ISSN No. 0976-5697

Volume 8, No. 4, May 2017 (Special Issue)

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Big Data and Hadoop challenges and issues

Er. Shalika Jaiswal
Research Scholar, CSE (Computer Science and Engineering)
SBBSU (Sant Baba Bhag Singh University)
Jalandhar, India
shalikajaswal@gmail.com

Er. Amandeep Singh Walia
AP, CSE (Computer Science and Engineering)
SBBSU (Sant Baba Bhag Singh University)
Jalandhar, India
Er.amanwalia@hotmail.com

***Abstract*—**:Data flow in the millions of computers and millions of process every moment of every day so today is the era of Big Data where data interrelate to the 3Vs volumes, velocity, and varietiey of data interrelate. Huge volume, various varieties and high velocity creates challenges and issues regarding its management and processing. Big Dataenables any organization to the collected, manages, analyzes and making decision of the incredibly from large data sets. Big data growing at an exponential rate but security feature not growing at an same rate. Therefore it becomes very important to develop new technologies to deal with security. This paper introduces the big data of technology along with its importance in the modern world and existing projects like hadoop which are very effective and important in changes the concept of science into big science. Hadoop, Map Reduces and No SQL are the major big data of technology. This paper is also include some other challenges and issues. The various challenges in adapting and accepting Big data are concept that make robust hadoop ecosystem without any processing overhead.

Keywords: Big data, hadoop, Map Reduce, HDFS

1.INTRODUCTION

The term **Big Data** are beingthe increasingly used almost everywhere on the planet – online and offline. And it is not related to computers only. It comes under a blanket term called Information Technology, which is now part of almost all other technology and fields of studies and businesses. Big Data is not a big deal. The type surrounding it is sure pretty big deal to confuse you. This article takes a look at what is a Big Data.

Big data is becoming one of the most talked about technology trends now days. The real challenge with the big organization is to get maximum out of the data already available and predict what kind of the data to collect in the future.[1]

Big data basically huge volume of a data that can't be store and process. In past time data store in mb or gb, now today's data store in terabyte and giga bytes according to the user requirements. Now understand of big data using world important example use in

Social networking sides like-face book, twitter, you tube, linked in google+ etc. each of this sides id huge volume of data in early base.

- Processing the data to analyse and decide to take appropriate action and get appropriate result.
- GB-RDBMS
- TB-Terabyte data
- More than petabyte>big data (Hadoop)

RDBMS means structural form of data i.e. row and columns. It can be represent in a format xml, excel files, sql files, oracle data etc. On the other hand in big data is a represent in structure, semi structure and unstructured form of data.[2]

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters.

The three ways of data can be identified::Structured data: - which is representing in a tabular format.e.g. : Database

Unstructured data: - which does not have pre-defined data model.e.g: Text filesIn Big Data 3 V's

Structured	**unstructured**	**Semi structured**
• **Data bases** • **Data Warehouses** • **Enterprise System** • **(CRM, ERP) etc**	• **Analog Data** • **GPS tracking info.** • **Audio/ Video Stream**	• **XML** • **E-mail**

Some others add few more Vs to the concept:

a. Veracity (Reliability)
b. Variability and
c. Volume of data
d. Variety of data
e. Velocity of data

a. Volume of data: Volume refers to amount of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabyte. Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden [3].

b.Variety of data: Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio, video, XML etc.

It is define as range of data type (include text ,image, audio, video) and source (web, sales, social media, mobile data so on) . Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text

CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)

documents, email, video, audio, stock ticker data and financial transactions. [4]

c. Velocity of data:Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value. Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.[5]

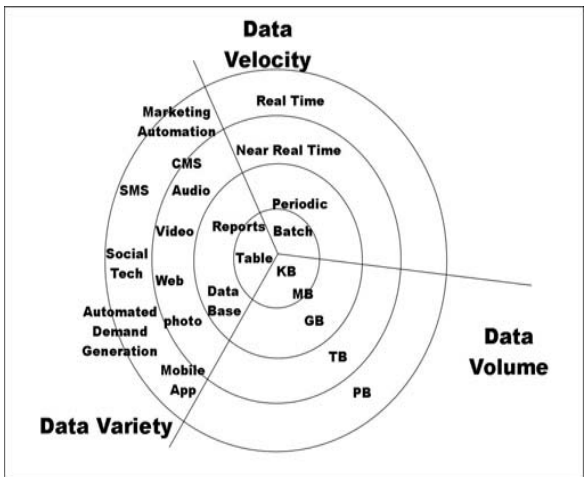

Fig: 1 Data Velocity

d.Veracity:- It refers to the messiness or trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable (just think of Twitter posts with hash tags, reliability and accuracy of content) but big data and analytics technology now allows us to work with these type of data. The volumes often make up for the lack of quality or accuracy. It include structure, semi-structure and unstructured.[6]

e. Variability:- In consistency of data set can process to handle and manage it.

2. Actual uses of big data

The definition of big data isn't really important and one can get hung up on it. Much better to look at 'new' uses of data. So, here are some examples of new and possibly 'big' data use both online and off-.

Netflix

This article from the Wall Street Journal details Netflix's well known Hadoop data processing platform. Cloud architecture is highly scalable and allows Netflix to quickly provision computing resources as its sees the need. Traffic patterns are analyze across device types and localities to help improve the reliability of video streaming and plan for growth.

a. Out of home advertising

I've previously covered Route, who have combined lots of data on footfall and traffic, including the tracked day-to-day movements of 28,000 people. It's hoped that the accuracy of predicting eyes on billboards will increase, leading to fairer pricing.

b. Weather

WeatherSignal works by repurposing the sensors in Android devices to map atmospheric readings. Handsets such as the Samsun S4, contain a barometer, hygrometer (humidity), ambient thermometer and light meter.[6]

c. Understanding and Targeting Customers

This is one of the biggest and most publicized areas of big data use today. Here, big data is used to better understand customers and their behaviors and preferences. Companies are keen to expand their traditional data sets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers.

d. Understanding and Optimizing Business Processes

Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts.

e. Personal Quantification and Performance Optimization

Big data is not just for companies and governments but also for all of us individually. We can now benefit from the data generated from wearable devices such as smart watches or smart bracelets. Take the Up band from Jawbone as an example: the armband collects data

f. Improving Healthcare and Public Health

The computing power of big data analytics enables us to decode entire DNA strings in minutes and will allow us to find new cures and better understand and predict disease patterns. Just think of what happens when all the individual data from smart watches and wearable devices can be used to apply it to millions of people and their various diseases. The clinical trials of the future won't be limited by small sample sizes but could potentially include everyone! Big data techniques are already being used to monitor babies in a specialist premature and sick baby unit

g. Optimizing Machine and Device Performance

Big data analytics help machines and devices become smarter and more autonomous. For example, big data tools are used to operate Google's self-driving car. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings. Big data tools are also used to optimize energy grids using data from smart meters. We can even use big data tools to optimize the performance of computers and data warehouses. [7]

3. Hadoop

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's Map Reduce that is a software framework where an application breaks down into various parts. The Current Apache Hadoop ecosystem consists of the Hadoop Kernel, Map Reduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. HDFS and Map Reduce [8]

CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)

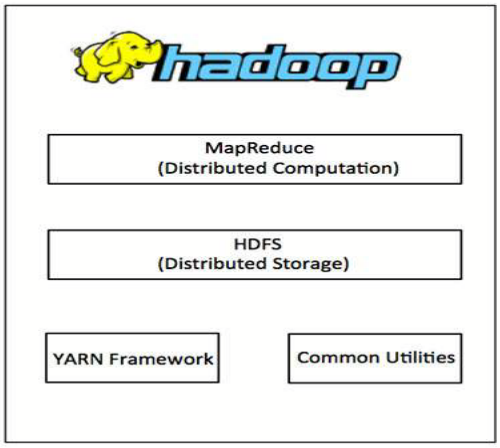

Fig 2: Handoop

- It is a open source framework developed by 2006.
- It is a managed by the apache software foundation
- After the name of hadoop is yellow stuff toy elephant which cutting same hat
- Hadoop is design to slave and process huge volume of data efficiently.
- Hadoop is a Programming framework used to support the
- processing of large data sets in a distributed computing
- Environment. Hadoop was developed by Google's Map Reduce
- that is a software framework where an application break down
- into various parts. The Current Apache Hadoop ecosystem
- consists of the Hadoop Kernel, Map Reduce, HDFS and
- numbers of various components like Apache Hive, Base and
- Zookeeper. HDFS and

Hadoop architecture

Hadoop framework comprise of two main components:

1) HDFS 2) Map Reduce

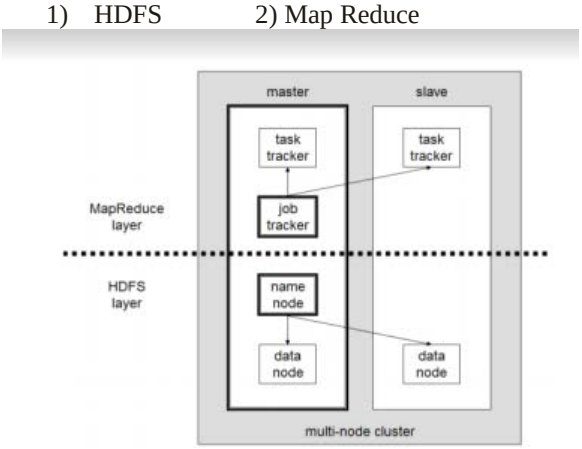

Fig: 3 Handoop Architecture

a. HDFS Architecture

Hadoop includes a fault-tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete of each file by copying each piece to three different servers.[9]

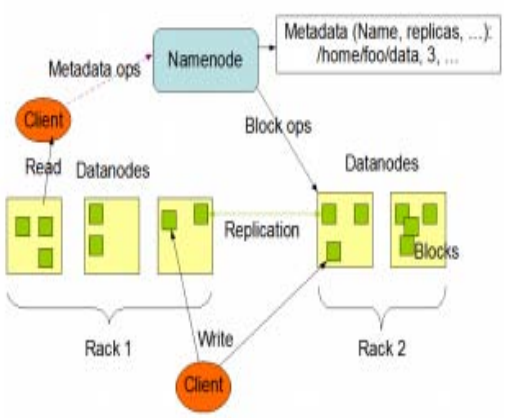

Fig: 4 HDFS Architecture

Map Reduce Architecture

The processing pillar in the Hadoop ecosystem is theMapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel. From an analyst's point of view, this can occur on multiple dimensions. For example, a very large dataset can be reduced into a smaller subset where analytics can be applied. In a traditional data warehousing scenario.Warehouse. There are two functions in Map Reduce as follows:

Map – the function takes key/value pairs as input and generates an intermediate set of key/value pairs.

Reduce – the function which merges all the intermediate values associated with the same intermediate key.[10]

Conclusion

We have now entered in an era of Big Data. The paper describes the concepts of Big Data along with 3 Vs, Volumes, Velocitiesand varieties of Big Data. The paper also focuses on various problems of Big Data processing. These technical challenges must be clearly addressed for efficient and fast processing ofthe Big Data. These technical challenges are common acrossa large variety of application domains, and therefore not cost effective to address in the context of one domain alone. Thepaper describingthe Hadoop which is an open sourceof the soft ware'sused forthe processing of the Big Data.Where data is the collected from the different sources and security is a measures of the issues, as there no any fixed sources of data and not any kind of securities mechanism. Hadoop adopted by the various industries to the process such demand strong,strong security solution. Thus authentication, authorization and encryption or decryption methods are much helpful to securethe Hadoop file system.

REFERENCES

[1] S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka,Japan, Apr 19{23 2013).

[2] Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data : survey" IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}

[3] Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013).

[4] Umasri.M.L, Shyamalagowri.D ,Suresh Kumar.S "Mining Big Data:- Current status and forecast to the future" Volume 4, Issue 1, January 2014 ISSN: 2277 128X

[5] Albert Bifet "Mining Big Data In Real Time" Informatica 37 (2013) 15–20 DEC 2012

[6] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013.

[7] Sameer Agarwal†, Barzan MozafariX, Aurojit Panda†, Henry Milner†, Samuel MaddenX, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data" Copyright © 2013ì ACM 978-1-4503-1994 2/13/04

[8] Yingyi Bu _ Bill Howe _ Magdalena Balazinska _ Michael D. Ernst "The HaLoop Approach to Large-Scale Iterative Data Analysis" VLDB 2010 paper "HaLoop: Efficient Iterative Data Processing on Large Clusters.

[9] Shadi Ibrahim⋆ _ Hai Jin _ Lu Lu "Handling Partitioning Skew in MapReduce using LEEN" ACM 51 (2008) 107–113

[10] Kenn Slagter · Ching-Hsien Hsu "An improved partitioning mechanism for optimizing massive data analysis using MapReduce" Published online: 11 April 2013

CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)