International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 4 Issue: 3 09-Apr-2015,ISSN_NO: 2321-3337

A Review on Big Data with its Issues, Challenges & Tools

Deshmukh S. P.[1], Mali M. V.[2]

Asst. Professor, Computer Science and Engg. Dept.,VVPIET, Solapur, India[1]
Student, Computer Science and Engg. Dept, VVPIET Solapur, India [2].

.

ABSTRACT- *Big data is nothing but large amount of data which requires new technologies and architectures so that it becomes possible to extract value from it by capturing and analysis process. Because of large size of data, it becomes very difficult to perform effective analysis by using the existing traditional techniques. Volume, velocity, variability, value and complexity are the some characteristics of big data. As big data is the recent upcoming technology in market, which results in huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology should brought into focus. In this paper, the introduction of big data technology along with its importance is provided, moreover its application to the modern world along with the existing projects are effectively enlisted, also this paper reports various challenges and issues in adapting big data technology ,along with its tool Hadoop.*

Keywords— Big Data, Hadoop, Hadoop Distributed File System, MapReduce.

1. INTRODUCTION

To describe the exponential growth of both structured and unstructured data , Big Data term is used. Big data may be as important to business – and society – as the Internet has become. The concept of big data has been introduced within computer science since the earliest days of computing. “Big Data” is nothing but the volume of data that could not be processed efficiently by traditional database methods and tools. Whenever a new storage medium was invented for such a data, the amount of data exploded further. The original definition was focused on structured data, but afterwards researchers came to know that the most of the world’s information exist in massive, unstructured information, large amount of data is in text or image form. The explosion of data has not been accompanied by a corresponding new storage medium.

“Big Data” can be defined as the amount of data which is just beyond technology’s capability to store, manage and process efficiently. Today, we are thinking in tens to hundreds of terabytes. Thus, big data is a moving target. The current growth rate in the amount of data collected is high. A major challenge for IT researchers and practitioners is that this growth rate is fast exceeding our ability to both:

(1) Design appropriate systems to handle the data effectively

(2) Analyzing data to extract relevant meaning for decision making.

2. LITERATURE SURVEY

Following are the terminologies needed to be discussed referring to big data.

2.1 Volume

There are many factors which causes to increase volume of data. Transaction based data is to be stored for number of years . Unstructured data is created from social media like facebook, twitter etc.. Large amounts of data is collected from sensors for weather forecasting etc. Before some years, excessive data volume was a major storage issue. But due to decreasing storage costs,it is not the problem now. How to use analytics to create value from relevant data and how to determine relevance within large data volumes is the major problem now.

2.2 Velocity

Velocity is nothing but the speed at which the new data is generated or the data moves around.

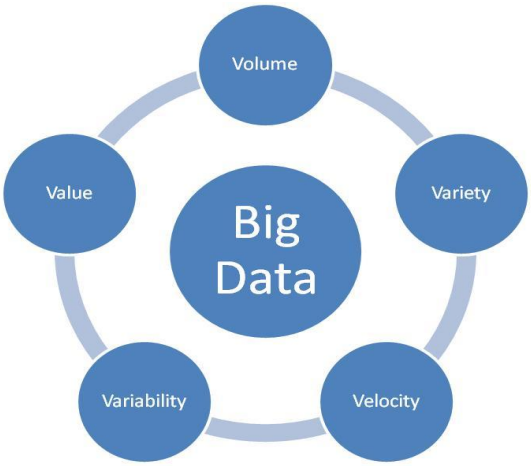

Fig. 1 Characteristics of Big data

2.3 Variety

Structured Data

It is the data which resides in a fixed field within a record or file. It includes data contained in relational databases and spreadsheets.

Unstructured Data

It is the data or information which do not reside in a traditional row-column database. It is the opposite of structured data. Unstructured data files also include text and multimedia content. Word processing documents, e-mail messages, , videos, images,maps,

audio files, power point presentations, webpages and many other kinds of business documents are the examples of unstructured data.

2.4 Value

Value creation from Big Data could become the major use. Value is nothing but the usefulness of data in making decisions. The purpose of computing is insight, not the numbers .

2.5 Complexity

Complexity is the degree of interconnectedness and interdependence in big data structures such that a small change in one or a few elements can result in very large changes.

3. MOTIVATION

Following factors motivates the evolution of a big data.

3.1 Business

Companies are able to gain a more complete understanding of their business, customers, products, competitors, etc. ,when big data is effectively and efficiently captured, processed, and analyzed, It leads to improve eficiency, to increase sales, to lower the costs, for better customer service, and to improve products and services.

3.2 Medical Field

An example from the medical field illustrates how and why big data and new analytics may be truly beneficial. Current data in a patient's medical record and current health situation is used to plan and target patient participation in wellness and disease management programs.

3.3 Log Storage in IT Industries

IT industries store large amount of data as Logs to deal with the problems. But the storage of this data is required for years because of their value. Because of their volume, raw and semi structured nature ,the Traditional Systems are not able to handle these logs. Moreover these logs go on changing with the S/w and H/w updates occurring.

3.4 Sensor Data

Massive amount of sensor data is also a big challenge for big data. Moreover sensor data is characterized by both data in motion and data at rest.

4. CHALLENGES AND ISSUES OF BIG DATA

The various issues and challenges regarding big data are enumerated as follows.

4.1 Privacy and Security

Big data is sensitive. It includes conceptual, technical as well as legal significance. When the personal information is combined with external large data sets leads to the

inference of new facts about that person and it is possible that these kinds of facts about the person are secretive and the person might not want the Data Owner to know or any person to know about them. To add value to the business of the organization ,information regarding the people is collected and used.

4.2 Data Access and Sharing of Information

If available data is to be used to make accurate decisions in time ,it becomes necessary that the data should be available in complete, accurate and timely manner. Expecting sharing of data between companies is awkward because of the need to get an edge in business. Sharing personal data about the clients and operations threatens the secrecy and competitiveness.

4.3 Storage and Processing Issues

The storage available is not enough for storing the large amount of data which is being produced by almost everything: Social Media sites are themselves a great contributor along with the sensor devices etc. Because of the rigorous demands of the Big data on networks, storage and servers outsourcing the data to cloud may seem an option. Uploading this large amount of data in cloud doesn't solve the problem. Since Big data insights require getting all the data collected and then linking it in a way to extract important information. Terabytes of data will take large amount of time to get uploaded in cloud and moreover this data is changing so rapidly which will make this data hard to be uploaded in real time. At the same time, the cloud's distributed nature is also problematic for Big data analysis. Thus the cloud issues with Big Data can be categorized into Capacity and Performance issues.

4.4 Skill Requirement

Since Big data is an emerging technology , it needs to attract organizations and youth with diverse new skill sets. These skills should not be technical oriented only, but also it should be extensible for research work including analytical, interpretive and creative tasks. These skills are required to be developed in related people.

4.5 Analytical challenges

The main challenging questions are as follows:

- What if the data volume gets big and varied and it is unknown how to deal with it?
- Whether all data need to be stored?
- Whether all data need to be analysed?
- Out of very large amount of total data, which data points are really important from the whole is also important to identify.
- Use of the data to its best advantage?

Big data brings along with it some huge analytical challenges. The type of analysis to be done on such a bulky amount of data which may be unstructured, semi structured or structured needs a large number of advanced skills

4.6 Technical Challenges

Fault Tolerance:

With the incoming of new technologies like Cloud computing and Big data it is always intended that whenever the failure occurs the damage done should be within acceptable threshold rather than beginning the whole task from the scratch. Fault-tolerant computing is extremely hard, involving intricate algorithms. It is simply not possible to devise absolutely foolproof, 100% reliable fault tolerant machines or software. Thus the main task is to reduce the probability of failure to an "acceptable" level. Unfortunately, the more we strive to reduce this probability, the higher the cost. Two methods which seem to increase the fault tolerance in Big data are as: First is to divide the whole computation being done into tasks and assign these tasks to different nodes for computation. One node is assigned the work of observing that these nodes are working properly. If something happens that particular task is restarted.

Scalability:

The processor technology has changed in recent years. The clock speeds have largely stalled and processors are being built with more number of cores instead. Previously data processing systems had to worry about parallelism across nodes in a cluster but now the concern has shifted to parallelism within a single node. In past the techniques which were used to do parallel data processing across data nodes aren't capable of handling intra-node parallelism. This is because of the fact that many more hardware resources such as cache and processor memory channels are shared across a core in a single node. The scalability issue of Big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively.

Quality of Data:

Collection of huge amount of data and its storage comes at a cost. More data if used for decision making or for predictive analysis in business will definitely lead to better results. Business Leaders will always want more and more data storage whereas the IT Leaders will take all technical aspects in mind before storing all the data. Big data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn.

Heterogeneous Data:

Unstructured data represents almost every kind of data being produced like social media interactions, to recorded meetings, to handling of PDF documents, fax transfers, to emails and more. Structured data is always organized into highly mechanized and manageable way. It shows well integration with database but unstructured data is completely raw and unorganized.

5. TOOLS

Hadoop is an open source project hosted by Apache Software Foundation. It consists of many mall sub projects which belong to the category of infrastructure for distributed computing. Hadoop mainly consists of:

- File System (The Hadoop File System)

- Programming Paradigm (Map Reduce)

The other subprojects provide complementary services or they are building on the core to add higher-level abstractions. There exist many problems in dealing with storage of large amount of data. Though the storage capacities of the drives have increased massively but the rate of reading data from them hasn't shown that considerable improvement. The reading process takes large amount of time and the process of writing is also slower. This time can be reduced by reading from multiple disks at once. Only using one hundredth of a disk may seem wasteful. But if there are one hundred datasets, each of which is one terabyte and providing shared access to them is also a solution. There occur many problems also with using many pieces of hardware as it increases the chances of failure. This can be avoided by Replication i.e. creating redundant copies of the same data at different devices so that in case of failure the copy of the data is available.

5.1 Hadoop Distributed File System

Hadoop comes with a distributed File System called HDFS, which stands for Hadoop Distributed File System. HDFS is a File System designed for storing very large files with streaming data access patterns, running on clusters on commodity hardware. HDFS block size is much larger than that of normal file system i.e. 64 MB by default. The reason for this large size of blocks is to reduce the number of disk seeks. A HDFS cluster has two types of nodes i.e. name node (the master) and number of data nodes (workers). The name node manages the file system namespace, maintains the file system tree and the metadata for all the files and directories in the tree.

An important characteristic of Hadoop is the partitioning of data and computation across many (thousands) of hosts, and the execution of application computations in parallel close to their data. A Hadoop cluster scales computation capacity, storage capacity and I/O bandwidth by simply adding commodity servers. Hadoop clusters at Yahoo! span 40,000 servers, and store 40 petabytes of application data, with the largest cluster being 4000 servers. One hundred other organizations worldwide report using Hadoop.

5.2 Hadoop Architecture

Name Node:

The HDFS namespace is a hierarchy of files and directories. Files and directories are represented on the Name Node by inodes. Inodes record attributes like permissions, modification and access times, namespace and disk space quotas. The file content is split into large blocks (typically 128 megabytes, but user selectable file-by-file), and each block of the file is independently replicated at multiple Data Nodes (typically three, but user selectable file-by-file).

Data Nodes*:*

Each block replica on a Data Node is represented by two files in the local native file system. The first file contains the data itself and the second file records the block's metadata including checksums for the data and the generation stamp. The size of the data file equals the actual length of the block and does not require extra space to round it up to the nominal block size as in traditional file systems. Thus, if a block is half full it needs only half of the space of the full block on the local drive.

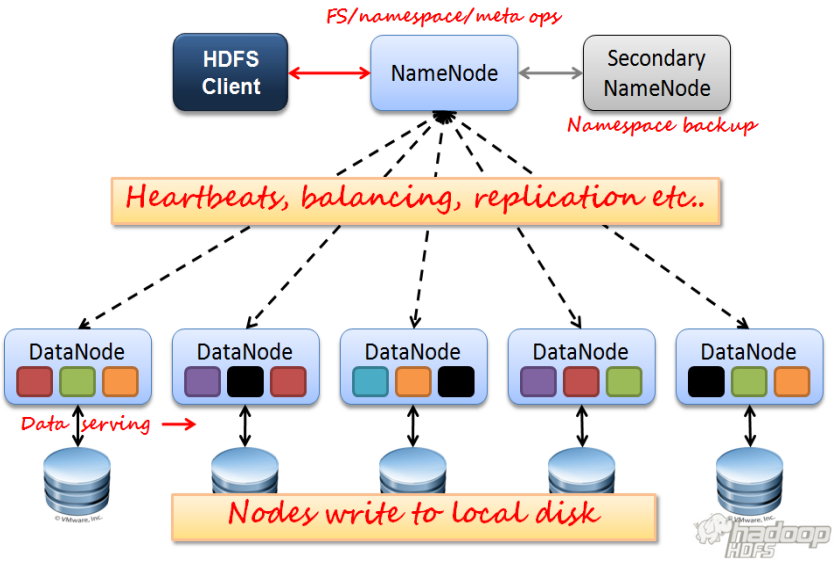

Fig. 2 HDFS Architecture

During startup each Data Node connects to the Name Node and performs a handshake. The purpose of the handshake is to verify the namespace ID and the software version of the Data Node. If either does not match that of the Name Node, the Data Node automatically shuts down.

The namespace ID is assigned to the file system instance when it is formatted. The namespace ID is persistently stored on all nodes of the cluster. Nodes with a different namespace ID will not be able to join the cluster, thus protecting the integrity of the file system. A Data Node that is newly initialized and without any namespace ID is permitted to join the cluster and receive the cluster's namespace ID.

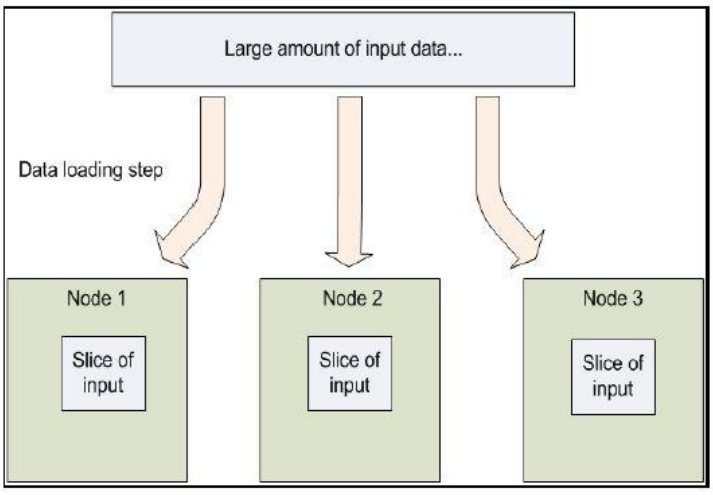

Fig.3 Division of Input Data

5.3 Mapreduce

MapReduce is the programming paradigm allowing massive scalability. The MapReduce basically performs two different tasks i.e. Map Task and Reduce Task. A map-

reduce computation executes as follows: Map tasks are given input from distributed file system. The map tasks produce a sequence of key-value pairs from the input and this is done according to the code written for map function. These value generated are collected by master controller and are sorted by key and divided among reduce tasks. The sorting basically assures that the same key values ends with the same reduce tasks. The Reduce tasks combine all the values associated with a key working with one key at a time. Again the combination process depends on the code written for reduce job. The Master controller process and some number of worker processes at different compute nodes are forked by the user.

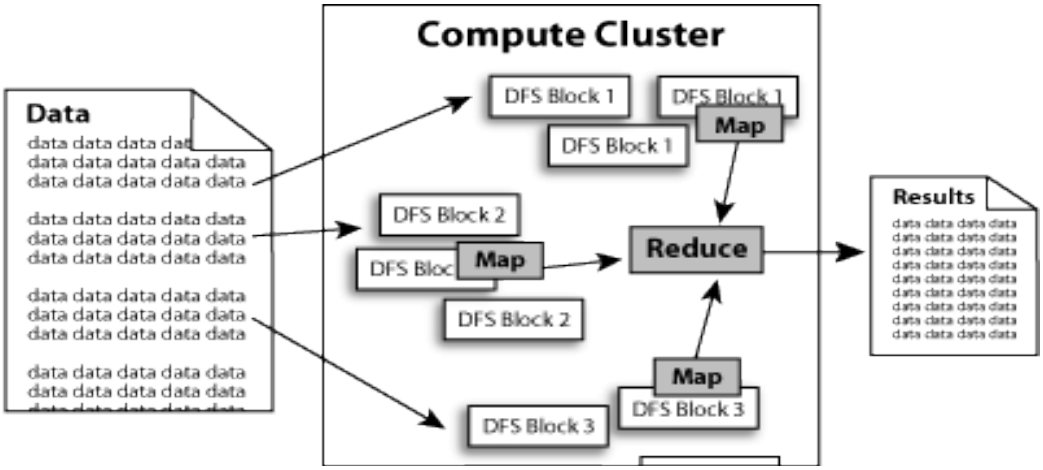

Fig. 4 MapRaduce

The status of each of these Map tasks is set to idle by Master. These get scheduled by Master on a Worker only when one becomes available. The Master must also inform each Reduce task that the location of its input from that Map task has changed.

6. CONCLUSION

In this paper, new concepts of big data are discussed to accept and adapt to this new technology with many challenges and issues which exist in the beginning before it is too late. These challenges and issues will help the business organizations which are moving towards this technology to increase the value of the business to a significant level Hadoop tool for Big data is described in detail focusing on the areas where it needs to be improved so that in future Big data can have technology as well as skills to work with.

REFERENCES

[1] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, “Big Data: Issues and Challenges Moving Forward”, *IEEE, 46th Hawaii International Conference on System Sciences,* 2013.

[2] Sam Madden, “ From Databases to Big Data”, *IEEE, Internet Computing,* May-June 2012.

[3] Kapil Bakshi, “Considerations for Big Data: Architecture and Approach”, *IEEE , Aerospace Conference,* 2012.

[4] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", *IEEE,International Conference on Communication, Information & Computing Technology (ICCICT)*, Oct. 19-20, 2012.

[5] Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono, Cees de Laat, "Addressing Big Data Challenges for Scientific Data Infrastructure", *IEEE , 4th International Conference on Cloud Computing Technology and Science*, 2012.

[6] Martin Courtney, "The Larging-up of Big Data", *IEEE, Engineering & Technology*, September 2012.

[7] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt , "Big Data Privacy Issues in Public Social Media", *IEEE, 6th International Conference on Digital Ecosystems Technologies (DEST)*, 18-20 June 2012.