International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 6, ISSN No. 2455-2143, Pages 113-118
Published Online October 2021 in IJEAST (http://www.ijeast.com)

BIG DATA ANALYSIS AND CLOUD IN THE APPLICATIONS OF HEALTHCARE

Gangadevi M D
Department of CS
BU, Coimbatore,
Tamil Nadu, India

Abstract— **Big Data that either is too large, grows too fast or does not fit into traditional architectures. Within such data, there can be valuable information that discovered through data analysis. Big data now available for analytics present complex, daunting challenges due to the vast number of digital data generated daily by different organizations. The vast amount of data has improved the global community's ability to defend and allow for progress of rights of vulnerable people around the globe. Moreover, if big data processing is to improve lives, its existing data gathering methods should assist humanitarian affairs and not replace them. Therefore, finding ways to increase humanitarian services with data, highlighting the importance of big data are critically important. Difficulties include capture, storage, search, sharing, analytics, and visualizing. In this paper, we analyze four different papers and reveal the benefits of Big Data Analytics and Cloud in the applications of Healthcare where the data flow to and from is in massive volume.**

Numerous organizations have escaped the frustrations of their first-generation data warehouses by replacing older database technologies with significant data which is unstructured in a scalable, error tolerant and efficient way. The purpose of this paper is to propose a big data platform for large-scale data analysis by using the Map Reduce framework for unstructured data stored into integrating distributed clustered systems such as NoSQL (NotOnlySQL) and Hadoop Distributed File System (HDFS).

Keywords— Analytics, Big data, Cloud, Big Data Analytics Hadoop, HDFS, Map Reduce, NoSQL

I. INTRODUCTION

In this Information Age, information unceasingly created all over the world around the clock. Businesses and organizations want to run most of its business processes using big data technology, created in the class of transactions and interactions. Through emails, videos, and images, for example, goods and services are producing the huge amount of data, with the Internet becoming a very significant user interface for interactions.

Telecommunications Network providers get a broad degree of data in the form of conversations. Also, social network sites like Facebook have begun acquiring terabytes (TBs) of data every day in the form of comments, blogs, tweets, photos, audio, and videos among others. Internet-based companies in this digital age generate the vast level of on-line streaming data daily. For example, in a medical setting, one can imagine a situation where a patient complained about a particular set of symptoms, the doctor could track the cases accounting for these across all past patients and understand such comparable symptoms and how an individual patient may respond to different treatment. Consequently, health data about patients, diseases and the data produced by various medical devices will be massive. In addition, data generated from different machines in the production sectors like transport, war ammunitions, finance and many more are similarly a source of massive data with each being stored for a different reason for future use. Organizations can process this data, analyse it and store it for intelligent decision-making to gain a highly competitive benefit over their contemporary. The influence question that comes up is how we process, store and manage such a great size amount of data most of which is Unstructured. Although there are major categorizations of big data platforms to store, process and manage them on a possible scale, functional, effective and error tolerant and efficient fashion.

II. RELATED WORK

In some of the current big data platforms for big scale data analysis, there are several types of vendor commodities to consider for big data analytics. Now vendors have contributed analytic platforms established upon Map Reduce and distributed file system. The Vertica Analytic Platform offers a robust and ever-growing set of Advanced In-Database Analytic functionality. It has a high-speed relational SQL database management system (DBMS) purpose-built in analytic and business intelligence. It offers a shared-nothing Massive Parallel Processing (MPP) column-oriented architecture [1]. IBM Info Sphere Big insights represents a fast, robust, and easy-to-use platform for analytic on Big Data at rest. For example, IBM offers a platform for big data including IBM Info Sphere Big insights and IBM Info Sphere Streams. IBM Info Sphere Streams are a powerful analytic computing platform that delivers a platform for analysing data in real time with micro-latency [2]. And EMC Greenplum is driving the future of data warehousing and analytics with discovery products including the Greenplum Data Computing Appliance, Greenplum Database, Greenplum HD enterprise-ready Apache Hadoop and Greenplum Chorus. The SAND Analytic Platform is a columnar analytic database platform that achieves linear data scalability through massively parallel processing (MPP), breaking the constraints of shared-nothing architectures with fully distributed processing and dynamic allocation of resources [3]. Pavlov et al. [4] this described and compared Map Reduce structure and parallel DBMS for large-scale data analysis and defined a benchmark consisting of tasks run on an open-source version of MR as well as on two parallel DBMS.

A new scientific paradigm is born as data intensive scientific discovery (DISD), also known as Big Data problems. Many fields and sectors ranging from economic and business activities to public administration from national security to scientific research in many areas involve with Big Data problems.

Another aspect of big data to consider is the cost to both store and process these massive amounts of data. Cloud computing can be a possible solution as it provides a solution that is cost efficient while meeting the need of rapid scalability - an important feature when dealing with big data. However, even in cloud computing, big data analysis is not without its problems. The Data Stream Management Systems (DSMS) aims to reduce the amount of bad data from being collected thus reducing future costs for storage and processing as well as finding potential valuable information and patterns sooner in the data analysis pipeline.

Datasets that stretch the limits of traditional data processing and storage systems often referred to as Big Data. The need to process and analyse such massive datasets has introduced a new form of data analytics called Big Data Analytics. It includes analysing huge measure of data of a mixture of types to reveal hidden blueprint, unidentified association, and other useful information. The distributed processing of outsized data sets across groups of systems facilitated by using simple computing models of the Apache Hadoop Framework. Achieving better outcomes at lower costs has become very important for healthcare and Big Data Analytics and Hadoop's presence are positively part of the solution in reaching that goal. HBase is a column-oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases.

When considering big data techniques and approaches to store the data need to be re-evaluated. For storing large volumes of data, distributed file systems are a possible solution. One example of this type of system would be the distributed file system coupled with the Map Reduce engine in Apache's Hadoop project [5]. Other challenges in data storage are due to the variety of structured and unstructured data. One approach to this problem addressed by NoSQL databases. NoSQL databases are characteristically non-relational and typically do not provide SQL for data manipulation.

In comparison to the traditional relational DBMS, NoSQL databases eschew the management to the application as it aims to provide a highly scalable database. Furthermore, these databases are schema less, allowing changes to the data structure rapidly rather than having to do table rewrites. This gives an advantage over relational DBMS regarding horizontal scalability through the cloud. As a response to the scalability that NoSQL introduced, a new class of New SQL databases has developed that follow the relational model but either distributes the data or transaction processing across nodes in a cluster to achieve comparable scalability [6].

Systems need to keep modular such that analytic tools and approaches that applied when needed. Other factors have outlined as to what it has considered when constructing a system for the consumption of big data. These would include the format and size of the data ingested the types of analytics conducted and the objectives of the system. The data size and format are a major factor for deciding the architectural components for storing data in the system. Doing so can avoid making decisions on a non-relational model for its scalability when the size of data does not warrant a need [7].

It demonstrates a close-up view about Big Data, including Big Data applications, Big Data opportunities and challenges, as well as the state-of-the-art techniques and technologies that currently adopt to deal with the Big Data problems also known as data- intensive scientific discovery (DISD). Data-intensive science [8] is emerging as the fourth scientific paradigm in terms of the previous three, namely empirical science, theoretical science, and computational science. Simulations in large of fields generate a huge volume of data from the experimental science, at the same time, more and more large data sets generated in many pipelines. There is no doubt that the world of science has changed just because of the increasing data-intensive applications. Therefore, data-intensive science has viewed as a new and fourth science paradigm for scientific discoveries [9].

The current trends and characteristics of Big Data, its analysis and how these are presenting challenges in data collection, storage and management in cloud computing are examined. The Data Stream Management Systems (DSMS) aims to reduce the amount of bad data from being collected, thus reducing future costs for storage and processing as well as finding potential valuable information and patterns sooner in the data analysis pipeline [10].

It analyses and reveals the benefits of Big Data Analytics and Hadoop in the applications of Healthcare where the data flow to and from is in massive volume.

Hadoop provides solution to analyze the piling up medical images from various sources and extracts the necessary data to give right diagnosis. Hadoop image processing interface (Hipi) gives an API for processing images in the distributed computing environment [11] [12].

It demonstrates the rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time". Hadoop Map Reduce is the best tool available for processing data and its distributed, column-oriented database, HBase that uses HDFS for its underlying storage and support provides more efficiency to the system [13].

III. ROLE OF BIG DATA AND BIG DATA ANALYTICS

These are large complex system requiring efficient algorithms to process these raw data and require huge computational power. Big data refers to the data generated from different sensors this includes medical, traffic and social data.

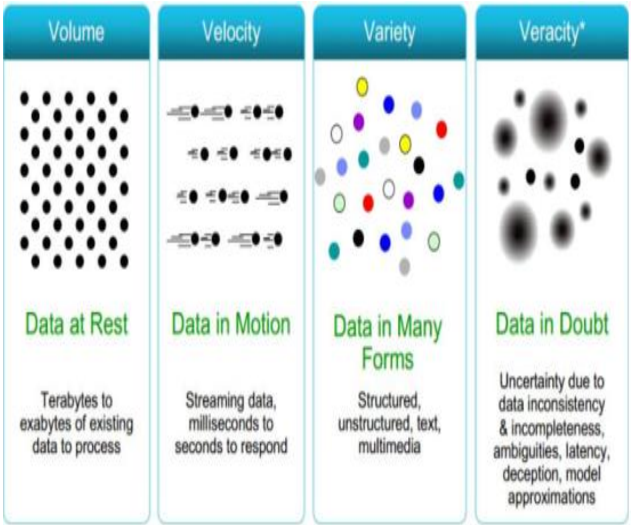
Volume
Velocity
Variety
Veracity*
Data at Rest
Data in Motion
Data in Many Forms
Data in Doubt
Terabytes to exabytes of existing data to process
Streaming data, milliseconds to seconds to respond
Structured, unstructured, text, multimedia
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Fig 1: Big Data – 4 "Vs"

3.1. Big Data

Progressively establishments today are facing increasingly Big Data stimulating situations. Big Data holds for the data that cannot processed or analysed employing traditional processes or carry out such practices. Many organizations have access to a huge amount of information, because they have no such idea whether it is deserving holding. There is major four dimensions (V's) features of Big Data: volume, variety, and velocity, veracity.

1) **Volume:** Volume is the first and most notorious feature because many organizations are giving rise to a very large volume of data internally or assembling other vast amounts of data from the outer side.

2) **Variety:** In an organization today, there are many ways, whereby data collection has increased. This has caused the rise inside and outside source of data non-structured, such as plain-text documents, electronic sensor, tweets, blogs, and social media.

3) **Velocity:** Data warehouse is traditional types of the result, therefore, in traditional methods, information has been often raw and inevitable to employ and acted according to share period frames to get the best potential value from it and this makes the real-time answers to a common need in advance establishments. There is major two types of big data: the data on balance, e.g., social media, emails, weblogs, and non-structured plain-text documents are all gathering of streamed function. So, data gathered in motion, e.g., twitters comments and sensor information (data).

4) **Veracity:** Big data like social media data (e.g., Tweets or Facebook Posts), how much should we put in the data, inaccurate data that is directly different in traditional data warehouses, where it was always the assumption that the data assure, clean, and correct. That is why so practically time on Data Lineage, Master Data Management, ETL/ELT, and Identity Insight/Assertion, etc.

3.2. Big Data Analytics

Big data analytics are a boosting analytic practical method to very big data sets because both structured and unstructured of big data can meet data from product data sources, which include network and mobile devices and the web technologies. Progress analytic is a gathering of practical techniques, which includes data mining, complex SQL, data visualization, artificial intelligence, and predictive analytics. For example, database techniques that hold analytic, such as, Map Reduce, columnar data stores and in-memory database.

3.3. Big Data Storage

Many establishments are finding it difficult to discuss the increasing data intensities i.e., big data plainly cause the problem more. Moreover, to solve this difficulty, big data establishments need to cut the measure of data being stored and manipulate to the benefit of the new storage technologies that better the operation and storage usage. Apparently, in a big data view, there are four important directions:

1) Reducing data storage requirements Using data compression and new physical storage structures such as columnar storage.

2) Potentially, big data (using an index that combines several quantitative metrics), will underpin new waves of productivity growth and consumer surplus.

3) Improving input/output (I/O) performance using solid-state drives (SSDs).

4) Increasing storage use by using tiered storage [14]

3.4. Big Data Outcomes

The new emergence of a big data solution shows the expert means to carry out operations with greater degree volumes of data in a limited period with the power to interact with many types of data from different sources. A good example of big data solutions for humanitarian is the timely intervention of a life-threatening condition that it can offer, when take a full advantage of a massive data available. Moreover, big data can make positive risk decisions based on its ability to offer a real-time trace of data. Furthermore, big data has the intensify ability to show threats and criminals different stations or a stream of data, audio, and video feeds, It can also predict weather patterns to design optimal wind turbine use, or multi –channel customer analysis and optimize capital expenditure on asset placement. The big data solution can offer these abilities for sectors like educations, hospitals, and governments for humanitarian purposes. They are:

1) Deep Analytic — a parallel, extensive, and extensible toolbox full of advanced and unique statistical and data mining capabilities
2) High Agility — the ability to create temporary analytic environments in an end-user driven, yet secure and scalable environment to deliver new and unique insights to the working business
3) Massive Scalability — the ability to scale analytics and sandbox to before unknown scales while leverage previously untapped data potential.
4) Low Latency — the ability to act based on these advanced analytic in working, production environments [15].

IV. EXISTING METHODS

(i) **Data Intensive Science**: Many fields and sectors ranging from economic and business activities to public administration from national security to scientific research in many areas involve with Big Data problems. It makes the data easy to find, accessible, and usable.

(ii) **Data Stream Management Systems**: Itreduce the amount of bad data from being collected thus reducing future costs for storage and processing as well as finding potential valuable information and patterns sooner in the data analysis pipeline. It handles very large amounts of data and to efficiently perform searches, comparisons, ordering and summarize etc.,

(iii) **Hadoop image processing interface**: It gives an API for processing images in the distributed computing environment. A set of images put together as one large file with Meta data of the images' layout in HipiImageBundle (HIB).

(iv) **HBase:** A column-oriented database management system runs on top of Hadoop Distributed File System (HDFS). It is well suited for sparse data sets. HBase does not support a structured query language like SQL.

V. BIG DATA PLATFORM

For big data analytics, generally, there are three major advances:

1) Direct analytic over massively parallel processing data warehouses
2) Indirect analytic over Hadoop
3) direct analytic over Hadoop

The proposed approach performs analytic over the Hadoop Map Reduce framework and distributed-clustered systems, such as NoSQL and Hadoop Distributed File System (HDFS). Altogether, the queried file for analytic are performed as Map Reduce problems across big unstructured data placed into NoSQL and Hadoop Distributed File System (HDFS). This approach can cause low-cost big data, result, highly scalable and fault tolerant achieved.

Apache Hadoop is big at storing, combining, and translating multi-structured data into greater useful and valuable arrangements. For instance, Apache Hive is a Hadoop-associated element that corresponds inside the Business Intelligence &Analytic class usually used for querying and analysing data within Hadoop in an SQL-like fashion. Apache Hadoop can also be included with the implied EDW, MPP and New SQL components such as HP Vertica, Teradata, EMC Greenplum, Aster Data, IBM Netezza, SAP Hana, and many others.

Furthermore, Apache HBase is a Hadoop-related NoSQL Key/Value store usually employed for building extremely reactive future-generation applications. Apache Hadoop can also be included with other SQL, NoSQL, and New SQL technologies such as MySQL, IBM DB2, Postgre SQL, Oracle, MongoDB, Terracotta, GemFire, SQLFire, VoltDB, Microsoft SQL Server and many others. In conclusion, data trends and integration applied science help in assuring data flows smoothly between the systems in the above plots.

S. No	Author Name	Method	Platform	Application
1	C.L. Philip Chen, Chun-Yang Zhang	Data Intensive Scientific Discovery (DISD)	Apache Hadoop	Big Data
2	Sanjay P. Ahuja & Bryan Moore	Data Stream Management Systems (DSMS)	Apache Hive, Apache Pig	Big Data in the Cloud
3	Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N Prasad .M. R	HBase	Hadoop Map Reduce	Column Oriented Database in Big Data
4	D. Peter Augustine	Hadoop Image Processing Interface (HIPI)	Horton works Data Platform	Big Data in Healthcare
5	Samson Oluwaseun Fadiyaa, Serdar Saydamb, Vanduhe Vany Zirac	Massively Parallel Processing (MPP)	NoSQL (Not Only SQL) and Hadoop Distributed File System (HDFS)	Big Data for Humanitarian Needs

VI. CONCLUSION

Big Data is an emerging problem for large companies and organizations, as massive volumes of data generated, examined, stored, and analysed. The demanding problems of big data categorized as issues pertaining to data variety, velocity (speed), volume, and veracity. To handle these very demanding challenges, many vendors have grown and modernized a big data platform. However, in this paper, we have suggested a big data platform for large-scale data analysis by using the Hadoop/Map Reduce Framework and NoSQL and HDFS file system over scale out NAS. Simply, Hadoop/Map Reduce is batch-like and not instantaneously desirable for real-time analysis, and not desirable to ad hoc queries. Hadoop figures out the volume and variety issues and so we still need to solve the speed outcome.

Today, the big data issue looms large over many healthcare stakeholders in developed and developing countries. Everyone seems to have realized that the capability to manage and create value from today's large stream of data, from various sources and in many forms (structured/stored, semi-structured/tagged; unstructured/in-motion), represents the new competitive differentiation. Their success will depend on the ability to develop technical capabilities to effectively integrate and analyse information-using new technologies (e.g., Hadoop), develop the right support systems (such as the establishment of big data control towers) and support effective decision making through analytics.

VII. REFERENCES

[1] Hewlett-Packard (2011). Sustainable Business Advantage is using Vertica Analytics. [ONLINE] Available at: http://www8.hp.com/ch/de/pdf/IM_A_Advanced_Services_Vertica_4AA3-8467ENW_tcm_179_1247469.pdf. [Last Accessed 10 February14].

[2] C. Eaton, T. Deutsch, D. Deroos, G. Lapisand P. Zikopoulos, "Understanding Big Data: Analytics for Enterprise Class Headband Streaming
Data", McGraw-Hill, 2011

[3] Philip Russom, (2011). Big Data Analytics. 4th Ed. Renton, WA: TDWI.

[4] Kyar Nyo Aye (2013). Big Data Analytics on Large Scale Shared Storage System. [ONLINE] Available at: https://www.academia.edu/3502343/Big_Data_Analytics_on_Large_Scale_Shared_Storage_System. [Last Accessed 8 February 14].

[5] Ji, C., Li, Y., Qiu, W., Awada, U., & Li, K. (2012). Big Data Processing in Cloud Computing Environments. Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on (17-23).

[6] Pokorny, J. (2011). NoSQL databases: a step to database scalability in web environment. In Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services (iiWAS '11) (pp. 278-283). New York, NY, USA: ACM.

[7] Begoli, E., & Horey, J. (2012). Design Principles for Effective Knowledge Discovery from Big Data. Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on (pp. 215-218).

[8] Gordon Bell, Tony Hey, Alex Szalay, Beyond the data deluge, Science 323 (5919) (2009) 1297–1298.

[9] Tony Hey, Stewart Tansley, Kristin Tolle, The fourth paradigm: data-intensive scientific discovery, Microsoft Research (2009).

[10] Ari, I., Olmezogullari, E., & Celebi, O. F. (2012). Data stream analytics and mining in the cloud. Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on (pp. 857-862).

[11] White, T. Hadoop: The Definitive Guide (2nd Edition) [M]. O'Reilly Media, 2010.

[12] Mukherjee, A. Datta, J. Jorapur, R. Singhvi, R. Haloi, S. Akram, W. 2012. Shared disk big data analytics with Apache Hadoop, High Performance Computing (HiPC), 19th International Conference.

[13] Apache Hadoop Project, http://hadoop.apache.org/, 2013.

[14] Colin White, (July 2011). BI Research. 1st Ed. England: IBM Corporation.

[15] Alex Popescu (e.g., 2011). Achieve the Impossible in Real-Time.
[ONLINE]Available at:http://nosql.mypopescu.com/post/6312810458/bigdata-achieve-the-impossible-in-real-time.
[Last Accessed 2 February 14].