IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. IV (Nov.-Dec. 2016), PP 13-18
www.iosrjournals.org

Big Data: Emerging Challenges of Big Data and Techniques for Handling

Dr.M.Padmavalli[1]

[1] *Department of Computer Science and Applications, Sri Krishnadevaraya University, Andhra Pradesh, India*

Abstract: *Big Data is the large amount of data that cannot be processed by making use of traditional methods of data processing. Due to widespread usage of many computing devices such as smartphones, laptops, wearable computing devices; billions of people are connected to internet worldwide, generating large amount of data at the rapid rate. The data processing over the internet has exceeded more than the modern computers can handle. Due to this high growth rate, the term Big Data is envisaged. However, the fast growth rate of such large data generates numerous challenges, such as data analysis, Storage, quering, inconsistency and incompleteness, scalability, timeliness, and security. Key industry segments are heavily represented; financial services, where data is plentiful and data investments are substantial, and life sciences, where data usage is rapidly emerging. This paper provides a brief introduction to the Big data technology and its importance in the contemporary world. This paper addresses various challenges and issues that need to be emphasized to present the full influence of big data. The tools used in Big data technology are also discussed in detail. This paper also discusses the characteristics of Big data and the platform used in handling the Big Data as well as the efforts expected on big data mining.*

Keywords: *Big Data, Big Table Hadoop, MapReduce, ETL Process, Big Data Mining*

I. Introduction

Big Data has gained much attention from the last few years in the IT industry. As we can witness billions of people are connected to internet worldwide, generating large amount of data at the rapid rate. The generation of this large amount of engenders various challenges.The generation of this large amount of engenders various challenges. Along with Big Data„s huge benefits to many organizations, the challenges and issues should also be brought into light. A forecast from International Data Corporation (IDC), the Big Data technology and services market represents a fast-growing multibilliondollar worldwide opportunity. In fact, a recent IDC forecast shows that the Big Data technology and services market will grow at a 26.4% compound annual growth rate to $41.5 billion through 2018, or about six times the growth rate of the overall information technology market. Additionally, by 2020 IDC believes that line of business buyers will help drive analytics beyond its historical sweet spot of relational (performance management) to the double-digit growth rates of realtime intelligence and exploration/discovery of the unstructured worlds.The services-related opportunity will account for more than half of all big data and business analytics revenue for most of the forecast period, with IT Services generating more than three times the annual revenues of Business Services. Software will be the second largest category, generating more than $55 billion in revenues in 2019. Nearly half of these revenues will come from purchases of End-User Query, Reporting, and Analysis Tools and Data Warehouse Management Tools. Hardware spending will grow to nearly $28 billion in 2019.

The theme of this paper is to provide an in-depth study on the issue of big data mining, its challenges and the perceivable opportunities. The point to a few topics those are either promising or much needed for solving the big data and big data mining problems. In order to make the discussion logical and smooth, need to start with a review of some essential and relevant concepts, including data mining, big data, big data mining, and the frameworks/platforms related to big data and big data mining.

II. Big Data Overview

Big Data is a compendium of big datasets that cannot be processed using traditional computing techniques. It is not a technique that can be worked on its own or in isolation; rather it involves many areas of business and Technology. The properties of signify Big Data are volume, Variety, Velocity, Variability and Complexity as discribed below.

1. Volume: Many factors contribute towards increasing volume streaming data, live streaming data and data collected from sensors etc.,

2. Variety: Data comes in all types of formats-from traditional databases, text documents, vedios, audios, emails, transactions etc.,

3. Velocity: This means how fast the data is being prodcued and how fast the data needs to be processed to meet the demands and the challenges which lie ahead in the path of growth and development.

4. Variability: Along with the Velocity, the data flows can highly inconsistent with periodic peaks. This is a factor which can be a problem for those who are analyse the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.
5. Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Big Data Data comes mainly in two forms- 1. Structured, and 2. Unstructured Data (there are also semi-structured data – eg. XML) structured data has semantic meaning attached to it whereas Unstructured data has no latent meaning. The growth in data that we are referring is most unstructured data. Below are few examples of unstructured data.

1. Calls, text, tweet, net surf, browse through various websites each day and exchange messages via several means.2. Social media usage my several million people for exchanging data in various forms also forms a part of Big Data. 3. Transactions made through card for various payment issues in large numbers every second across the world also constitutes the Big Data.
Big data involves the data produced by different devices and applications. Sources of Big Data can be broadly classified into six different categories as shown in Fig 1.

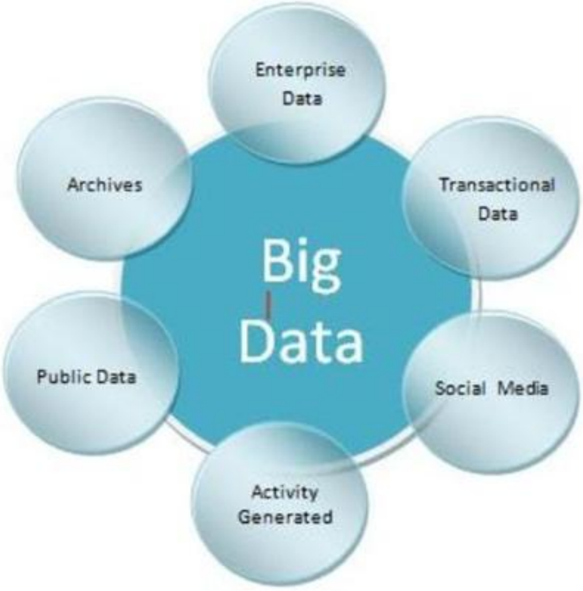


Fig1. Sources of Big Data

III. Big data Mining

Big data mining is referred to the collective data mining or extraction techniques that are performed on large sets/volume of data or the big data. Big data mining is primarily done to extract and retrieve desired information or pattern from humongous quantity of data.

The goals of big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters. Analyzing fast and massive stream data may lead to new valuable insights and theoretical concepts. Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous big data has the potential to maximize our knowledge and insights in the target domain. However, this brings a series of new challenges to the research community. Overcoming the challenges will reshape the future of the data mining technology, resulting in a spectrum of groundbreaking data and mining techniques and algorithms. One feasible approach is to improve existing techniques and algorithms by exploiting massively parallel computing architectures (cloud platforms in our mind). Big data mining must deal with heterogeneity, extreme scale, velocity, privacy, accuracy, trust, and interactiveness that existing mining techniques and algorithms are incapable of. The need for designing and implementing very-large-scale parallel machine learning and data mining algorithms (ML-DM) has remarkably increased, which accompanies the emergence of powerful parallel and very-large-scale data processing platforms, e.g., Hadoop MapReduce.

IV. Trends And Technologies In Big Data Processing

Lifecycle for Big Data processing and classifies various available tools and technologies in terms of the lifecycle phases of Big Data, which include data acquisition, data storage, data analysis, and data exploitation of the results.

Before processing Big data it must be recorded from various data generating sources. After recording, it must be filtered and compressed. Only the relevant data should be recorded by means of filters that discard useless information. In order to facilitate this work specialized tools are used such as ETL. ETL tools represent the means in which data actually gets loaded into the warehouse. The figure 3 demonstrates different stages in the process.

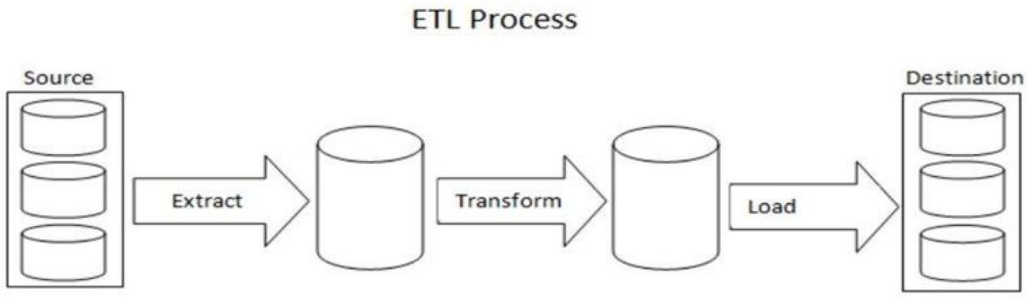


Figure3.ETLprocess

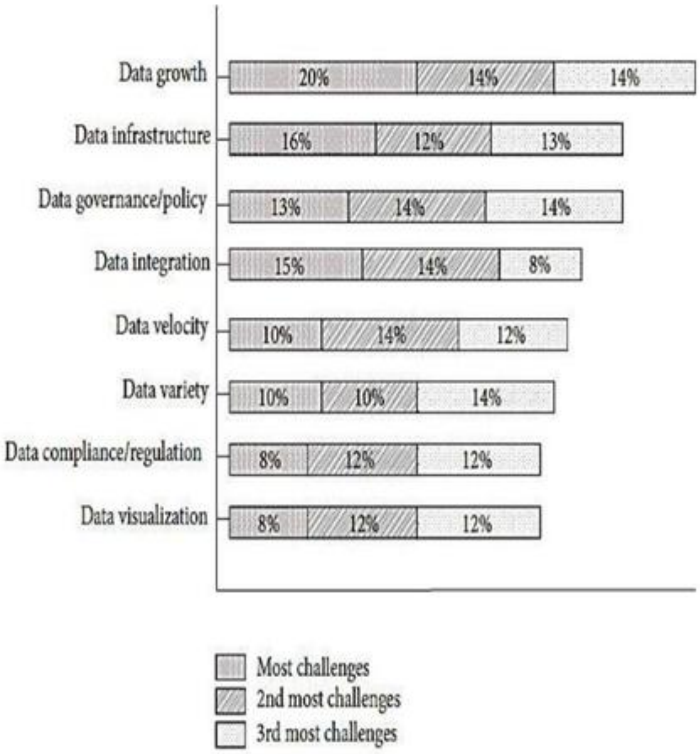
Phases in ETL Process:

1. Extraction: In this phase relevant information is extracted. To make this phase efficient, only the data source that has been changed since recent last ETL process is considered.
2. Transformation: Data is transformed through various phases.The phases are 1. Data analysis; 2. Definition of transformation workflow and mapping rules; 3. Verification; 4. Transformation; and 5. Backflow of cleaned data.
3. Loading: At the last, after the data is in the required format, it is then loaded into the data warehouse/Destination.

V. Big data Challenges

Big data is set to offer companies tremendous insight. But with terabytes and petabytes of data pouring in to organizations today, traditional architectures and infrastructures are not up to the challenge. Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. Considering the business impacts of these challenges suggests some serious risks to successfully deploying a big data program. Below is the list of some of the challenges in Big data along with its challenge, impact and risks involved.

Challenge	**Impact**	**Risk**
Uncertainty of the market Landscape	Difficulty in choosing technology components Vendor lock-in	Committing to failing product or failing vendor
Big data talent gap	Steep learning curve Extended time for design, development, and implementation	Delayed time to value
Big data loading	Increased cycle time for analytical platform data population	Inability to actualize the program due to unmanageable data latencies
Synchronization	Data that is inconsistent or out of date	Flawed decisions based on flawed data
Big data accessibility	Increased complexity in syndicating data to end-user discovery tools	Inability to appropriately satisfy the growing community of data consumers



VI. Techniques For Big Data Handling

Efficiently managing today's big data challenges requires a robust data integration strategy backed by leading-edge data technologies and services that lets you easily connect to and access your data, wherever it resides.

This can be achieved through the development of informed processes that take advantage of best-of-breed data integration technologies that address your evolving challenges and eliminate the correlated risks. Some key characteristics of these technologies include:

- Accessing data stored in a variety of standard configurations.
- Relying on standard relational data access methods.
- Enabling canonical means for virtualizing data accesses to consumer applications „
- Employ push-down capabilities of a wide variety of data management systems (ranging from conventional RDBMS data stores to newer NoSQL approaches) to optimize data access „
- Rapid application of data transformations as data sets is migrated from sources to the big data target platforms.

There are many techniques available for data management. The Big Data handling techniques and tools include Hadoop, MapReduce, Simple DB, Google BigTable, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort. Out of these, Hadoop is one of the most widely used technologies.

1. Predictive analytics: software and/or hardware solutions that allow firms to discover, evaluate, optimize, and deploy predictive models by analyzing big data sources to improve business performance or mitigate risk.

2. NoSQL databases: key-value, document, and graph databases.

3.Search and knowledge discovery: tools and technologies to support self-service extraction of information and new insights from large repositories of unstructured and structured data that resides in multiple sources such as file systems, databases, streams, APIs, and other platforms and applications.

4. Stream analytics: software that can filter, aggregate, enrich, and analyze a high throughput of data from multiple disparate live data sources and in any data format.

5. In-memory data fabric: provides low-latency access and processing of large quantities of data by distributing data across the dynamic random access memory (DRAM), Flash, or SSD of a distributed computer system.

6. Distributed file stores: a computer network where data is stored on more than one node, often in a replicated fashion, for redundancy and performance.

7. Data virtualization: a technology that delivers information from various data sources, including big data

sources such as Hadoop and distributed data stores in real-time and near-real time.
8. Data integration: tools for data orchestration across solutions such as Amazon Elastic MapReduce (EMR), Apache Hive, Apache Pig, Apache Spark, MapReduce, Couchbase, Hadoop, and MongoDB.
9.Data preparation: software that eases the burden of sourcing, shaping, cleansing, and sharing diverse and messy data sets to accelerate data's usefulness for analytics.
10. Data quality: products that conduct data cleansing and enrichment on large, high-velocity data sets, using parallel operations on distributed data stores and databases.

Hadoop
Hadoop is an Apache open source framework which is written in java. High volumes of data, in any structure, are processed by Hadoop. Hadoop allows distributed storage and distributed processing for very large data sets. The main components of Hadoop are:

1. Hadoop distributed file system (HDFS)
2. MapReduce

Hadoop has three layers. The two major layers are MapReduce and HDFS. The architecture of Hadoop is shown in the figure 4.

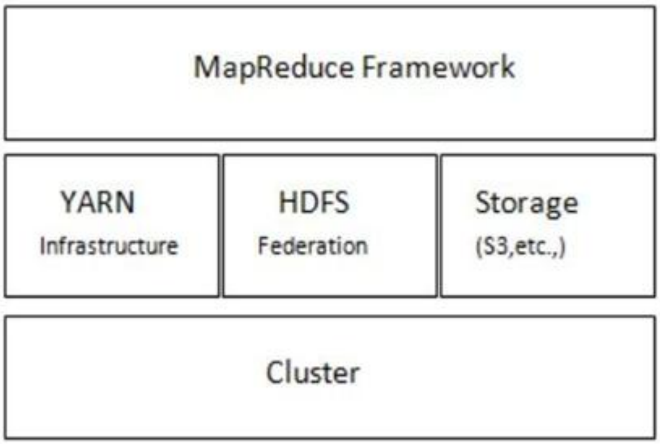


Figure 4 .Architecture of Hadoop

HDFS (Storage layer):- Hadoop has a distributed File System called HDFS, which stands for Hadoop Distributed File System. It is a File System used for storing very large files with streaming data access patterns, running on clusters on commodity hardware. There are two types of nodes in HDFS cluster ,namely namenode and datanodes. 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 name node resilient to failure.

Map Reduce (Processing/Computation layer):- It is a programing paradigm which is meant for managing applications on multiple distributed servers. In MapReduce divide and conquer method is used to break the large complex data into small units and process them. It reads the data from HDFS in an optimal way. However, it can read the data from other places too; including mounted local file systems, the web, and databases. It divides the computations between different computers (servers, or nodes). It is also fault-tolerant. If some of nodes fail, Hadoop knows how to continue with the computation, by re-assigning the incomplete work to another node and cleaning up after the node that could not complete its task. It also knows how to combine the results of the computation in one place]. The other core components in Hadoop architecture includes Hadoop YARN, it is a framework for job scheduling and cluster resource management. The other component is the cluster which is the set of host machines (nodes).

Big table: Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers. Many projects at Google store data in Bigtable, including web indexing, Google Earth, and Google Finance. These applications place very different demands on Bigtable, both in terms of data size (from URLs to web pages to satellite imagery) and latency requirements (from backend bulk processing to real-time data serving). Despite these varied demands, Bigtable has successfully provided a flexible, high-performance solution for all of these Google products.

VII. Conclusion

As there are huge volumes of data that are produced every day, so such large size of data it becomes very challenging to achieve effective processing using the existing traditional techniques. Big data is data that exceeds the processing capacity of conventional database systems. In this paper fundamental concepts about Big

Data are presented. These concepts include Big Data characteristics, challenges and techniques for handling big data and Big Data Mining. Big data discloses the limitations of existing data mining techniques, resulted in a series of new challenges related to big data mining. Big data mining is a promising research area, still in its infancy. In spite of the limited work done on big data mining so far, we believe that much work is required to overcome its challenges related to heterogeneity, scalability, speed, accuracy, trust, provenance, privacy, and interactiveness. This paper also provides an overview of frameworks/platforms for processing and managing big data as well as platforms and libraries for mining big data.

References

[1]. Golfarelli, M., & Rizzi, S. (2009). Data warehouse design: modern principles and methodologies. Columbus: McGraw-Hill

[2]. https://www.idc.com/prodserv/4Pillars/bigdata

[3]. www.Wikibon.org

[4]. Berkovich, S., Liao, D.: On Clusterization of big data Streams. In: 3rd International Conference on Computing for Geospatial Research and Applications, article no. 26. ACM Press, New York (2012)

[5]. A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices." Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.]

[6]. Almeida, F., and Calistru, C, "The Main Challenges and Issues of Big Data Management", International Journal of Research Studies in Computing, 2(1), 2013, pp. 11-20.

[7]. ABADI, D. J., MADDEN, S. R., AND FERREIRA, M. C. Integrating compression and execution in columnoriented database systems. Proc. of SIGMOD (2006).

[8]. https://www.progress.com

[9]. M. Chen, S. Mao, and Y. Liu, “Big data: a survey,” Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014

[10]. Apache Hadoop (2013). HDFS Architecture Guide [Online]. Available: https://hadoop. apache.org/docs/r1.2.1/hdfs_design.ht

[11]. Amrit pal, Pinki Aggrawal, Kunal Jain, Sanjay Aggrawal “A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data using Hadoop” Forth International Conference on Communication Systems and Network Technologies, 2014. [10] Rahm, E., & Hai Do, H. (2000). Data cleaning: problems and current approaches. Bulletin of the Technical Committee on Data Engineering, 23(4), 3-13.),

[12]. Apache Hadoop (2013).HDFS Architecture Guide [Online]. Available: https://hadoop. apache.org/docs/r1.2.1/hdfs_design.ht Intel,“BigdataAnalaytics,”2012,http://www.intel.com/content/dam/www/public/ us/en/documents/reports/data-insightspeerresearch-report.pdf [13]https://www.progress.com/docs/default-source/default-document-library/Progress/Documents/Papers/Addressing-Five-Emerging-Challenges-of-Big-Data.pdf

[13]. https://www.sas.com/resources/asset/five-big-data-challenges-article.pdf