iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** XI **Month of publication:** November 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 3 Issue X, November 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Processing Weather Sensor Data Using Iterative MapReduce Approach

B. Vasudevan[1], A. S. Shanthi[2]
[1]PG Scholar, [2]Assistant Professor
Tamilnadu College of Engineering, Coimbatore

Abstract — Big Data is the most promising research area in the field of Cloud Computing. In areas like social networking applications, e- commerce, finance, health care, education, etc, huge amount of data is being accumulated. As new data and updates are constantly arriving, the results of data mining applications become stale and obsolete over time. Hence another approach is required for mining big data. Hence in this paper, an iterative approach is introduced which is an extension of already existing map reduce approach. This iterative approach is experimented with the real time weather sensor data for particular area. Furthermore, the performance is compared with the existing data mining applications of big data.
Key words— Cloud computing; Big Data; MapReduce

I. INTRODUCTION

Now-a-days, in Information communication Technology (ICT), cloud computing plays a vital role to perform complex computing and large scale operations. Cloud computing has the advantages like virtualized resources , parallel processing, security and data service integration with scalable data storage. Cloud computing can not only minimize the cost and restriction for automation and computerization by individuals and enterprises but can also provide reduced infrastructure maintenance cost, efficient management, and user access . As a result of the these advantages, a number of applications that influence various cloud platforms have been developed and resulted in a tremendous increase in the scale of data generated and consumed by such applications. Some of the first adopters of big data in cloud computing are users that deployed Hadoop clusters in highly scalable and elastic computing environments provided by vendors, such as IBM, Microsoft Azure, and Amazon AWS. [1]
In recent years, Big data is the new emerging concept in the field of information technology. The term Big data refers increases in the volume of data which are difficult to store, process and analyze through traditional database technologies. The data may be structured, semi-structured or unstructured. Big data can be characterized as four Vs. Volume, Variety, Velocity and Value. Thus big data can also be defined as a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex and of massive scale.

II. CLASSIFICATION OF BIG DATA

Cloud has large scale data and it can be classified into different categories based on the following aspects: data sources, content format, data stores, data staging and data processing. Each of these categories has its own characteristics and complexities. Figure 1 shows the classification of big data.

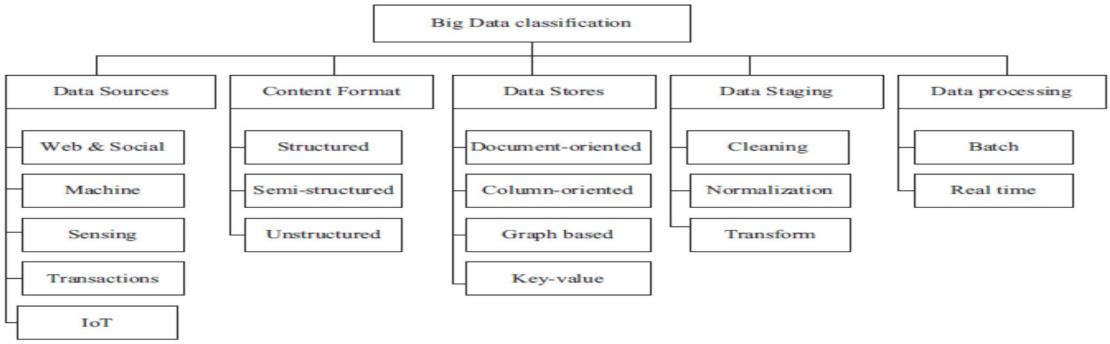


Fig. 1 Big data classification

III. CHARACTERISTICS OF BIG DATA

The characteristics of big data are explained as follows

A. Volume

It refers to the amount of all types of data generated from different sources and continues to expand. The benefit of gathering large amounts of data includes the creation of hidden information and patterns through data analysis.

B. Variety

It refers to the different types of data like video, image, text, audio and data logs. These data can be collected via sensors, smart phones or social networks. The data can be in either structured or unstructured format. Mostly mobile applications generate unstructured data such as text messages, online games, blogs and social media. The data generated by internet users is a combination of both structured and semi structured data.

C. Velocity

It refers the data transfer speed. The contents of data is constantly changed because of the absorption of complementary data collections, introduction of previously achieved data or legacy collections, and streamed data arriving from multiple sensors.

D. Value

It is an important aspect of big data and it refers to the process of discovering huge hidden values from large datasets with various types and rapid generation

E. Variability

The inconsistency the data can show at times which can hamper the process of handling and managing the data effectively.

F. Veracity

The quality of captured data, which can vary greatly. Accurate analysis depends on the veracity of source data.

G. Complexity

Data management can be very complex, especially when large volumes of data come from multiple sources. Data must be linked, connected, and correlated so users can grasp the information the data is supposed to convey.

There are several platforms and tools that are available to support big data. Some of them are Hadoop, mapReduce, Cassandra, HBase, MongoDB, pig, sqoop, etc. Among these technologies that supporting big data, this work focus on mapReduce.

IV. RELATED WORKS

To perform complex computations and massive scale operations cloud computing is the powerful technology. Because it eliminates the need to maintain expensive computing hardware, dedicated space and software. Big data generated through cloud computing has been observed. In [2] the rise of big data is reviewed. The definition characteristics and classification of big data along with some discussions on cloud computing are introduced. The relationship between big data and cloud computing, big data storage systems and Hadoop technology are also discussed. Furthermore, research challenges are investigated, with focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues and governance. Lastly, open research issues that require substantial research efforts are also summarized.

Despite the advances in hardware for hand-held mobile devices, resource-intensive applications (e.g., video and image storage and processing or map-reduce type) still remain off bounds since they require large computation and storage capabilities. Recent research has attempted to address these issues by employing remote servers, such as clouds and peer mobile devices. For mobile devices deployed in dynamic networks (i.e., with frequent topology changes because of node failure/unavailability and mobility as in a mobile cloud), however, challenges of reliability and energy efficiency remain largely unaddressed. To the best of our knowledge, we are the first to address these challenges in an integrated manner for both data storage and processing in mobile cloud, an approach we call k-out-of-n computing. In our solution, mobile devices successfully retrieve or process data, in the most energy-efficient way, as long as k out of n remote servers are accessible. Through a real system implementation we prove the feasibility of

Volume 3 Issue X, November 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

our approach. Extensive simulations demonstrate the fault tolerance and energy efficiency performance of our framework in larger scale networks [3].

Data analysis is an important functionality in cloud computing which allows a huge amount of data to be processed over very large clusters. Map Reduce is recognized as a popular way to handle data in the cloud environment due to its excellent scalability and good fault tolerance. However, compared to parallel databases, the performance of Map Reduce is slower when it is adopted to perform complex data analysis tasks that require the joining of multiple data sets in order to compute certain aggregates. A common concern is whether Map Reduce can be improved to produce a system with both scalability and efficiency. Map-Join-Reduce, a system that extends and improves Map Reduce runtime framework to efficiently process complex data analysis tasks on large clusters is introduced in [4]. They first proposed a filtering-join-aggregation programming model, a natural extension of Map Reduce's filtering-aggregation programming model. Then, presented a new data processing strategy which performs filtering-join-aggregation tasks in two successive MapReduce jobs. The first job applies filtering logic to all the data sets in parallel, joins the qualified tuples, and pushes the join results to the reducers for partial aggregation. The second job combines all partial aggregation results and produces the final answer. The advantage of their approach is joining multiple data sets in one go and thus avoid frequent check pointing and shuffling of intermediate results, a major performance bottleneck in most of the current Map Reduce-based systems.

In recent days sensors plays a vital role and they are becoming ubiquitous in collecting information and having a great influence in industrial applications to intelligent vehicles, smart city applications, and healthcare applications, etc. but each and every applications uses different types of sensors depending upon their usage. The rate of increase in the amount of data produced by these sensors is much more dramatic since sensors usually continuously produce data. It becomes crucial for these data to be stored for future reference and to be analyzed for finding valuable information, such as fault diagnosis information. So in [5] a scalable and distributed architecture is described for sensor data collection, storage, and analysis. The system uses several open source technologies and runs on a cluster of virtual servers. GPS sensors are used as data source and run machine-learning algorithms for data analysis.

V. MAP REDUCE

Map Reduce is the heart of Hadoop. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The term Map Reduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map Reduce implies, the reduce job is always performed after the map job.

Figure 2 is the pictorial representation of map reduce. The process begins by slicing the input text files and maps each slice to the mapper. The mapped slices are sorted and reduced by the reducer. Finally the reduced results are then aggregated. This has been shown in the following diagram.

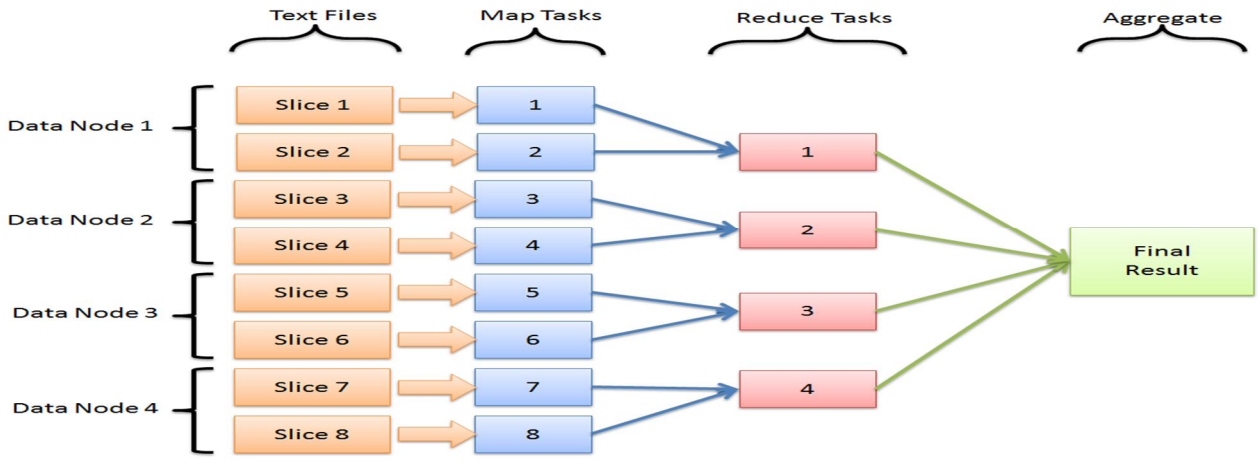


Fig. 2 Map reduce

Volume 3 Issue X, November 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

VI. SYSTEM DESIGN

The weather sensors are placed in each and every area in order to forecast the weather of that particular area. Here, real time weather sensor data for a year of a particular area is taken as input data. In other words we can call it as big data. This work is performed in the Hadoop open source framework and with the help of eclipse. Initially the real time weather sensor data is moved the HDFS which is installed in four machines. Then the weather data is splitted and passed to the mapper function and converted into key, value pairs (K, V). The key is unique whereas value may have repeated values. This key, value pair is given as input to the reduce function so that the reduced output is represented as the final result. This can be represented in figure 3.

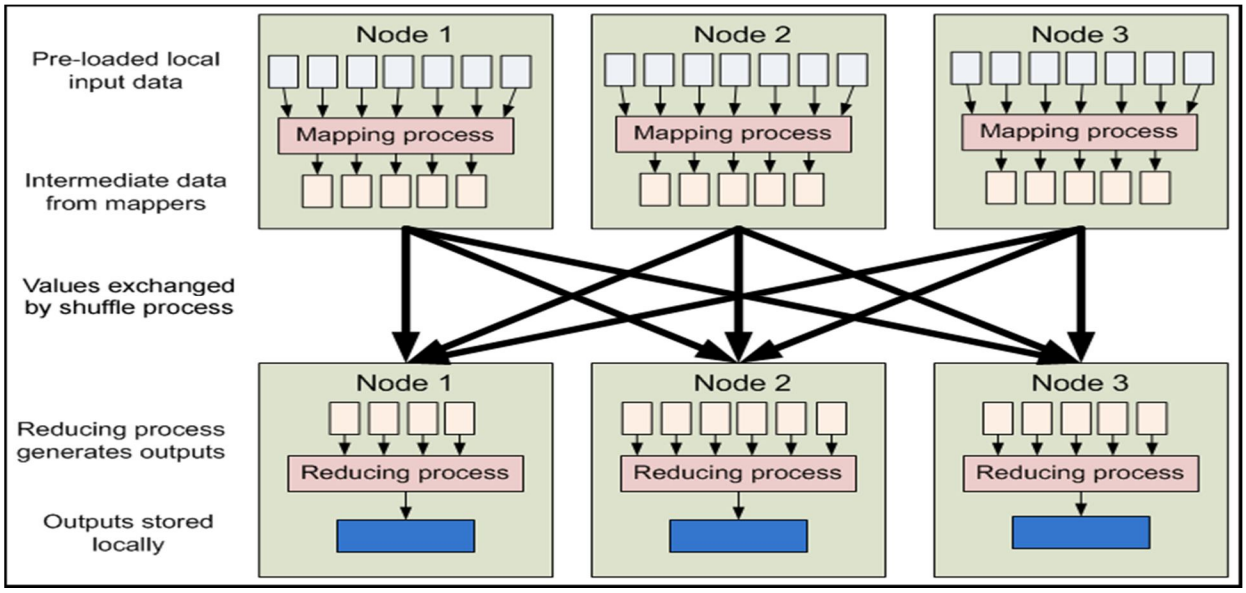


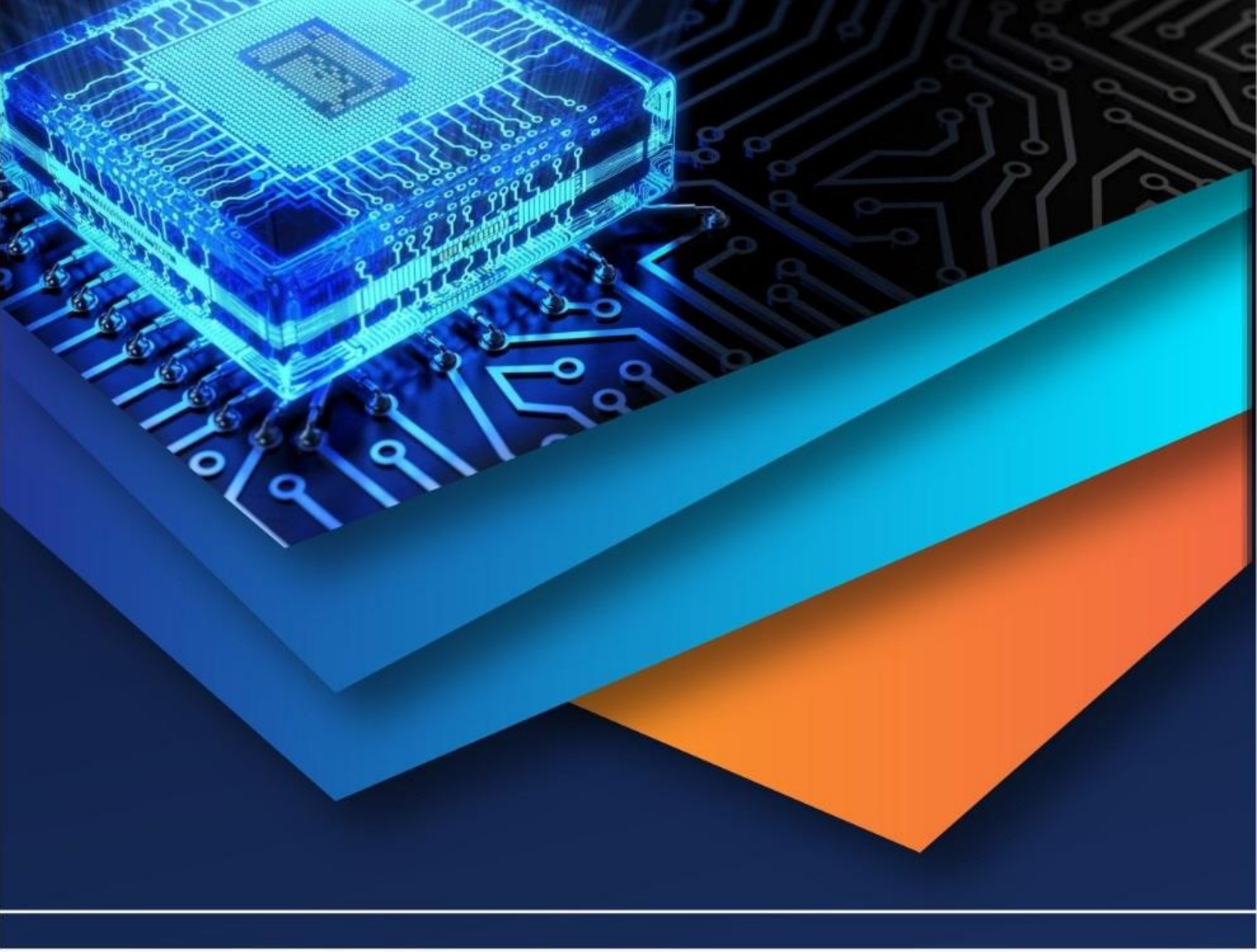
Fig. 3 System Architecture

VII. CONCLUSION

Big data plays the major role while considering the applications such as social networking applications. Iterative mining is required to obtain the updated data. This is performed in this work. The real time weather sensor data is taken for year of a particular area and the operations are performed. The big data is processed iteratively so that resulting in an accurate mining of the updated and new growing data. The average temperature for an year is the obtained result. This process is fast when comparing with the existing approaches.

REFERENCES

[1] Q. Yang, “Introduction to the IEEE Transactions on Big Data”, IEEE Transactions on Big data, no. 1, vol. 1, pp. 2-14, January 2015.
[2] S. Chen, Q. Wang, G. Yu and Y. Zhang, “i2MapReduce: Incremental MapReduce for Mining Evolving Big Data”, IEEE transactions on Knowledge and Data Engineering, vol. 27, no. 7, pp. 1906-1919, July 2015.
[3] Hashem, I. Yaqoob, S. Mokhtar, A. Gani and S. Khan, “The rise of “big data” on cloudcomputing: Review and open research issues”, Elsevier, no. 47, pp. 98-115, 2015.
[4] C. Chen, W. Myounggyu , R. Stoleru and G. G. Xie, “Energy-Efficient Fault-Tolerant Data Storage and Processing in Mobile Cloud”, IEEE Transactions on cloud computing, pp. 28-41, March 2015.
[5] D. Dahiphale, R. Karve, A. V. Vasilakos, H. Liu, “An Advanced MapReduce: Cloud MapReduce, Enhancements and Applications”, IEEE Transactions on Network and service Management, pp. 101-115, April 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY