IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-10 E-ISSN: 2347-2693

Big Data: Challenges and Solutions

M.A. Srinuvasu[1*], A. Koushik[2], E.B. Santhosh[3]

[1*]Dept. of Computer Science and Engineering, Raghu Institute of Technology, JNTUK, Visakhapatnam, INDIA
[2] Dept. of Computer Science and Engineering, Raghu Institute of Technology, JNTUK, Visakhapatnam, INDIA
[3] Dept. of Computer Science and Engineering, Raghu Institute of Technology, JNTUK, Visakhapatnam, INDIA

[*]*Corresponding Author: srinuvasu.mutti@gmailmail.com,*

Available online at: www.ijcseonline.org

Received: 22/Sep/2017, Revised: 05/Oct/2017, Accepted: 17/Oct/2017, Published: 30/Oct/2017

***Abstract*—** Big data is huge amount of data which is beyond the processing capacity of conventional data base systems to manage and analyze the data in a specific time interval. The data is too big to store and processed by a single machine. New innovative methods are necessary to process and store large volumes of data. This paper endows with overview of big data, its size, nature, 12Vs of Big data and some technologies to handle it.

***Keywords*—**Big Data, Hadoop, Map Reduce, YARN.

I. INTRODUCTION

Big data became a hot topic today because of massive amount of data. Conventional database management tools find it difficult to manage and analyze in specific time interval [1][2]. The conventional method of psychoanalysis needs two things: 1. Structured form of data, which can fit easily in the relational database. 2. The analysis to be done by a single machine [4].

Based on changes of trend in technology, the nature of data has been processed. Today, most common forms of data are images, text, audio and videos which are beyond the processing capability of conventional data to manage and analyze in precise time period. Big data comes from many sources; some of them are social media, cell phone signals, records of ecommerce etc [6]. For illustration, 10TB (Tera Bytes) data of image file where processing needs to be done such as resizing and its enhancement within the given time interval. Conventional methods will not be able to finish this task within the given time interval because the computing resources would not be enough to complete this task [2][5].

Big data is supposed to handle structured, semi-structured and unstructured data. Relational database can easily hold structured data and can be developed using Structured Query Language (SQL) queries. Semi-structured and Unstructured data do not fit in the relational database; explanation is given in further sections [1][7]. The main challenge of big data is storing and processing of data is done at a specified time span. Hadoop technology is developed to analyze and store Big data.

Big data which is gigantic to any corporation, distributed on several machines. Its size makes it natural in the involvement of many complex data structures to analyze the big data. In Figure 1 shows, few of the companies are estimated to manage huge and complex data on number of servers [4].

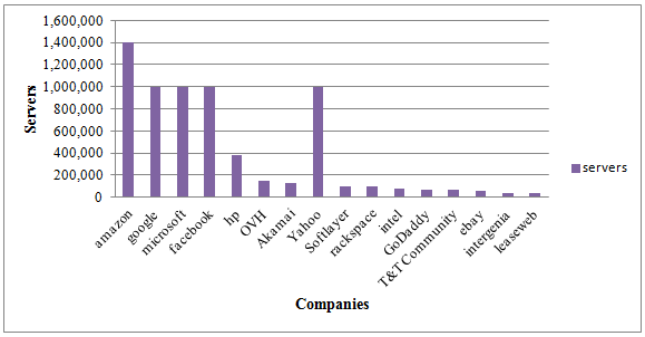

Figure1. Estimated Number of Servers by Various Companies

The next section of the paper deals with generation of data size, classifications of big data, significance of big data. The next further section deals with the challenges of big data and its solutions.

The contributions of this research paper are extracting useful knowledge from 12Vs of Big data and also explained the typical processing of big data using Map-reduce YARN framework.

II. GENERATION OF DATA SIZE

2.5 quintillion bytes of data are created every day, produced by everything from sensors, posts to social media websites, digital pictures and videos, purchase transaction records, and cell phone GPS signals etc [5][7]. If the data generation is increasing day by day then it will become more complex to

store and process these datasets via traditional approaches. The various data sizes are shown in below table 1. In future it is estimated that 40 Zettabytes of data will be produced by 2020 [8].

Table 1: Various data sizes

Notations of Data	Size	Comes Under
Bit	1/8 Byte	Data
Nibble	½ Byte	Data
Byte	1 Byte	Data
Megabyte	1,024 Kilobytes	Data
Gigabyte	1,024 Megabytes	Big Data
Terabyte	1,024 Gigabytes	Big Data
Petabyte	1,024 Terrabytes	Big Data
Exabytes	1,024petabytes	Big Data
Zettabyte	1,024 Exabytes	Bigger Than Big Data
Yottabyte	1,024 Zettabytes	Bigger Than Big Data
Googolbyte	10+1000's Bytes	Bigger Than Big Data

III. BIG DATA CLASSIFICATIONS

Big data can be classified into three categories. They are

Structured data: Structured data refers to any kind of data exist in relational databases and spreadsheets that resides in a fixed field within a record or file [4][6].

Unstructured data: The phrase unstructured data usually refers to information that doesn't reside in a traditional row-column database. As you might expect, it's the opposite of structured data - the data stored in fields in a database [3][4].

Semi structured data: Semi-structured data is data that has not been organized into a specialized repository, such as a database, but nevertheless has associated information such as metadata, that makes it more amenable for processing than raw data [1][2].

Figure 2 gives information about various types of data formats.

A. Merits and Demerits of Data

Merits:

1. Programmers persist objects from their application to a database do not need to worry about object-relational impedance mismatch, but often serialize objects via a light-weight library.
2. Support for nested or hierarchical data often simplifies data models representing complex relationships between entities.
3. Support for lists of objects simplifies data models by avoiding translations of lists into a relational data model[8][9]

Demerits:

1. The traditional relational data model has a popular and ready-made query language, SQL.
2. Prone to "garbage in, garbage out"; by removing restraints from the data model, where there is less fore-thought that is necessary to operate a data application [5].

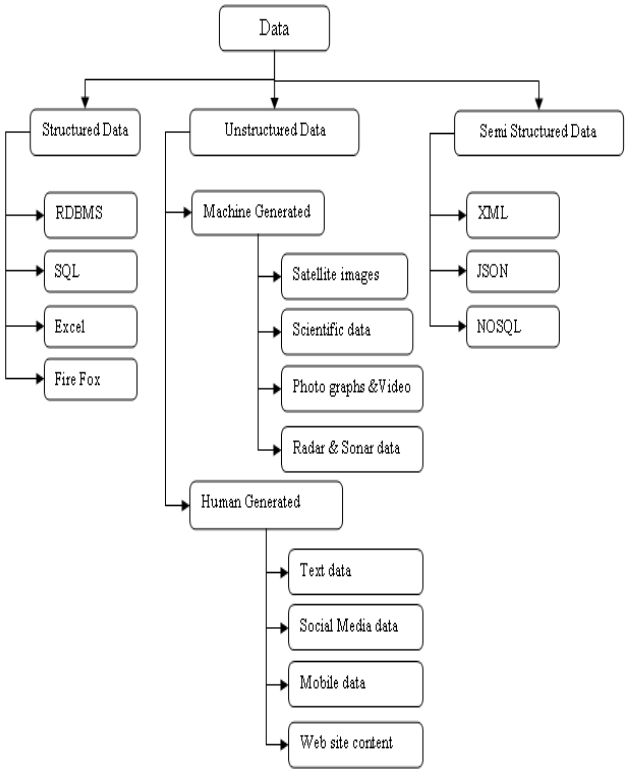

Figure 2: Various types of data formats

B. Difference between structured, unstructured and semi structured data:

Unstructured data has not been organized into a format that makes it easier to access and process. In reality, very little data is completely unstructured. Even things that are often considered unstructured such as documents and images, are structured to some extent [1][3]. Structured data is basically the opposite of unstructured: It has been reformatted and its elements are organized into a data structure so that elements can be addressed, organized and accessed in various combinations to make better use of the information. Semi-structured data lies somewhere between the two. It is not organized in a complex manner that makes sophisticated access and analysis possible; however, it may have information associated with it, such as metadata tagging that allows elements contained to be addressed [6].

IV. SIGNIFICANCE OF BIG DATA

The final objective of big data is to recommend some business solutions that can help corporation to gain insight into business. This rationale itself formulates the analysis of importance of big data. For illustration any corporation can benefit if they could figure if consumer purchases "P" then it is likely that he/she would be interested in purchasing "Q" also. This type of analysis at execution-time can significantly get benefit by growing business. Social networking sites analyze web logs to suggest users what he/she may be fascinated in. Big data endeavours at spectacular cost reduction and generous developments in the time required to achieve a computing task [2].

V. BIG DATA CHALLENGES

In 2001, the 3Vs is a term used to define the characteristics of big data –volume, variety and velocity. Most of the literature review illustrates that there are additional Vs that Information technology, business and data scientists need to concerned with, most particularly big data veracity. Other big data V's getting attention at the high point are: Volume, Velocity, Variety, Variability, Veracity, Validity, Viscosity, Visualization, Value, Virility, Visibility, and Volatility. Figure 3 shows various characteristics of Big data [3].

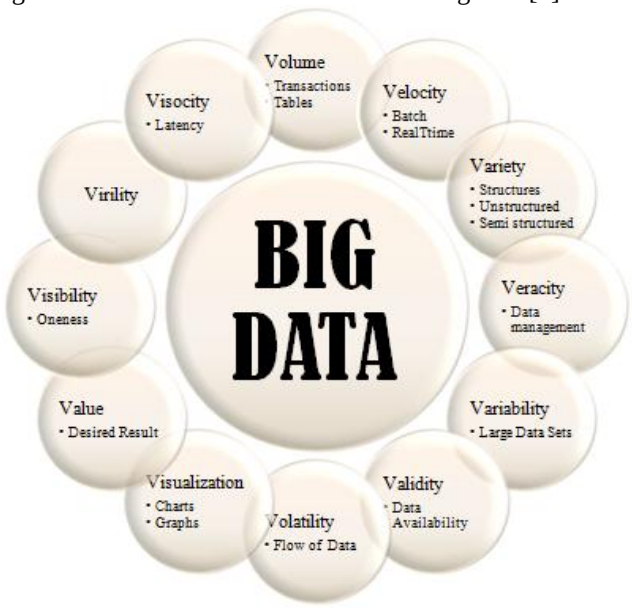

Figure3. Various Characteristics of Big Data

1. *Volume:* volume is the amount of data we have – deliberated in gigabytes is currently measured in Zettabytes or even Yottabytes. The internet of things is generating exponential development in data.
2. *Velocity:* Velocity is the speed in which data is accessible. I memorize the times of daily clusters, now in the event that it's not ongoing it's normally not sufficiently quick.
3. *Variety:* Variety describes one of the biggest challenges of big data. It can be unstructured and it can incorporate such a large number of data from XML to video to SMS. Organizing the data genuinely is no straightforward undertaking, particularly when the data itself changes rapidly.
4. *Variability:* Variability is not similar the same as variety. A coffee shop may offer 6 distinct intermingle of coffee, but if you get the same blend every day and it tastes different every day, that is variability. The same is true of data; if the meaning is constantly changing it can have a huge impact on data homogenization.
5. *Veracity:* Veracity is all about making sure the data is accurate, which requires processes to keep the bad data from accumulating in your systems. The simplest example is contacts that enter your marketing automation system with false names and inaccurate contact information. How many times have you seen Mickey Mouse in your database? It's the classic "garbage in, garbage out" challenge.
6. *Validity:* The interpreted data having a sound basis in logic or fact – is a result of the logical inferences from matching data. One of the most common errors being the confusion between correlation and causation. Volume -Validity = Worthlessness?
7. *Viscosity:* This term is sometimes used to describe the latency or lag time in the data relative to the event being described. We found that this is just as easily understood as an element of Velocity.
8. *Visualization:* Visualization is critical in today's world. Using charts and graphs to visualize huge amounts of complicated data is much more effectual in conveying meaning than spreadsheets and reports chock-full of numbers and formulas.
9. *Value:* Value is the end game. After addressing volume, velocity, variety, variability, veracity, and visualization – which take a lot of time, effort and resources – you want to be sure your organization is getting value from the data.
10. *Virility:* Defined by some users as the rate at which the data spreads; how often it is picked up and repeated by other users or events.
11. *Visibility:* The state of being able to see or be seen – is implied. Data from incongruent sources need to be sewing together where they are visible to the

technology stack making up Big Data. Critical data that is otherwise available, but not visible to the processes of Big Data may be one of the Achilles Heels of the Big Data paradigm. Conversely, unauthorized visibility is a risk. Big Data – visibility = Black Hole?

12. *Volatility:* Big data volatility refers to how long is data valid and how long should it be stored. In this world of real time data you need to determine at what point is data no longer relevant to the current analysis.

VI. BIG DATA SOURCES

Big data sources are available in areas like astronomy, atmospheric science, social networking web sites, life sciences, medical science, government data, natural disaster, resource management, web logs, mobile phones, sensor networks, scientific research, and telecommunications [1][7].

VII. BIG DATA SOLUTIONS

With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for fast and more efficient ways of analyzing such data. Having piles of data on hand is no longer enough to make efficient decisions at the right time [2][6].

Such data sets can no longer be easily analyzed with traditional data management techniques. Therefore, there arises a need for new tools and methods specialized for big data analytics, as well as required architectures for storing and managing such data. Accordingly, the emergence of big data has an effect on everything from the data itself and its collection, to the processing, to the final extracted decisions [7][8].

A. Big Data Storage and Management

YARN (Yet Another Resource Negotiator) is a framework responsible for providing computational resources CPU, RAM etc) needed for execution of application. The important elements present in YARN are [7]

a. *Resource Manager*- It is the master which knows where the slaves are located and also it keeps track of how many resources are utilized by the nodes. It executes numerous services and uses Resource Scheduler module which decides how to assign the resources

b. *Node Manager*- It is the slave node which periodically sends the heartbeat signal to the Resource Manager and also offers several resources to the cluster. The Resource scheduler mainly decides how to use the resource capacity i.e., determined by the amount of memory and the number of vcores at run time.

1. Benefits of YARN

- New Applications and services
- Improved utilization of clusters
- Scalability
- Shared services

2. Explanation of Processing of Map-Reduce job using YARN

The different steps involved in processing of map-reduce job using YARN framework is described below [7][8]

a. A client program submits the application which includes the necessary specifications for launching the application-specific ApplicationMaster itself.

b. The ResourceManager takes the responsibility for negotiating a specified container to start the ApplicationMaster

c. The ApplicationMaster, on boot-up, registers with the ResourceManager –which allows the client program to query the ResourceManager for details allowing communicating with its own ApplicationMaster.

d. During normal operation the ApplicationMaster negotiates resource containers via the resource-request protocol.

e. On successful allocation of the container, the ApplicationMaster launches it by providing the container launch specification to the NodeManager.

f. The application code which executes within the container provides necessary information (progress, status etc.) to its ApplicationMaster via an application-specific protocol.

g. During the application execution, the client communicates directly with the ApplicationMaster to get status, progress updates etc. via an application-specific protocol.

h. Once the application is complete, and all work has been finished, the ApplicationMaster deregisters with the ResourceManager and gets shut down.

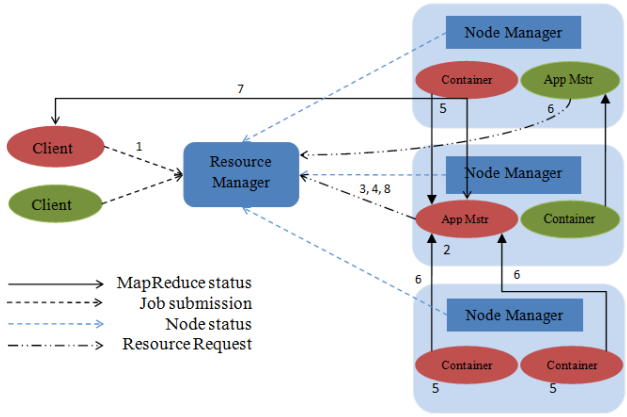

Figure 4. Entire data flow of yarn

B. Big Data Processing

After the big data storage, next stage is the analytical processing. Consequently, there are four significant necessities for big data processing [8] [9].

a. Data loading must be fast, because the disk and network traffic interferes with the query executions during data loading, it is necessary to reduce the data loading time.

b. Query Processing must be fast: in order to satisfy the requirements of heavy workloads and real-time requests since many queries are response-time critical. Thus, the data placement structure must be capable of retaining high query processing speeds as the amounts of queries rapidly increases.

c. Storage space utilization is highly well-organized: Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space necessitates that data storage be well managed during processing, and issues on how to store the data so that space utilization is maximized.

d. Dynamic workload patterns are strongly adaptive: As big data sets are analyzed by different applications and users, for different purposes, and in various ways, the underlying system should be highly adaptive to unexpected dynamics in data processing, and not specific to certain workload patterns

Map Reduce is a parallel programming model, inspired by the "Map" and "Reduce" of functional languages, which is suitable for big data processing. It is the core of Hadoop, and performs the data processing and analytics functions. The fundamental idea of MapReduce is breaking a task down into stages and executing the stages in parallel in order to reduce the time needed to complete the task [3][4].

The first phase of the MapReduce job is to map input values to a set of key/value pairs as output. The "Map" function accordingly partitions large computational tasks into smaller tasks, and assigns them to the appropriate key/value pairs. Thus, unstructured data, such as text, can be mapped to a structured key/value pair, for example, the key could be the word in the text and the value is the number of occurrences of the word. This output is then the input to the "Reduce" function. Reduce then performs the collection and combination of this output, by combining all values which share the same key value, to provide the final result of the computational task [3][10].

The MapReduce function within Hadoop depends on two different nodes: the Job Tracker and the Task Tracker nodes. The Job Tracker nodes are the ones which are responsible for distributing the mapper and reducer functions to the available Task Trackers, as well as monitoring the results. The MapReduce job starts by the Job- Tracker assigning a portion of an input file on the HDFS to a map task, running on a node. On the other hand, the Task Tracker nodes actually run the jobs and communicate results back to the Job Tracker. Communication between nodes is often through files and directories in HDFS, so inter-node communication is minimized [6][10].

Hadoop makes it popular for big data by loading data as files into the distributed file system, and running parallel MapReduce computations on the data. Data is loaded into Hadoop simply by copying files into the distributed file system, and MapReduce interprets the data at processing time rather than loading time [7][9]. Thus, it is capable of attracting all data sources, as well as adapting its engines to any evolutions that may occur in such big data sources. The entire process is summarized in the figure 5

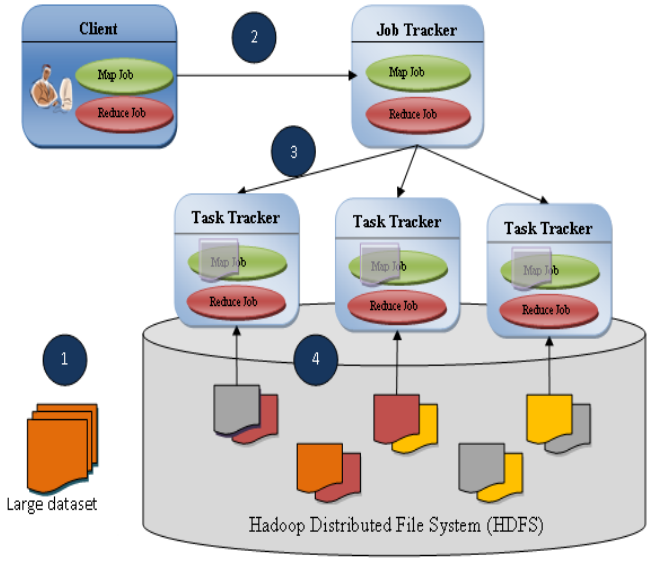

Figure 5. Hadoop Distributed File System

VIII. CONCLUSION

Big data is huge amount of data which is beyond the processing capability of conventional data base systems to manage and analyze the data in a specific time interval. New innovative methods are necessary to process and store large volumes of data.

This paper explores generation of data size, importance of big data and its challenges, solutions regarding the processing of big data in real world such as explaining its processing using Hadoop Map-reduce V2 YARN framework.

REFERENCES

[1] R. Gupta,S. Gupta, A singhal," *Big Data : Overview*", International Journal of Computer Trends and Technology, Vol 9, Issue 5, pp. 266-268, 2014.

[2] S. Sagiroglu, D. Sinanc, "*Big data: A review*", In the Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, pp. 42-47, 2013.

[3] S. Sathyamoorthy, "*Data Mining and Information Security in Big Data*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.3, pp.86-91, 2017.

[4] Cuzzocrea, I. Song, K.C. Davis "*Analytics over Large-Scale Multidimensional Data: The Big Data Revolution*", In the Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101-104, 2011.

[5] Palaghat Yaswanth Sai, Pabolu Harika, "*Illustration of IOT with Big Data Analytics*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.9, pp.221-223, 2017.

[6] Raju Din, Prabadevi B. , "*Data Analyzing using Big Data (Hadoop) in Billing System*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.5, pp.84-88, 2017.

[7] V.K. Vavilapalli, A.C. Murthy, Ch. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, Owen O'Malley, S.Radia, B. Reed, and E. Baldeschwieler, "*Apache Hadoop YARN: yet another resource negotiato*r", In Proceedings of the 4th annual Symposium on Cloud Computing (SOCC '13). ACM, New York, NY, USA, , Article 5 , pp.16, 2013.

[8] P. Sharma, V. Garg, R, Kaur, S, Sonare, " *Big Data in Cloude Environment*" International Journal of Computer Sciences and Engineering, Vol 1, Issue 3, pp.15-17, 2013.

[9] Prakash Singh , "*Efficient Deep Learning for Big Data: A Review*", International Journal of Scientific Research in Computer Science and Engineering, Vol.4, Issue.6, pp.36-41, 2016.

[10] M. Bhanu Sridhar, A. Koushik, "*A Study of Big Data Analytics in Clouds with a Security Perspective*", International Journal of Engineering Research and Technology, Vol 6, Issue 1, pp.5-9, January 2017.

Authors Profile

Mr. Appala Srinuvasu Muttipati pursued MCA from Andhra University, Visakhapatnam in 2008 and M.Tech in Computer Science and Technology from GITAM University, Visakhapatnam. Further pursuing Ph.D and currently working as Assistant Professor in Department of Computer Science and Engineering at Raghu Institute of Technology. He has published 6 research papers from reputed international journals which are approved by UGC. His main research work focuses on Data Mining, Big data, Cloud computing and IoT. He has 9 years of teaching experience.

Mr.Koushik Akkinapalli pursued B,Tech from JNTUK, Kakinada in 2009 and M.Tech in Computer Science and Engineering from JNTUK,Kakinada. Further pursuing Ph.D and currently working as Assistant Professor in Department of Computer Science and Engineering at Raghu Institute of Technology. He has published 4 research papers from reputed international journals. His main research work focuses on Big data, Cloud computing and Network Security. He has 5 years of teaching experience.

Mr.Eagala Bhaskara Santhosh pursued B,Tech from JNTUK, Kakinada in 2011 and M.Tech in Cyber Security from JNTUK,Kakinada and currently working as Assistant Professor in Department of Computer Science and Engineering at Raghu Institute of Technology. His main research work focuses on IOT Security, Cloud computing and Network Security. He has 2 years of teaching experience.

