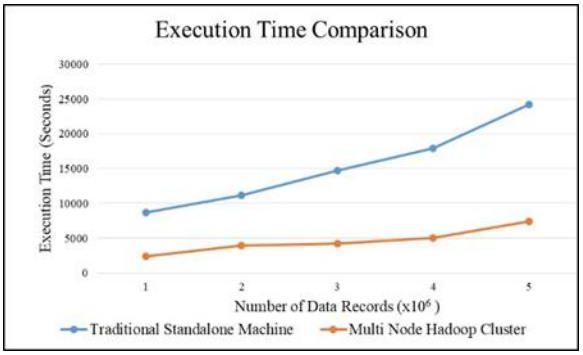

Fig. 3 Execution Time Comparison between Two Platforms

VI. DISCUSSIONS AND FUTURE WORKS

As the use of internet is increasing, the extremely large amount of data being generated by everywhere makes traditional technologies and techniques unable to store and analyze it. This unpredictable increase in volume and complexity of data is challenging for existing database management technologies. Moreover, the massive data volume of unstructured text documents has become text data analytics to extract meaningful insights from text content in these documents. In this paper, the experiments were performed on Apache Hadoop MapReduce Multi Node Cluster applying "Bag of Words" dataset from UCI Machine Learning Repository. According to the experimentation, applying Hadoop platform offers not only faster execution time but also better data scalability performance for managing and processing unstructured big text data. In future works, we will extend the research work by experimenting diverse big datasets with different data sizes.

REFERENCES

[1] B. Gully, H. Eduard, -Extracting data records from unstructured biomedical full text,| in Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007, pp. 837-846.

[2] B. Peter, B. L. Michael, Christian, -The meaningful use of big data: four perspectives--four challenges,| ACM Sigmod, vol. 40, pp. 56-60, 2012.

[3] D. Mitchell, M. Neha, S. Vinaya. (2015). Big data vs Traditional data. International Journal for Research in Applied Science & Engineering Technology (IJRASET).

[4] F. George, -Database Trends and Directions: Current Challenges and Opportunities,| in Proc. DATESO, 2010, pp. 163-174.

[5] G. Shankar, R. Siddarth, -Big data analysis using Apache Hadoop,| in Proc. International Conference on IT Convergence and Security (ICITCS), 2014, pp. 1-4.

[6] K. A. Bhadani, J. Dhanya, -Big data: challenges, opportunities, and realities,| in Proc. Effective big data management and opportunities for implementation, 2016, pp. 1-24.

[7] K. Kanimozhi, M. Venkatesan, -Unstructured data analysis-a survey,| International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, pp. 223-225, 2015.

[8] M. Rahul, K. Rashmi, et al., -Efficient analysis of big data using map reduce framework,| International Journal of Recent Development in Engineering and Technology, vol. 2, 2014.

[9] O. Carlos, -Can we analyze big data inside a DBMS?,| in Proc. 16th international workshop on Data warehousing and OLAP, 2013, pp. 85-92.

[10] P. Jaroslav, -How to Store and Process Big Data: Are Today's Databases Sufficient?,| in Proc. IFIP International Conference on Computer Information Systems and Industrial Management, Springer, Berlin, 2015, pp. 5-10.

[11] P. Rabi, Padhy, - Big data processing with Hadoop-MapReduce in cloud systems,| International Journal of Cloud Computing and Services Science, vol. 2, pp. 1-16, 2013.

[12] S. Naga, R. Monika, et al., - Data Migration from RDBMS to Hadoop, 2016.

[13] S. Vikram, R. E. Madhusudhan, -Big Data-solutions for RDBMS problems-A survey,| in Proc. 12th IEEE/IFIP Network Operations & Management Symposium, Osaka, Japan, 2013.

[14] T. Abderrahim, B. Abdessamad, et al., -A Comparative Study of Hadoop-based Big Data

Architectures,| International Journal of Web Applications (IJWA), vol. 9, pp. 129-137, 2017.

[15] Z. Ahmed, -Data management and big data text analytics,| in Proc. National Conference on Novel Trends in Computer Science (TECHSA-17), 2017, pp. 140-144.