Statistical Analysis of Big Data Using Hadoop: A Review

Shaili[1], Durgesh Srivastava[2,] Deepak Sinwar[3]
[1]Research Scholar, [2]Asstt. Professor, [3]Asstt. Professor
[1,2,] Department of Computer Science & Engineering
BRCM College OF Engineering and Technology Bahal (Haryana)
[3]Amity School of Engineering and Technology, Delhi
shaili522@gmail.com[1], dsrivastava@brcm.edu.in[2],deepak.sinwar@gmail.com[3]

Abstract: -Data analysis is the combination of concepts, tools and algorithms of machine learning and statistics to analyze very large data sets, so as to gain insight, understanding and effective knowledge and it is applied for this purpose in many companies.There is no doubt that "Big Data" analytics is still in the initial stage of development, since existing "Big Data" techniques and tools are very limited to solve the real problems. For energy optimization using MAP preduce applications, we will use Apache Hadoopframework that is used for distributed processing of large data sets across clusters ofsystems using simple programming models.

Keywords: - Big Data, HDFS, CaPC, EMRSA, DBMS, MFF, SQL, HPCC, BI

I. INTRODUCTION

Big data analysis is the process of inspecting large and varied data sets i.e., big data to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions. The emerging Big Data Science term, showing its broader impact on our society and in our business life cycle, has insightful transformed our society and will continue to attract diverse attentions from technical experts and as well as public in general [1] [2]. It is obvious that we are living in Big Data era, shown by the sheer volume of data from a variety of sources and its rising rate of generation. For instance, an IDC report predicts that, from 2005 to 2020, the global data dimensions will grow by a factor of 300, from 130 Exabyte's to 40,000 Exabyte's, representing a double growth every two years. This is focuses on accessible big-data systems that include a set of tools and technique to load, extract, and improve dissimilar data while leveraging the immensely parallel processing power to perform complex transformations and analysis[3][4].

II. BIG-DATA SYSTEM ARCHITECTURE

A big-data system is complex, providing functions to deal with different phases in the digital data life cycle, ranging from its birth to its destruction. At the same time, the system usually involves multiple distinct phases for different applications.

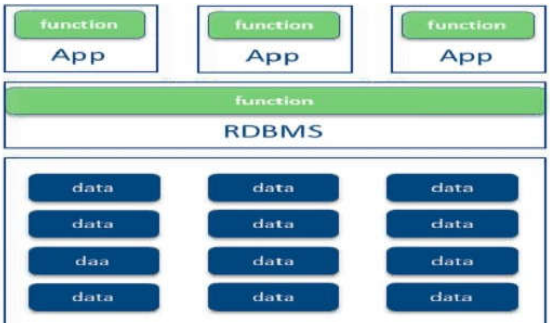

Fig1: Architecture of big Database

III. ADVANTAGES AND DISADVANTAGES

Advantages:

- It can point the way to various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals.
- It can analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs.
- It encompasses a mix of semi-structured and unstructured data.
- Data analytics technologies and techniques provide a means of analyzing data sets and drawing conclusions about them to help organizations make informed business decisions.

Disadvantage

- Heterogeneity is the big challenge in data Analysis and analysts need to cope with it.
- Managing with large data sets is a big problem from decades.
- Another challenge with size is speed. If the data sets are large in size, longer the time it will take to analyze it.
- Privacy of data is another big problem with big data.
- In spite of the advanced computational models, there are many patterns that a computer cannot detect. A new method of harnessing human ingenuity to solve problem is crowd-sourcing.

IV. OPPORTUNITIES TO BIG DATA

Technology:
Almost every top organization like Facebook, IBM, and yahoo have adopted Big Data and are investing on big data. Facebook handles 50 Billion photos of users.
Government: Big data can be used to handle the problems faced by the government Obama government announced big data research and development initiative in 2012. Big data analysis played an important role of BJP winning the elections in 2014 and Indian government is applying big data analysis in Indian electorate

Healthcare:
Healthcareorganizations are adapting big data technology to get the complete information about a patient

Science and Research: Big data is a latest topic of research. Many researchers are working on big data. NASA center for climate simulation stores 32 peta bytes of observations

Media:
Media is using big data for the promotions and selling of products by targeting the interest of the user on internet.

V. TECHNOLOGIES AND TOOLS

Unstructured and semi-structured data types typically don't fit well in traditional data warehouses that are based on relational databases oriented to structured data sets. Some important technologies and tools for big data are:

- **YARN:** A cluster management technology and one of the key features in second-generation Hadoop.
- **MapReduce:** A software framework that allows researchers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.
- **Spark:** An open-source parallel processing framework that enables users to run large-scale data analytics applications across clustered systems.

- **HBase:** A column-oriented value data store built to run on top of the Hadoop Distributed File System (HDFS).
- **Hive:** An open-source data warehouse system for querying and analyzing large datasets stored in Hadoop files.
- **Kafka:** A distributed publish-subscribe messaging system designed to replace traditional message brokers.
- **Pig:** An open-source technology that suggest a high-level mechanism for the parallel programming of Map Reduce jobs to be executed on Hadoop clusters.

VI. WHY BIG DATA

Use of big data is very effective on high level and handling of this data is very challenging in today's life. Instead of these challenges importance of big data never reduces because:

Cost reduction: Big data technologies are cost effective when it comes to storing large amounts of data and identification more efficient ways of doing business.

Faster, better decision making: The in-memory analytics, combined with the ability to analyze new sources of data and speed .Businesses are able to analyze information immediately and make decisions based on what they've learned.

New products and services: The ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

Fig 2: Source of Big Data

VII. PREVIOUS WORK

Previously research on big data analysis is discussed below:

In paper 1] author focused on Hadoop and HDFS by Apache is for storing and managing BigData.DBMS techniques like Joins and Indexing and other techniques like graph search is used for classification and clustering of Big Data. Map Reduce is used for minimization [1]

In paper [2] author focused on For big data various techniques like Hadoop, Map Reduce, Apache Hive, No SQL and HPCC are used.These technologies handle massive amount of data in MB, PB, YB, ZB, KB and TB.The need of big data generated from facebook, yahoo, Google, YouTube etc. [2]

In paper [3] author focused on Energy-efficient routing problem is resoled using greedy heuristic called GEERA.Proposed scheme can optimize energy at large scale as compared with the traffic engineering based solutions. [3]

In paper [4] author focused on Author focused on the energy consumption by Map Reduce scheduling methods and proposed heuristic based methods.Different workloads with various benchmarks applications Tera-Sort, Page Rank, and K-means clustering etc. [4]

In paper [5] author focused on Author focused on the energy consumption by Map Reduce scheduling methods.It reduce the energy consumption up to 40%. [5]

In paper [6] author focused on Author developed a function called memory fast-forward (MFF).It process the graph computations with optimal memory requests. It reduce 54.6% energy consumption due to low memory traffics. [6]

In paper [7] author focused on Author analyze and construct the information regarding energy consumption of each user, called ESA. Authors did a real time analysis to show its performance in terms of accurate decision making about power consumption. [7]

In paper [8] author focused on Author proposed a framework to enhance the memory consumption for data.Author focused on the processing of the data as per their sensitivity, in cloud environment .Proposed scheme can be extended for data value characterization algorithm. [8]

In paper [9] author focused on defined a Content-aware, Partial Compression (CaPC) for text using a dictionary-based method.The performance interms of size reduction upto30% and performance enhancement of I/O jobs upto 32%. [9]

In paper [10] Author use energy-aware scheduling of Map Reduce jobs and used a greedy method, called Energy-aware Map Reduce Scheduling Algorithm (EMRSA). Proposed scheme can be extended to provide the support for multiple Map Reduce jobs. [10]

In paper [13] author focused on A new scientific paradigm is born as data-intensive scientific discovery (DISD), also known as Big Data problems.The main objective of this paper is emphasizing the significance and relevance of Big Data in business system, society administration and scientific research.Author purposed potential techniques to solve the problem, including cloud computing, quantum computing and biological computing. [13]

In paper [14] author focused on the various challenges and issues in adapting and accepting Big data technology.Author also discussed in detail along with the problems Hadoop is facing.Author shown tool to resolve this problem by showing Comparison of Hadoop Technique with other system Techniques. [14]

In paper [15] author focused on in this paper classifications of four types of inconsistencies in big data and point out the utility of inconsistency.Learning about tool for big data analysis.Analysis is desired to establish correspondence between big data analysis tasks and types of inconsistencies impacting or affecting those tasks. [15]

VIII. CONCLUSION

The high-performance data visualization is providing a better way to analyze data more quickly than ever. The trends of big data generation can be characterized by the data generation rate. Specifically, the data generation rate is increasing due to technological advancements. Systems were widely adopted; many management systems in various organizations were storing large volumes of data, such as bank trading transactions, shopping mall records, and government sector archives. These data sets are structured and can be analyzed through database-based storage management systems.

REFERENCES

[1] M. Dhavapriya, et al. , "Big Data Analytics: Challenges and Solutions Using Hadoop, MapReduce and Big Table", International Journal of Computer Science Trends and Technology(IJCST) – Volume 4 Issue 1, PP 5-14 2016.

[2] Varsha B.Bobade , "Survey Paper on Big Data and Hadoop" International Research Journalof Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 03 Issue: 01 PP 861ISSN: 2395-0072 2016

[3] Rahul Beakta, "Big Data And Hadoop: A Review Paper" Baddi University of EmergingSciences& Technology, Baddi, India Volume 2, Spl. Issue 2 ISSN: 1694-2329 2015.

[4] Ms. Gurpreet Kaur Et Al., "Review Paper On Big Data Using Hadoop" International Journal of Computer Engineering & Technology (IJCET) Volume 6, Issue 12, pp. 65-71,ISSN0976–6375 2015.

[5] Jennifer Ortiz, Victor Teixeira de Almeida, Magdalena Balazinska, "Changing the Face ofDatabase Cloud Services with Personalized Service Level Agreements", 2015.

[6] AnantBhardwaj, SouvikBhattacherjee, AmitChavan, AmolDeshpande, Aaron J.Elmore, Samuel Madden, AdityaParameswaran, "DataHub: Collaborative DataScience& Dataset Version Management at Scale", 2015.

[7] Andrea Acquaviva, Daniele Apiletti, Antonio Attanasio, Elena Baralis, Lorenzo Bottaccioli, "Energy signature analysis: knowledge at your fingertips", IEEE International Congress on Big Data, IEEE-2015, pp.543-550

[8] ThiThao Nguyen Ho, Barbara Pernici, "A Data-Value-Driven Adaptation Framework for Energy Efficiency for Data Intensive Applications in Clouds", IEEE Conference on Technologies for Sustainability (SusTech), IEEE-2015,pp.47-52

[9] Dapeng Dong, John Herbert, "Content-aware Partial Compression for Big TextualData Analysis Acceleration", International Conference on Cloud Computing Technology and Science, IEEE-2014, pp. 320-325

[10] Lena Mashayekhy, MahyarMovahedNejad, Daniel Grosu, Dajun Lu, Weisong Shi, "Energy-aware Scheduling of MapReduce Jobs", IEEE-2014, pp.32-39

[11] KarthiDuraisamy, Ryan Gary Kim, Wonje Choi, Guangshuo Liu, ParthaPratimPande, RaduMarculescu, Diana Marculescu, "Energy Efficient MapReduce with VFI-enabled Multicore Platforms", Design Automation Conference (DAC), ACM/EDAC/IEEE-2015, pp.1-6

[12] Han hu (Fellow, IEEE)," Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", IEEE 2169-3536(2014),PP 652-687

[13] C.L. Philip Chen, Chun-Yang Zhang, "Data intensive applications, challenges, techniques and technologies: A survey on Big Data" Information Science 0020-0255 (2014), PP 341-347, elsevier

[14] AvitaKatal, Mohammad Wazid, R H Goudar, "Big Data: Issues, Challenges, Tools and Good Practices", IEEE 978-1-4799-0192-0/13,PP 404-409

[15] Du Zhang," Inconsistencies in Big Data", IEEE 978-1-4799-0783-0/13, PP 61-67