International Journal of Business Quantitative Economics and Applied Management Research

Volume-3, Issue-4, September-2016 ISSN No: 2349-5677

A BRIEF STUDY ON USE OF BIG DATA IN CLOUD COMPUTING ENVIRONMENT

Ketaki S. Pathak
Asst. Prof. Ashoka Center for Business and Computer studies, Nashik, India
Ketaki.s.pathak@gmail.com

Abstract

Cloud computing is a powerful technology to perform massive-scale and complex computing. It eliminates the need to maintain expensive computing hardware, dedicated space, and software. Massive growth in the scale of data or big data generated through cloud computing has been observed. Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The rise of big data in cloud computing is reviewed in this study. 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 summarized. This paper introduces several big data processing techniques from system and application aspects here provide an organized picture of challenges that are focused by the application developers and DBMS designers in developing cum deployment of the internet scale applications. Then we see about the security issues in the cloud computing along with the big data and Hadoop. We show some possible solutions for the issues of the cloud computing and Hadoop.

Keywords-Big Data; Cloud Computing;

I. INTRODUCTION

The successful paradigm for the service oriented programming is the cloud computing. It has revolutionized the way of computing infrastructure’s abstraction and usage. The elasticity, pay per use, low upfront investment, transfer of risks are few of the major enabling characteristics that makes the cloud computing the ubiquitous platform for deploying economically feasible enterprise infrastructure settings. Distributed databases had been the boon of vision for research for few decades. But changes in the data patterns and applications has made way for the new type of storage called key value storage which are now being widely used by various enterprises. In the domain of Map reduce and open source implementation of the same known as the Hadoop has been used by majority of the industry and academics. Hadoop increases the usability and performance. HDFS has become a Very helping tool to maintain and store the complex data. Big data has becoming more available and understandable to computers. What is big data? The question arrives. Big data is the representation of progress of the human cognitive processes, usually includes data sets with sizes that is beyond the current technology's capability. The data which is very fast, has various varieties and requires new type of the processing forms to enable decision making, insight discovery and optimization of process. In order for analyzing the data and for identification of patterns it is very important for us to store the data securely, manage and sharing of complex data on cloud. Since cloud involves extensive complexity, we feel its ideal to make enhancements in securing cloud than showing holistic solutions.

In this paper we provide a comprehensive background study of state of art systems. Identification of critical aspects in design of various systems and scope of the systems. We show up some approaches in security provision through a scalable system to handle large number of sites and also has the capability to process large and massive amounts of data. We also provide the status of big data studies and related works, aiming at providing a overview of managing big data and its applications.

II. BIG DATA

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 tech ologies 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.

The big data is defined using Three V's

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

2) Variety: various types of data is to be supported.

3) Velocity: the speed at which the files are created and processes are carried out refers to the velocity.

Fig 1 shows a typical bug data representation./ The areas for example that comes in big data are shown.

Technologies not only supports the collections of large amounts such data effectively. Transactions that are made all over the world in a Bank, Walmart customer transactions, and Facebook users generating social interaction data Are few examples for big data usage.

III. HADOOP

This is a freely available java based programming framework supporting for the processing of large sets of data in a distributed computing environment. Using Hadoop, big amount of data sets can be processed over cluster of servers and apps may be run on system with thousands of nodes involving terabyes of information. This lowers the risk of system failure even when a huge number of nodes fail.it enables a scalable, flexible, fault tolerant computing solution. HDFS, a file system spanning all nodes in a Hadoop cluster for data storage links the file systems on local nodes to make it onto a very large file system thus improving the reliability.

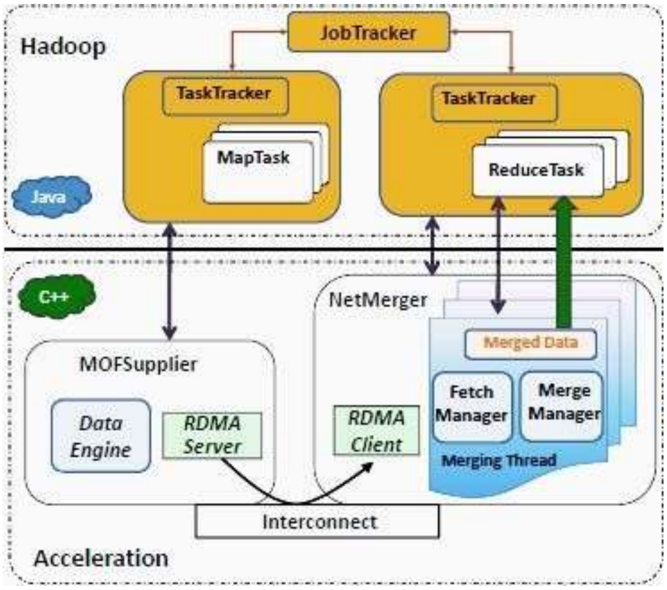

Fig 2: Hadoop structure

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

a. Job trackers has two primary responsibilities which are managing the cluster resources and scheduling all user jobs.
b. Data engine consists of all the information about the processing the data.
c. Fetch manager helps to fetch the data while particular task is running.

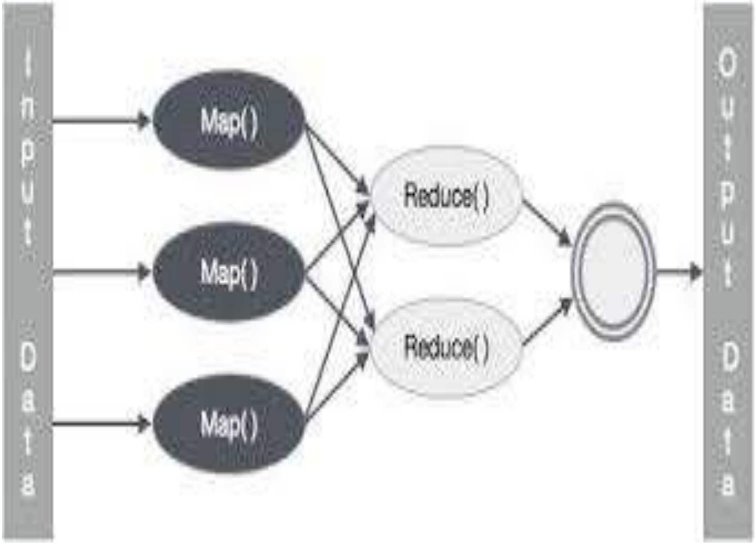

Fig 3: Map reduce

IV. MAP REDUCE

Map reduce framework 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 framework and then sent to the reduce tasks. The monitoring is taken care by the framework.

The input data is divided into individual chunks and are 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 the reduce task where the results that are generated in parallel by the map task are consolidated and the reduced report is given as output.

Big data applications:

1. 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 big data. Map reduce is a very popularly used tool or model used in industry and academics.
2. 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 the reading point.
3. Map reduce has received a lot of attentiveness in the fields of data mining, information retrieval, image retrieval etc.
4. The computation becomes difficult to be handled by traditional data processing which triggers the development of big data apps.

5. 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.

V. ADVANTAGES OF BIG DATA

1. Scalable

Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data.

2. Cost effective

Hadoop also offers a cost effective storage solution for businesses' exploding data sets. The problem with traditional relational database management systems is that it is extremely cost prohibitive to scale to such a degree in order to process such massive volumes of data. In an effort to reduce costs, many companies in the past would have had to down-sample data and classify it based on certain assumptions as to which data was the most valuable. The raw data would be deleted, as it would be too cost-prohibitive to keep. While this approach may have worked in the short term, this meant that when business priorities changed, the complete raw data set was not available, as it was too expensive to store. Hadoop, on the other hand, is designed as a scale-out architecture that can affordably store all of a company's data for later use. The cost savings are staggering: instead of costing thousands to tens of thousands of pounds per terabyte, Hadoop offers computing and storage capabilities for hundreds of pounds per terabyte.

3. Flexible

Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. This means businesses can use Hadoop to derive valuable business insights from data sources such as social media, email conversations or clickstream data. In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection.

4. Fast

Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing. If you're dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes, and petabytes in hours.

5. **Resilient to failure**

A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use.

The MapR distribution goes beyond that by eliminating the NameNode and replacing it with a distributed No NameNode architecture that provides true high availability. Our architecture provides protection from both single and multiple failures.

When it comes to handling large data sets in a safe and cost-effective manner, Hadoop has the advantage over relational database management systems, and its value for any size business will continue to increase as unstructured data continues to grow.

The big data allows an individual to analyze the threats he/she faces internally by naooing onto the entire data landscape over the company using the rich set of tools that the software supporting the big data provides. This is an important advantage of big data since it allows the user to make the data safe and secure. The speed, capacity and scalability of cloud storage provides a mere advantage for the company and organization. Big data even allows the end users to visualize the data and companies can find new business opportunities. Data analytics is one more notable advantage of the big data where in which the individual is allowed to personalize the content or to look and feel the real time websites.

VI. CHALLENGES AND DISCUSSIONS

We live in the period of the big data where we can gather more information from daily life of human being. So far, researchers are unable to unify the features that are more essential to big data, many think that big data is something which we cannot process using existing technology, theory or any methods of such kind. However the world has become helpless since enormous amount of data is being generated by science,business and even society. Big data has posed many challenges to the IT industry.

VII. BIG DATA 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 exist now and thus we need to even design a very efficient algorithm for the effective handling of the heterogeneous data.

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 threat occurs to big data, it may come out with even more serious issue. Nowadays, companies use this technology to store data of petabyte range regarding to the company, business and customers. This result in severe criticality for classification of information.to secures the data we either need to encrypt, log or use honeypot techniques. The challenge of detecting threats and malicious intruders, must be solved using big data style analysis.

Analysis and computation of big data:
Speed is the main thing when we look up for querying in the big data. However the process may be time consuming only because of the reason that it cannot traverse all related data in the whole database in a short time. While the big data is getting complicated, the indices in the big data are aiming at the simple type of the data. The traditional serial algorithm is inefficient for this big data.

VIII. PROPOSED APPROACHES FOR SECURITY OF BIG DATA IN CLOUD COMPUTING ENVIRONMENT

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

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 a serious issue for any company or organization to safeguard their data. To avoid this, we may go for encrypting the data. Different encryption mechanisms can be used in different systems and the keys generated should be stored secretly behind firewalls. By choosing this method the data of the user may be kept securely.

2. Nodes authentication:

The node must be authenticated whenever it joins the cluster. If the node turns out to be a malicious cluster then such nodes must not be authenticated.

3. Honeypot nodes:

The honeypot nodes appears to be like a regular node but is a trap. It automatically traps the hackers and will not allow any damage to happen to the data.

4. Access control:

The differential privacy and access control in the distributed environment will be a good measure of security. To prevent the information from leaking we use a SELinux[17]. The Security Enhanced Linux is a feature that provides the mechanism for supporting access control security policy through the use of linux Security modules in linux kernels.

IX. CONCLUSION

This paper gave a description of a systematic flow of survey of the big data in the environment of cloud computing. We discussed about the applications, advantages and challenges faced by big data when used over a cloud computing environment. We proposed few solutions to safeguard the data in the cloud computing environment. In future, the challenges are need to be overcome and make way for the even more efficient use of the big data by the user on a cloud computing environment. It is very much needed that the computer scholars and IT professionals to cooperate and make a successful and long term use of cloud computing nd explore new ideas for the usage of the big data over cloud environment.

REFERENCES

1. D. Borthakur, "The hadoop distributed file system: Architecture and design," Hadoop Project Website, vol. 11, 2007.
2. The Apache Hadoop Project. http://hadoop.apache.org/core/, 2009.
3. Abouzeid, K. B. Pawlikowski, D. J. Abadi, A. Rasin, and A. Silberschatz. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. PVLDB,2(1):922–933, 2009.
4. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony,H.Liu, P. Wyckoff, and R. Murthy. Hive - A Warehousing Solution Over a Map-Reduce Framework. PVLDB, 2(2):1626–1629, 2009.
5. A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug 2013.
6. K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.
7. Wie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.
8. F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Busines on big dataapplications." Big Data, 2013 IEEE International Conference, Silicon Valley, CA, Oct 6-9, 2013, pp.32 - 37.
9. Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." Open Cirrus Summit (OCS), 2012 Seventh, Beijing, Jun 19-20, 2012, pp. 48 – 52
10. Venkata Narasimha Inukollu , Sailaja Arsi and Srinivasa Rao Ravuri "Security issues associated with big data in cloud computing "International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014
11. Andrew Pavlo, "A Comparison of Approaches to Large-Scale Data Analysis", SIGMOD, 2009.
12. Apache Hadoop: http://Hadoop.apache.org
13. shilpa Manjit Kaur," BIG Data and Methodology- A review" ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.
14. Dean, J. and Ghemawat, S., "MapReduce: a lexible data processing tool", ACM 2010.
15. DeWitt & Stonebraker, "MapReduce: A major step backwards", 2008.

16. Hadoop Distributed File System, http://hadoop.apache.org/hdfs
17. HadoopTutorial: http://developer.yahoo.com/hadoop/tutorial/module1.html
18. J. Dean and S. Ghemawat, "Data Processing on Large Cluster", OSDI '04, pages 137–150, 2004
19. J. Dean and S. Ghemawat,"MapReduce: Simplified Data Processing on Large Clusters", p.10, (2004).
20. Jean-Pierre Dijcks, "Oracle: Big Data for the Enterprise", 2013.