Proceedings of Second Shri Chhatrapati Shivaji Maharaj QIP Conference on Engineering Innovations
Organized by Shri Chhatrapati Shivaji Maharaj College of Engineering, Ahmednagar
In Association with Novateur Publications JournalNX-ISSN No: 2581-4230
February, 22nd and 23rd, 2019

A Study on use of Big Data in Cloud Computing Environment

Miss . Thombare Mayuri

TE, Computer Engineering,
Pravara Rural Engineering College, Loni,
Maharashtra, India
mayurethombare@gmail.com

Miss . Amale Rohini

TE, Computer Engineering,
Pravara Rural Engineering College, Loni,
Maharashtra, India
rohiniamale01@gmail.com

***Abstract*— The purpose of this paper is to present definition, characteristics, and classification of big data along with environment for carrying net analytics in clouds for Big Data application. Big data and Cloud computing are the hot issues in current vast Information Technology. Effective management and analysis of large-scale data present an interesting but critical challenge. Now a days handling of big data is one of the main problem and this can be sorted with the help of cloud computing. The connection between the big data, cloud computing, big data storage systems, and Hadoop technology are also presented. This paper provides a complete review of the big data state of the art, characteristics, management and research challenging aspects.**

***Keywords*— Big data, Big Data in Cloud, Hadoop, Map Reduce.**

I. INTRODUCTION

[3]Big data is a word used for description of massive amounts of data which are either structured, semi structured or unstructured. The data if it is not able to be handled by the traditional databases and software technologies then we categorize such data as big data. The term big data is originated from the web companies who used to handle loosely structured or unstructured data."Every day, we create 2.5 quintillion bytes of data so much that 90% of the data in the world today has been created in the last two years alone. Lots of data is being collected and warehoused Web data, e-commerce

Bank/Creditcard

transactions

– Social Network

Society is becoming increasingly more instrumented and as a result, organizations are producing and storing vast amounts of data. [6]Data is becoming more valuable.

2 .DATA STORAGE

Presently, the dealer or customer produces large amount of information. The database system is used in the best way to store

.

The information or to create data. The data may be unstructured or structured. The very first step after creating data base is analyzing the data i.e; data is moved to the data warehouse. The main reason we use data warehouse is we have many deal. Who produce a bulk of data or the information. So we have too many information's to get the particular information regarding the particular dealer/customer we use data warehouse which analyze the complete information. One of the application of the database is it not only stores or creates but also they design the database for analysis. This overall database is supported by SQL.

Another distinct trend in cloud computing is, it increases the use of NOSQL Database that are prepared for storing and retrieving information. The survey plays an important role, it generally explores the data model that studies NOSQL system support. support the data model, types of query and support for concurrency, consistency, replication and partitioning it is compared with different

NOSQL system.

There are some of the challenges we are facing in Big Data Management.

They done research on this topic and also they analyzed what are the issues of this and some of the challenges.

Primarly the Data Variety: It is nothing but handling the data. We have different variety of data's to be transferred or it can also be of how to send multiple data. But, the important point is the data what we send should be meaningful.

Data Storage :This defines where we store he data or how efficiently we store and recognize the data. Storing is the important aspect of Big Data. It is also defined how to store large volume of data or information. We also see how to store

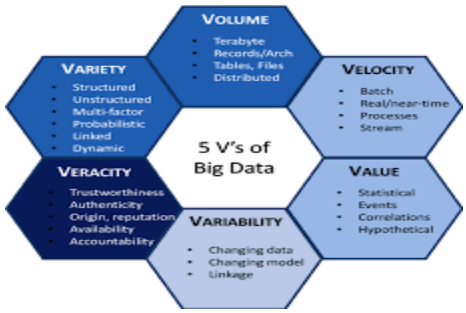
information and the way it can be easily ported between data centers.

Data Integration : The new protocols and interfaces are integrated and able to manage the data that may be structured, unstructured, semi-structured. Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error---free difference resolution.

Data Processing and Resource Management : It defines how the data is communicated and optimized and also the new programming models for streaming system. It also enables to combine the application from multiple programming models.

3. [5]FIVE V'S OF BIG DATA

There are many properties associated with big data. The prominent aspects are volume, variety, velocity, variability and value.



volume: many factors contribute for the increase in volume like storage of data, live streaming etc.

variety: various types of data is to be supported.

velocity: the speed at which the files are created and processed are carried out refers to the velocity.

variability: it describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set.

value: all enterprises and e-commerce systems are keen in improving the customer relationship by providing value added services.

Categorization of big data falls with major aspects, since this technology involves with multiple diversified fields and ir-related types of information handling. Some of the classes can be framed like storing, sourcing, formatting, staging and processing.

Sourcing: Data sources are identified are internet web pages, discussion forums, chats and messages shared in and among social networks, remote sensing networks. All kinds of day to day transactions done through internet based applications.[4 Formats: Unstructured, partially structured, and structured.

Storing: Image based, graph based, documents, key value stores.

4. DATA MANAGEMENT

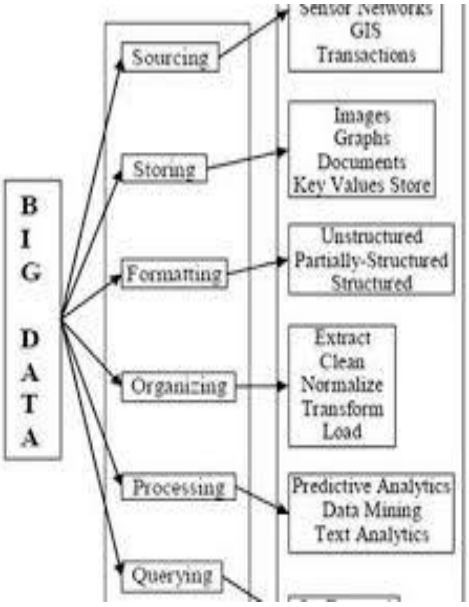
This is one of the time-consuming tasks of analytics i.e; to prepare the data for analysis; Analytics are performed on large volumes of data that requires efficient methods to store, filter, transform, and retrieve the data. Cloud analytics solutions need to consider the multiple Cloud deployment models adopted by enterprises are :

Private: These mainly work on the private network, managed by the organisation itself or by the third party. A private Cloud is suitable for businesses that require the highest level of control of security and data privacy.

Public : These work with off-site over the Internet and available to the general public. Public Cloud offers high efficiency and shared resources with low cost. The quality of services such as Privacy, security, and availability is specified in a contract.

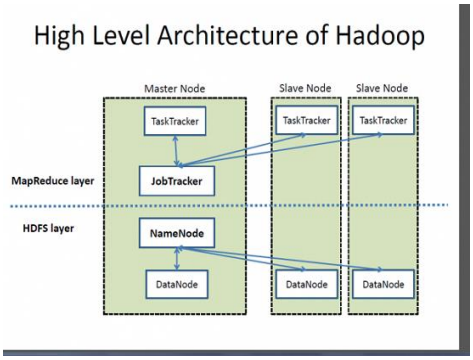
Hybrid : combines both Clouds where additional resources from a public Cloud can be provided as needed to a private Cloud.

Customers can develop and deploy analytics applications using a private environment.



5. HADOOP

Hadoop which is a free, Java based programming frame work supports the processing of large sets of data in a distributed computing environment. It is a part of the apache project sponsored by the apache software foundation. Hadoop cluster uses a
master/slave structured. Using hadoop, large data sets can be processed across a cluster of servers and applications run on systems with thousands of nodes involving thousands after bytes. Distributed file system in hadoop helps in rapid data transfer rates and allows the
system to continue its normal operation even in case of some node failures.



Task trackers are responsible for running the tasks that the job tracker assigns them.

Job trackers has two primary responsibilities which are managing the cluster resources and scheduling all user jobs.

Data engine consists of all the information about processing the data.

Fetch manager has to fetch the data while particular task is running.

6. ADVANTAGES OF BIG DATA

1. Cost Savings : Some tools of Big Data like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are to be stored and these tools also help in identifying more efficient ways of doing business.
2. Time Reductions :The high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analyzing data immediately and make quick decisions based on the learnings.

3. New Product Development : By knowing the trends of customer needs and satisfaction through analytics you can create products according to the wants of customers.

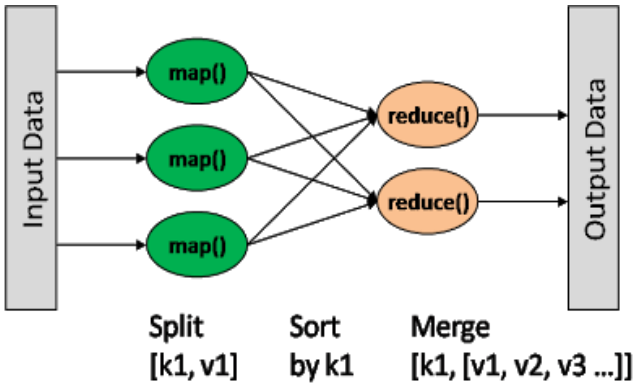
7. BIGDATA APPLICATIONS

In the current age of data explosion, parallel processing is very much essential for performing a massive volume of data in a timely manner. Parallelization techniques and algorithms are used to achieve better scalability and performance for processing of big data. Map reduce if a very popularly used tool or model used in industry and academics. The two major advantages of map reduce are encapsulation of data storage distribution, replication details. It is very simple for use by the programmers to code for the map reduce task. Since the map reduce is schema free and index free, it requires parsing of each records at reading point.map reduce has received a lot of attentiveness in the fields of data mining, information retrieval, image retrieval etc.

The computation becomes difficult to be handled by traditional data processing which triggers the development of big data apps. Big data provides an infrastructure for maintaining transparency in manufacturing industry, which has been having the ability to unreveal uncertainties that exists in the component performance and availability. Another application of the big data is the field of bioinformatics which requires large scale data analysis.

8. MAP REDUCE

Map reduce frame work is used to write apps that process a large amounts of data in a reliable and fault tolerant way. The application is initially divided into individual chunks which are processed by individual map jobs in parallel. The output of map sorted by a frame work and then sent to reduce the tasks. The monitoring is taken care by the framework.
The input data is divided into individual chunks and re provided for processing by the map task. These map task process the data in parallel and the result from the map task is then provided to reduce the task where the results that are generated in parallel by the map task are consolidated and the reduced report is given as output.



9. BIGDATA MANAGEMENT

The needs of the big data are not being satisfied by the current technologies and the speed of increasing storage capacity is much less compared to the data. Thus a revolution reconstruction of information framework is needed very much for this we need to design a hierarchical architecture for storage. The heterogeneous data are not efficiently handled by the efficient.
Algorithms that exit now and thus we need to even design a very efficient algorithm for the effective handling of the heterogeneous data.

[5]9.1 NECESSITY OF SECURITY IN BIG DATA

The big data is used by many of the business but they may not have assets from perspective of the security .if any security occurs the big data it may come out with even more serious issues .now a day's companies uses this technology to store data of peta byte range regarding to the company business and customers. This result in severe criticality for as in classification of information to secure the data we either need to encrypt log or use honey pot techniques. The challenge of ducting threats and malicious introduces must be solved using big data style analysis.

9.2 PROPOSED APPROACHES FOR SECURITY OF BIG DATA IN CLOUD COMPUTING ENVIRONMENT:

Here we present few security measurements that can be used to improve the cloud computing environment.

9.2.1 Encryption:
since the data in any system will be present in a cluster ,a hacker can easily steal the data from the system. This may become serious issue for any company or organization to safeguard their data. To avoid this we may go for encrypting the data.

9.2.2 Nodes authentication:
The node must be authenticated whenever it joins the cluster. If the nodes turns out to be a malicious cluster then such node must be authenticated.

9.2.3 Honey pot nodes:
The honey pot nodes appear to be like a regular node but is a trap. It automatically tracks the hackers and will not allow any damage to happen to the data.

9.2.4 Access control:
[3]The differential privacy and access control in the distributer environment will be a good measure of security. To prevent the information from leaking we use a SELinux. The security Enhanced Linux is a feature that provides the mechanism for supporting access control security policy through the use of linux security, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Hence security issues of these systems and technologies are applicable to cloud computing.
The challenges of security in cloud computing environment can be categorized into network level, user authentication level,
data level, and generic issues.
Networklevel*:* The challenges can be categorized under a network level deal with network protocols and network securitysuch as distributed nodes ,distribute data and inter *node communication.*

Authentication level: *T*he challenges that can be categorized under data level deals with data integrity and availability such asdata protection and distributed data.

Data level: The challenges that can be categorized under data level deals with data integrity and availability such as dataprotection and distributed data.

Generic types: The challenges that can be categorized under general level are traditional security tools, and use of different technologies

A. ***10. TECHNICAL CHALLENGES IN BIG DATA***

[5]Whenever new technologies evolve, they meet with new challenges in all the aspects. Once the functional challenges are in place, the next kin is the technical challenges. The big data faces many technical challenges which are on the road way of the research.

10.1 Failure handling :

Devising 100% reliable systems on the go is not the easy task system can be devised in such away that the probability of failure must fall within the permitted threshold .fault tolerance is the technology challenge in the big data. when process started it may involve with numerous network nodes and the whole computation process becomes cumbersome retaining the checkpoints and fixing the threshold level for process restart in case of failure, are greater concerns.

10.2 Data heterogeneity:

Big data deals with unstructured, semi structured and structured data. Linking unstructured data with the structured data, converting data from one form into another required form needs a lot of research.

11. BIG DATA IN CLOUD

Storing and processing big volumes of data requires scalability, fault tolerance and availability. Cloud computing delivers all these through hardware virtualization. Thus, big data and cloud computing are two compatible concepts as cloud enables big data to be available, scalable and fault tolerant. Business opportunity. As such, several new companies such as Cloudera, Terradata and many others, have started to focus on delivering Big Data as a service(BDaaS) or database as a service(DBaaS). Companies such as Google ,

[4]IBM, Amazon and Microsoft also provide ways for consumers to consume big data on demand. Next, we present to examples, Nokia and RedBus, which discuss the successful use of big Data within Cloud environment. Cloud comes with an explicit security challenge that is the data owner might not have any control of where the data is placed.
The reason behind this control issue is that if one wants to get the benefits of cloud computing, he/she must also utilize the allocation of resources and also the scheduling given by the controls. Hence it is required to product the data in the midst of untrustworthy processes.
Since cloud involves extensive complexity, we believe that rather than providing a holistic solution to securing the cloud, it would be ideal to make noteworthy enhancements in securing the cloud that will ultimately provide us with a secure cloud.

11.CONCLUSION

Cloud computing enables small to medium sized business to implement big data technology with a reduced commitment of company resources. The processing capabilities of the big data model could provide new insights to the business pertaining to performance improvement, decision making support, and innovation in business models, products, and services. Benefits of implementing big data technology through cloud computing are cost savings in hardware and processing, as well as the ability to experiment with big data technology before making a substantial commitment of company resources. Several models of cloud computing services are available to the businesses to consider, with each model having trade-offs between the benefit of cost savings and the concerns data security and loss of control.

References

[1]. The Search for Analysts to make Sense of Big Data. Yuki Noguchi. National Public Radio, Nov.30,2011

[2]. S. Justin Samuel, Koundinya RVP, Kotha Sashidhar and C.R. Bharathi, "A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES"Asian Research Publishing Network Vol.10, No.8, May 2015.

[3]. Jui-Chien Hsieh , Ai-Hsien Li and Chung-Chi Yang "Cloud, and Big Data Computing",Int.J.Environ.Res.Public Health 2013,10,6131-6153.

[4] Venkatesh H, Shrivatsa D Perur, NiveditaJalihal "A Study on Use of Big Data in CloudComputing Environment", international journal of computer science and information technology Vol. 6 (3) , 2015, 2076-2078.

[5] Santosh Kumar Majhi1 and Gyanaranjan Shial "Challenges in Big Data Cloud Computing", smart computing review, vol. 5, no. 4, August 2015.

[6].Marcos D. Assunção, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto, RajkumarBuyya "Big Data computing and clouds:Trends and future directions", journal of parallel and distributed Computing. Aug 25 2014.