Big Data: Issues, Challenges, Tools and Good Practices

Avita Katal
Department of CSE
Graphic Era University
Dehradun, India
avita207@gmail.com

Mohammad Wazid
Department of CSE
Graphic Era University
Dehradun, India
wazidkec2005@gmail.com

R H Goudar
Department of CSE
Graphic Era University
Dehradun, India
rhgoudar@gmail.com

***Abstract*— Big data is defined as 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. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. Since Big data is a recent upcoming technology in the market which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology are brought into light. This paper introduces the Big data technology along with its importance in the modern world and existing projects which are effective and important in changing the concept of science into big science and society too. The various challenges and issues in adapting and accepting Big data technology, its tools (Hadoop) are also discussed in detail along with the problems Hadoop is facing. The paper concludes with the Good Big data practices to be followed.**

Keywords— Big data; Hadoop; Hadoop Distributed File System; MapReduce.

I. Introduction

Data is growing at a huge speed making it difficult to handle such large amount of data (exabytes).The main difficulty in handling such large amount of data is because that the volume is increasing rapidly in comparison to the computing resources. The Big data term which is being used now a days is kind of misnomer as it points out only the size of the data not putting too much of attention to its other existing properties.

Big data can be defined with the following properties associated with it:

A. Variety

Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents, sensor devices data both from active passive devices. All this data is totally different consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems.

B. Volume

The Big word in Big data itself defines the volume. At present the data existing is in petabytes and is supposed to increase to zettabytes in nearby future. The social networking sites existing are themselves producing data in order of terabytes everyday and this amount of data is definitely difficult to be handled using the existing traditional systems.

C. Velocity

Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example the data from the sensor devices would be constantly moving to the database store and this amount won't be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.

D. Variability

Variability considers the inconsistencies of the data flow. Data loads become challenging to be maintained especially with the increase in usage of the social media which generally causes peak in data loads with certain events occurring.

E. Complexity

It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control.

F. Value

User can run certain queries against the data stored and thus can deduct important results from the filtered data obtained and can also rank it according to the dimensions they require. These reports help these people to find the business trends according to which they can change their strategies.

As the data stored by different organizations is being used by them for data analytics. It will produce a kind of gap in-between the Business leaders and the IT professionals the main

978-1-4799-0192-0/13/$31.00 ©2013 IEEE

concern of business leaders would be to just adding value to their business and getting more and more profit unlike the IT leaders who would have to concern with the technicalities of the storage and processing. Thus the main challenges that exist for the IT Professionals in handling Big data are:

- The designing of such systems which would be able to handle such large amount of data efficiently and effectively.
- The second challenge is to filter the most important data from all the data collected by the organization. In other words we can say adding value to the business.

In this paper we have presented the main issues and challenges along with the complete description of the technologies/methods being employed for tackling the storage and processing problems associated with Big Data. The paper concludes with the good Big data practices to be followed.

II. RELATED WORK

In paper [1] the issues and challenges in Big data are discussed as the authors begin a collaborative research program into methodologies for Big data analysis and design. In paper [2] the author discusses about the traditional databases and the databases required with Big data concluding that the databases don't solve all aspects of the Big data problem and the machine learning algorithms need to be more robust and easier for unsophisticated users to apply. There is the need to develop a data management ecosystem around these algorithms so that users can manage and evolve their data, enforce consistency properties over it and browse, visualize and understand their algorithm results. In paper [3] architectural considerations for Big data are discussed concluding that despite the different architectures and design decisions, the analytics systems aim for Scale-out, Elasticity and High availability. In paper [4] all the concepts of Big data along with the available market solutions used to handle and explore the unstructured large data are discussed. The observations and the results showed that analytics has become an important part for adding value for the social business. This paper [5] proposes the Scientific Data Infrastructure (SDI) generic architecture model. This model provides a basis for building interoperable data with the help of available modern technologies and the best practices. The authors have shown that the models proposed can be easily implemented with the use of cloud based infrastructure services provisioning model. In paper [6] the author investigates the difference in Big data applications and how they are different from the traditional methods of analytics existing from a long time. In paper [7] authors have done analysis on Flickr, Locr, Facebook and Google+ social media sites. Based on this analysis they have discussed the privacy implications and also geo-tagged social media; an emerging trend in social media sites. The proposed concept in this paper helps users to get informed about the data relevant to them in such large social Big data.

III. IMPORTANCE OF BIG DATA AND VARIOUS PROJECTS

Big data is different from the data being stored in traditional warehouses. The data stored there first needs to be cleansed, documented and even trusted. Moreover it should fit the basic structure of that warehouse to be stored but this is not the case with Big data it not only handles the data being stored in traditional warehouses but also the data not suitable to be stored in those warehouses. Thus there comes the point of access to mountains of data and better business strategies and decisions as analysis of more data is always better.

A. Log Storage in IT Industries

IT industries store large amount of data as Logs to deal with the problems which seem to be occurring rarely in order to solve them. But the storage of this data is done for few weeks or so though these logs need to be stored for longer duration because of their value. The Traditional Systems are not able to handle these logs because of their volume, raw and semi structured nature. Moreover these logs go on changing with the s/w and H/w updates occurring. Big data analytics not only does analysis on the whole /large data available to pinpoint the point of failures but also would increase the longevity of the log storage.

B. Sensor Data

Massive amount of sensor data is also a big challenge for Big data. All the industries at present dealing with this large amount of data make use of small portion of it for analysis because of the lack of the storage infrastructure and the analysis techniques. Moreover sensor data is characterized by both data in motion and data at rest. Thus safety, profit and efficiency all require large amount of data to be analyzed for better business insights.

C. Risk Analysis

It becomes important for financial institutions to model data in order to calculate the risk so that it falls under their acceptable thresholds. A lot amount of data is potentially underutilized and should be integrated within the model to determine the risk patterns more accurately.

D. Social Media

The most use of Big data is for the social media and customer sentiments. Keeping an eye on what the customers are saying about their products helps business organizations to get a kind of customer feedback. This feedback is then used to modify decisions and get more value out of their business.

TABLE I. VARIOUS BIG DATA PROJECTS

Domain	Description
Big Science	1. The Large Hadron Collider (LHC) is the world's largest and highest-energy particle accelerator with the aim of allowing physicists to test the predictions of different

	theories of particle physics and high-energy physics. The data flow in experiments consists of 25 petabytes (as of 2012) before replication and reaches upto 200 petabytes after replication. 2. The Sloan Digital Sky Survey is a multi-filter imaging and spectroscopic redshift survey using a 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. It is Continuing at a rate of about 200 GB per night and has more than 140 terabytes of information.
Government	1. The Obama administration project is a big initiative where a Government is trying to find the uses of the big data which eases their tasks somehow and thus reducing the problems faced. It includes 84 different Big data programs which are a part of 6 different departments. 2. The Community Comprehensive National Cyber Security initiated a data center, Utah Data Center (United States NSA and Director of National Intelligence initiative) which stores data in scale of yottabytes. Its main task is to provide cyber security.
Private Sector	1. Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB. 2. Walmart is estimated to store about more than 2.5 petabytes of data in order to handle about more than 1 million customer transactions every hour. 3.FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.
International Development	Information and Communication Technologies for Development (ICT4D) uses the Information and Communication Technologies (ICTs) for the socioeconomic development, human rights and international development. Big data can make important contributions to international development.

IV. BIG DATA CHALLENGES AND ISSUES

A. Privacy and Security

It is the most important issue with Big data which is sensitive and includes conceptual, technical as well as legal significance.

- The personal information of a person when combined with external large data sets leads to the inference of new facts about that person and it's 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.
- Information regarding the users (people) is collected and used in order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of.
- Another important consequence arising would be Social stratification where a literate person would be taking advantages of the Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse.
- Big Data used by law enforcement will increase the chances of certain tagged people to suffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated.

B. Data Access and Sharing of Information

If data is to be used to make accurate decisions in time it becomes necessary that it should be available in accurate, complete and timely manner. This makes the Data management and governance process bit complex adding the necessity to make Data open and make it available to government agencies in standardized manner with standardized APIs, metadata and formats thus leading to better decision making, business intelligence and productivity improvements.

Expecting sharing of data between companies is awkward because of the need to get an edge in business. Sharing data about their clients and operations threatens the culture of secrecy and competitiveness.

C. 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.

The transportation of data from storage point to processing point can be avoided in two ways. One is to process in the storage place only and results can be transferred or transport only that data to computation which is important. But both these methods would require integrity and provenance of data to be maintained.

Processing of such large amount of data also takes large amount of time. To find suitable elements whole of data Set needs to be Scanned which is somewhat not possible .Thus Building up indexes right in the beginning while collecting and storing the data is a good practice and reduces processing time considerably.

D. Analytical challenges

The main challenging questions are as:

- What if data volume gets so large and varied and it is not known how to deal with it?
- Does all data need to be stored?
- Does all data need to be analyzed?
- How to find out which data points are really important?
- How can the data be used to best advantage?

Big data brings along with it some huge analytical challenges. The type of analysis to be done on this huge amount of data which can be unstructured, semi structured or structured requires a large number of advance skills. Moreover the type of analysis which is needed to be done on the data depends highly on the results to be obtained i.e. decision making. This can be done by using one using two techniques: either incorporate massive data volumes in analysis or determine upfront which Big data is relevant.

E. Skill Requirement

Since Big data is at its youth and an emerging technology so it needs to attract organizations and youth with diverse new skill sets. These skills should not be limited to technical ones but also should extend to research, analytical, interpretive and creative ones. These skills need to be developed in individuals hence requires training programs to be held by the organizations. Moreover the Universities need to introduce curriculum on Big data to produce skilled employees in this expertise.

F. Technical Challenges

1) 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.

But sometimes it's quite possible that that the whole computation can't be divided into such independent tasks. There could be some tasks which might be recursive in nature and the input of the previous task is the input to the next computation. Thus restarting the whole computation becomes cumbersome process. This can be avoided by applying Checkpoints which keeps the state of the system at certain intervals of the time. In case of any failure, the computation can restart from last checkpoint maintained.

2) 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. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters. These factors combined put the concern on how to express the programs, even complex machine learning tasks.

There has been a huge shift in the technologies being used. Hard Disk Drives (HDD) are being replaced by the solid state Drives and Phase Change technology which are not having the same performance between sequential and random data transfer. Thus what kind of storage devices are to be used is again a big question for data storage.

3) 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.

This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it etc.

4) 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. Working with unstructured data is cumbersome and of course costly too. Converting all this unstructured data into structured one is also not feasible.

Structured data is the one which is organized in a way so that it can be managed easily. Digging through unstructured data is cumbersome and costly.

V. TOOLS AND TECHNIQUES AVAILABLE

The following tools and techniques are available:

A. Hadoop

Hadoop is an open source project hosted by Apache Software Foundation. It consists of many small 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.

The main problem is of combining the data being read from different devices. Many a methods are available in distributed computing to handle this problem but still it is quite challenging. All the problems discussed are easily handled by Hadoop. The problem of failure is handled by the Hadoop Distributed File System and problem of combining data is handled by Map reduce programming Paradigm. Map Reduce basically reduces the problem of disk reads and writes by providing a programming model dealing in computation with keys and values.

Hadoop thus provides: a reliable shared storage and analysis system. The storage is provided by HDFS and analysis by MapReduce.

B. Hadoop Components in detail

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. namenode (the master) and number of datanodes (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. The datanode stores and retrieve blocks as per the instructions of clients or the namenode. The data retrieved is reported back to the namenode with lists of blocks that they are storing. Without the namenode it is not possible to access the file. So it becomes very important to make namenode resilient to failure.

These are areas where HDFS is not a good fit: Low-latency data access, Lots of small file, multiple writers and arbitrary file modifications.

2) 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. Worker handles map tasks (MAP WORKER) and reduce tasks (REDUCE WORKER) but not both.

The Master controller creates some number of map and reduce tasks which is usually decided by the user program. The tasks are assigned to the worker nodes by the master controller. Track of the status of each Map and Reduce task (idle, executing at a particular Worker or completed) is kept by the Master Process. On the completion of the work assigned the worker process reports to the master and master reassigns it with some task.

The failure of a compute node is detected by the master as it periodically pings the worker nodes. All the Map tasks assigned to that node are restarted even if it had completed and this is due to the fact that the results of that computation would be available on that node only for the reduce tasks. 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.

C. Comparison of Hadoop Technique with other system Techniques

1) Comparison with HPC and Grid Computing Tools: The approach in HPC and Grid computing includes the distribution of work across a cluster and they are having a common shared File system hosted by SAN. The jobs here are mainly compute intensive and thus it suits well to them unlike as in case of Big data where access to larger volume of data as network bandwidth is the main bottleneck and the compute nodes start becoming idle. Map Reduce component of Hadoop here plays an important role by making use of the Data Locality property where it collocates the data with the compute node itself so that the data access is fast.

HPC and Grid Computing basically make use of the API's such as message passing Interface (MPI). Though it provides great control to the user, the user needs to control the mechanism for handling the data flow. On the other hand Map Reduce operates only at the higher level where the data flow is implicit and the programmer just thinks in terms of key and value pairs. Coordination of the jobs on large distributed systems is always challenging. Map Reduce handles this problem easily as it is based on shared-nothing architecture i.e. the tasks are independent of each other. The implementation of Map Reduce itself detects the failed tasks and reschedules them on healthy machines. Thus the order in which the tasks run hardly matters from programmer's point of view. But in case of MPI, an explicit management of check pointing and recovery system needs to be done by the program. This gives more control to the programmer but makes them more difficult to write.

2) Comparison with Volunteer Computing Technique: In Volunteer computing work is broken down into chunks called work units which are sent on computers across the world to be analyzed. After the completion of the analysis the results are sent back to the server and the client is assigned with another work unit. In order to assure accuracy, each work unit is sent to three different machines and the result is accepted if atleast two of them match. This concept of Volunteer Computing makes it look like MapReduce. But there exists a big difference between the two the tasks in case of Volunteer Computing are basically CPU intensive. This tasks makes these tasks suited to be distributed across computers as transfer of work unit time is less than the time required for the computation whereas in case of MapReduce is designed to run jobs that last minutes or hours on trusted, dedicated hardware running in a single data center with very high aggregate bandwidth interconnects.

3) Comparison with RDBMS: The traditional database deals with data size in range of Gigabytes as compared to MapReduce dealing in petabytes. The Scaling in case of MapReduce is linear as compared to that of traditional database. In fact the RDBMS differs structurally, in updating, and access techniques from MapReduce.

VI. BIG DATA GOOD PRACTICES

- Creating dimensions of all the data being store is a good practice for Big data analytics. It needs to be divided into dimensions and facts.
- All the dimensions should have durable surrogate keys meaning that these keys can't be changed by any business rule and are assigned in sequence or generated by some hashing algorithm ensuring uniqueness.
- Expect to integrate structured and unstructured data as all kind of data is a part of Big data which needs to be analyzed together.
- Generality of the technology is needed to deal with different formats of data. Building technology around key value pairs work.
- Analyzing data sets including identifying information about individuals or organizations privacy is an issue whose importance particularly to consumers is growing as the value of Big data becomes more apparent.
- Data quality needs to be better. Different tasks like filtering, cleansing, pruning, conforming, matching, joining, and diagnosing should be applied at the earliest touch points possible.
- There should be certain limits on the scalability of the data stored.
- Business leaders and IT leaders should work together to yield more business value from the data. Collecting, storing and analyzing data comes at a cost. Business leaders will go for it but IT leaders have to look for many things like technological limitations, staff restrictions etc. The decisions taken should be revised to ensure that the organization is considering the right data to produce insights at any given point of time.
- Investment in data quality and metadata is also important as it reduces the processing time.

VII. CONCLUSION

This paper described the new concept of Big data, its importance and the existing projects. To accept and adapt to this new technology many challenges and issues exist which need to be brought up right in the beginning before it is too late. All those issues and challenges have been described in this paper. These challenges and issues will help the business organizations which are moving towards this technology for increasing the value of the business to consider them right in the beginning and to find the ways to counter them. 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.