Monali Sahoo *et al*, International Journal of Computer Science and Mobile Applications,
Vol.6 Issue. 6, June- 2018, pg. 23-32 **ISSN: 2321-8363**
UGC Approved Journal
Impact Factor: 5.515

The Emergence of Big Data: A Survey

Monali Sahoo[1], Siddharth Swarup Rautaray[2], Manjusha Pandey[2]

[1] eCentric Solutions Pvt. Ltd., India
mmonali.sahoo@gmail.com
[2] School of Computer Engineering, KIIT University, India
siddharthfcs@kiit.ac.in, manjushapandey82@gmail.com

ABSTRACT

Since last two decades, data is growing exponentially due to the rapid evolution of new technologies, devices and communication means. This voluminous data with diverse variety generated with high velocity transform Data into Big Data. It is extensively used in marketing field, sales, banking, finance sector, healthcare, social media, tourism etc. But due to its 'Big' features in every aspect, it becomes difficult to handle it by traditional data processing applications. There are so many challenges while handling big data as difficulties lie in data capture, storage, searching, sharing, analysis and visualization. As large datasets are usually non-relational or unstructured, thus processing such data sets poses a significant challenge. Therefore, Big Data Analytics becomes the demanding field for researchers. It is not a single technology, but a data-driven approach used to develop and deploy customized solutions as it analyzes a large amount of data to uncover hidden patterns, correlations and other insights.

This survey paper, discusses the explicit characteristics of data which make it big data, different technologies required for processing of big data, applications and challenges in those applications for usage of big data. The objective is to explore challenges and identify the research gaps which will help researchers to provide effective and efficient solutions for real life problems in handling big data.

KEYWORDS: Data, Big Data, Data analytics, Data Product, Hadoop.

1. Introduction

Data plays vital role in every field of real world. Data is of three types: structured, un-structured, semi-structured [1]. Few decades ago most of the data used were in structured and semi-structured format which were easy to analyze, process and store using traditional database applications. But with the fast growing digital world different dimensions of data are also growing rapidly. This leads to the transformation of Data into Big Data. In current scenario there are 80% of un-structured data and 20% of structured data. That means un-structured data are growing more in comparison to structured data. According to various survey now the size of the data is reached to 4.4 zettabytes and is supposed to reach around 44 zettabytes or 44 trillion gigabytes by 2020[2].

There are some characteristics that differentiate 'Big Data' from 'Data' like volume, velocity, and variety (3 V's) [2]. Volume refers to the amount of data, variety refers to heterogeneous formats of data and velocity refers to the speed of data processing. But this definition of big data is not confined to only these three characteristics rather the characteristics are augmenting day by day based on ongoing research work.

Hence it becomes a tedious task to manage such big data with on-hand database system. Big data analytics comes into picture to analyze large amount of data to uncover hidden patterns, correlations and other insights. Hadoop is one of the big data analytical tools that allows us to store and process large data sets in parallel and distributed fashion. Apache hive, Apache pig, Apache spark, Apache HBase are the basic components of Hadoop framework [3, 4].

Big data analytics is used in wide range of areas such as Banking & Securities, Communications, Media and Entertainment, Healthcare Providers, Education, Manufacturing and Natural Resources, Government, Insurance

 ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

etc. For example, in banking sectors the Securities Exchange Commission (SEC) is using big data to monitor financial market activity. In healthcare domain Big data is used for analyzing data in the electronic medical record (EMR) system with the goal of reducing costs and improving patient care [5]. Big data is changing the media and entertainment industry, giving users and viewers a much more personalized and enriched experience.

The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools such as data mining and statistical analysis. Though the mining of Big Data offers many attractive opportunities, however, researchers and professionals are facing numerous challenges while discovering Big Data sets and while mining value and knowledge from such information. The obstacles lye at different stages including: data capture, storage, searching, sharing, analysis, management and visualization [6]. In this paper we provide a depth insight into big data challenges as well as research challenges for future research work.

This paper is organized as follows: Section 2 represents data to big data evolution. In section 3 different characteristics of big data are explained. Section 4 describes various technologies used for solving the big data problems. Section 5 describes the implementation of big data analytics in different application domains. Section 6 represents the challenges in big data and research direction. The conclusion and the future work is presented in Section 7.

2. Data to Big Data

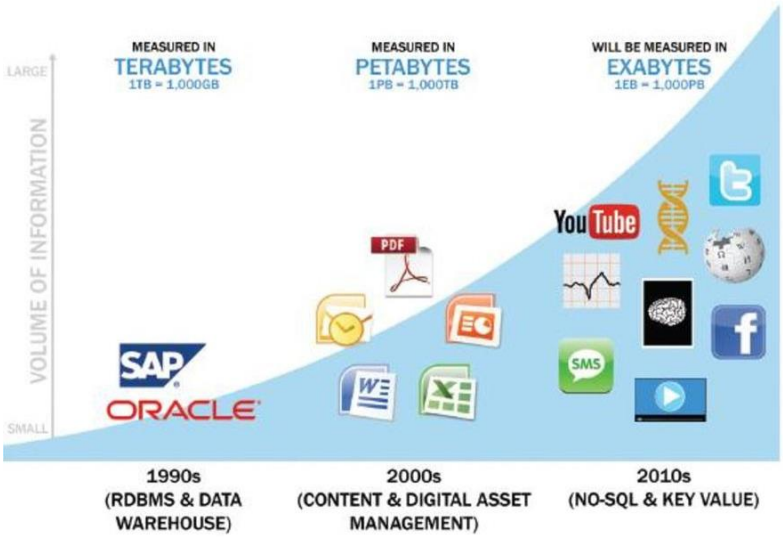

Figure 1: Data Evolution and the Rise of Big Data Sources
[Source: EMC Education Services, Data science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015]

In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. The term big data is used for data that go beyond the processing power of traditional database systems [6]. In 90's structured data were generated and stored in RDBMS and data warehouse. Gradually semi-structured and un-structured data are generated more in comparison to structured data. The rapid growth of new technologies, devices and communication means are responsible for creating such a huge amount of data. Internet of things contribute a major part in the generation of big data through internet devices and smart devices. In social media, for example in Facebook we are getting around 42 likes in every minute, at Twitter almost 3.5 lakhs tweets happen per minute, on YouTube almost 300 hours of data are being generated per minute[7]. Likewise massive amount of data are also generated through Instagram and through other social media.

 ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

From the above few examples it is understood how fast data size is increasing having heterogeneous format. To get appropriate business insight for growth in business, these data should be analyzed and used in the proper manner.

3. 9 V's

Initially Big data was defined with its 3V's characteristics. But it is no more confined with these three characteristics due to the growth of data in various dimensions. Till now 9 V's characteristics are discovered from various research work for defining big data [8]. These 9V's characteristics are: Volume, Variety, Velocity, Veracity, Validity, Variability, Volatility, Visualization and Value.

Volume: The term 'Big Data' itself refers to a huge amount of data. Volume is one characteristic which needs to be considered while dealing with 'Big Data' as it helps in determining the data to be considered as big data or not.

Variety: It refers to heterogeneous type and nature of data (structured, semi-structured, unstructured). During earlier days, most of the applications were using spreadsheets and databases to store the data as most of the data were structured. But now days, data are in the form of emails, photos, PDFs, log, audio, videos, etc. This variety of unstructured data leads to certain issues of storage, mining and analyzing data.

Velocity: The term velocity states about the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

Veracity: Big Data veracity refers to the noise, inconsistencies and uncertainty in data. As reliability of the data source depends on veracity, knowledge of the data's veracity helps us better understand the risks associated with analysis and business decisions based on the particular data set.

Validity: Big Data veracity is a matter of validity. It refers to how accurate and correct the data is for its intended use. Valid data is responsible for making the right decisions. Data validation ensures uncorrupted transmission of data.

Variability: Variability in big data's context refers to a few different things. One is the number of inconsistencies in the data. Big data is also variable because of the multitude of data dimensions resulting from multiple dissimilar data types and sources. Variability can also refer to the inconsistent speed at which big data is loaded into your database.

Volatility: volatility means retention period of datasets. Once it expires, we can easily destroy the data. For example: an online ecommerce company may not want to keep a one year customer purchase history. Because after one year as default warranty on their product expires so there is no possibility of these data restoration.

Visualization: Visualization means complex graphs that can include several variables of data while still remaining understandable and readable. It is the hard part of Big Data which makes that huge amount of data comprehensible, easy to understand and readable. With the right analysis and visualizations, raw data can be put to use, otherwise it remains essentially useless.

Value: Value is one of the most significant characteristics of big data. The other characteristics of big data are meaningless if the business value can't be derived from the data. Substantial value can be found in big data, including understanding the customers of different business domains better, targeting them accordingly, optimizing processes, and improving machine or business performance.

These characteristics play crucial role in data analysis and if considered properly during analysis, it can lead to an optimum result.

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

4. Application domains in Big Data

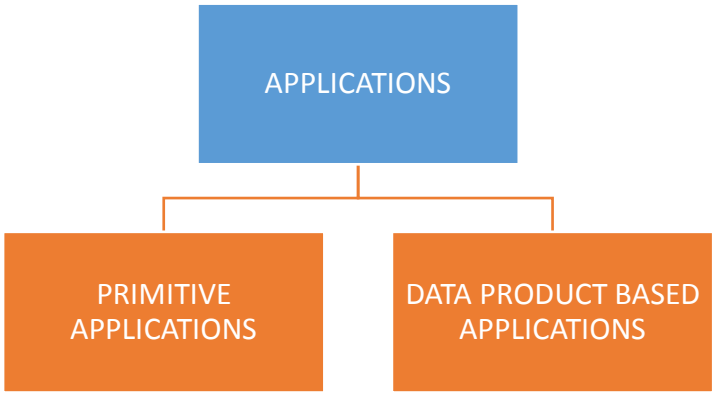

Figure 2: Categorization of application domains

Applications are categorized into two types: Primitive Applications and Data product based applications.

4.1 Primitive applications

These applications are developed in a very early period and are going through various analysis. Examples: Banking and Securities, Healthcare Providers, Manufacturing and Natural Resources, Retail and Whole sale trade, Transportation.

Banking and Securities

The Securities Exchange Commission (SEC) is taking help of big data to monitor financial market activity [9]. Retail traders, big banks and hedge funds in the financial markets use big data for trade analytics used in high frequency trading, pre-trade decision-support analytics, sentiment measurement, predictive analytics etc. This industry also greatly relies on big data for risk analytics including; anti-money laundering, demand enterprise risk management, "Know Your Customer", and fraud mitigation.

Healthcare Providers

Due to unavailability, inadequacy and unusability of electronic data, the healthcare sector has been facing failures in utilizing the data to curb the cost of rising healthcare. Additionally, the healthcare databases that hold health-related information have made it difficult to link data that can show patterns useful in the medical field. Thus big data analytics becomes a demanding technology in healthcare. Some instances are listed below:
Beth Israel hospitals are using data collected from a cell phone app, from millions of patients, to permit doctors to use evidence-based medicine as opposed to administering several medical lab tests to all patients who go to the hospital.
Free public health data and Google Maps have been used by the University of Florida to create visual data that allows for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease.

Manufacturing and Natural Resources

In the natural resources industry, big data used for predictive modeling to support decision making that has been utilized to integrate large amounts of data from geospatial data, graphical data, text and temporal data. Areas of interest where this has been used include; seismic interpretation and reservoir characterization.

 ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

Retail and Whole sale trade

From traditional brick and mortar retailers and wholesalers to current day e-commerce traders, the industry has gathered a lot of data over time. This data, derived from customer loyalty cards, POS scanners, RFID etc. is not being used enough to improve customer experiences on the whole. Any changes and improvements made have been quite slow. In New York's Big Show retail trade conference in 2014, companies like Microsoft, Cisco and IBM pitched the need for the retail industry to utilize big data for analytics and for other uses like optimized staffing through data from shopping patterns, to reduced fraud and timely analysis of inventory [10].

Transportation

Government use of big data in transportation are traffic control, route planning, intelligent transport systems, congestion management. Private sector use of big data in transport are revenue management, technological enhancements, logistics and for competitive advantage. Individual use of big data are route planning to save on fuel and time, for travel arrangements in tourism etc.

4.2 Data product based applications

These are the applications that use data products built through the analysis of data in an effort to engage the consumer and to solve a problem with data. Examples: Communications, Media and Entertainment, Education, Insurance.

Proper and correct analysis of available data can transform major business processes with the help of big data analytics. Such business sectors are:

Communications, Media and Entertainment

Organizations in this sector analyze customer data along with behavioral data to create detailed customer profiles in order to create content for different target audiences, recommend content on demand, and measure content performance. Some well know examples are; Big data helped Donald Trumph (the president of US) to win against Hillary Clinton in the US election. In 2012 FIFA world cup Germany won because they add a 12^{th} man and that 12^{th} man was big data analytics. Spotify, an on-demand music service, uses Hadoop big data analytics, to collect data from its millions of users worldwide and then uses the analyzed data to give informed music recommendations to individual users. Amazon Prime, which is driven to provide a great customer experience by offering, video, music and kindle books in a one-stop shop also heavily utilizes big data [9, 10].

Education

Big data is used quite significantly in higher education. For example, The University of Tasmania, an Australian university with over 26000 students, has deployed a Learning and Management System that tracks among other things, when a student logs onto the system, how much time is spent on different pages in the system, as well as the overall progress of a student over time [10].

Insurance

In these industries big data helps in analyzing and predicting customer behavior through the derived data of social media, GPS-enabled devices and CCTV footage to provide customer insights for transparent and simpler products. It also allows for better customer retention from insurance companies [11].

In current scenario big data analytics has been emerged as an inseparable part of most of the business process. To get the most out of big data opportunities one needs to familiarize with and understand industry-specific challenges.

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

5. Technologies in Big Data

Distributed architecture is the most suitable architecture to efficiently handle big data in order to increase the storage capacity and the processing power. It enables users to add multiple systems or nodes when there is an increase in data. Thus, making the performance of the processing data very high in comparison to those running in a single system (Vertical Scaling). This procedure of storing and processing data in a distributed architecture is known as Horizontal Scaling [12].

One of the most significant big data analytical tools is Hadoop framework that has adopted horizontal scaling as key design principle for storing and processing large data sets in parallel and distributed fashion. It is divided into two parts:

a) Hadoop Distributed File System(HDFS)
b) MapReduce

a) Hadoop Distributed File System (HDFS)

For storage purpose most commonly Hadoop Distributed File System (HDFS) is used. HDFS is the foundation for many big data frameworks, since it provides scalable and reliable storage. It allows to dump any kind of data across the cluster. It is a Java-based file system that provides scalable and reliable data storage, and it was designed to cover large clusters of commodity servers. HDFS has proven good in scaling and forming clusters of servers, which can support billions files and blocks. HDFS creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit. It has two core components i.e NameNode and DataNode. The NameNode is the main node that contains metadata about the data stored. Data is stored on the DataNodes which are commodity hardware in the distributed environment [12].

b) MapReduce

MapReduce is commonly used for big data processing. It allows parallel processing of the data stored in HDFS. It is a computational approach that involves breaking large volumes of data down into smaller batches, and processing them separately. The computation is done in terms of map and a reduce function. A cluster of computing nodes which are built on commodity hardware will scan the batches and aggregate their data. Then the multiple nodes' output gets merged to generate the final result data [12].

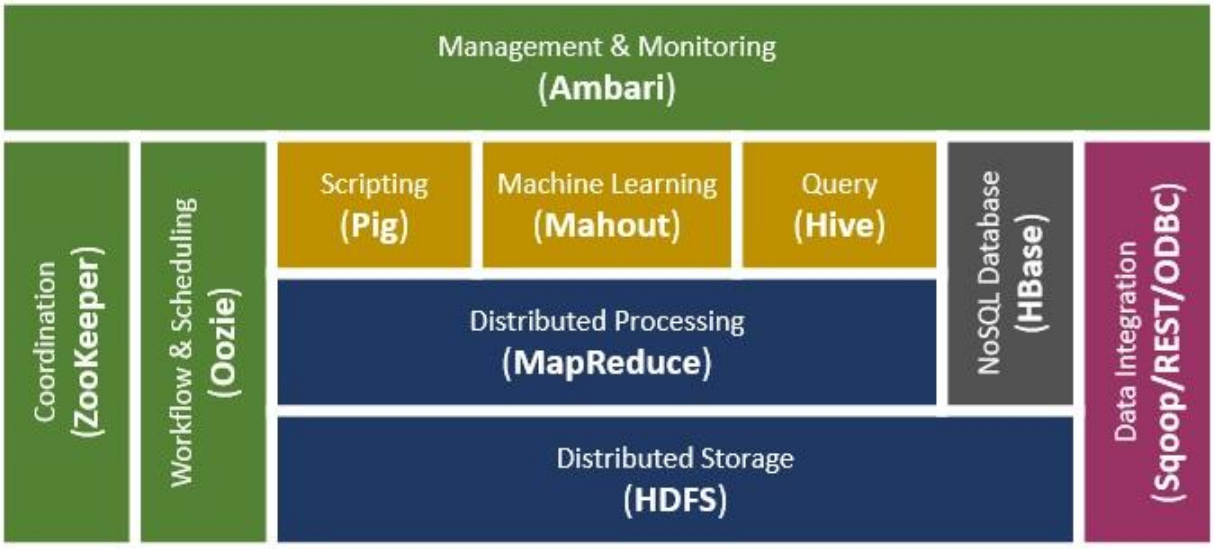

Figure 3: Apache Hadoop Ecosystem
[Source: https://www.dezyre.com/article/hadoop-ecosystem-components-and-its-architecture/114]

 ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

The Hadoop Ecosystem comprises of 4 core components: Hadoop common, Hadoop Distributed File System (HDFS), MapReduce, YARN [13].

Hadoop Common: Apache Foundation has pre-defined set of utilities and libraries that can be used by other modules within the Hadoop ecosystem. For example, if HBase and Hive want to access HDFS they need to make of Java archives (JAR files) that are stored in Hadoop Common.

YARN: YARN forms an integral part of Hadoop 2.0.YARN is great enabler for dynamic resource utilization on Hadoop framework as users can run various Hadoop applications without having to bother about increasing workloads.

HDFS and MapReduce components are already discussed above.

Other than these four main components there are several other Hadoop components that form an integral part of the Hadoop ecosystem with the intent of enhancing the power of Apache Hadoop like; providing better integration with databases, making Hadoop faster or developing novel features and functionalities. Some of the important Hadoop components used by enterprises extensively explained in Section 5.1 to 5.4.

5.1 Data Access Components of Hadoop Ecosystem- Pig and Hive

Pig: It is a convenient tool developed by Yahoo for analyzing huge data sets efficiently and easily [13]. It provides a high level data flow language Pig Latin which is easy to use. The most outstanding feature of Pig programs is that their structure is open to considerable parallelization making it easy for handling large data sets.

Hive: Hive developed by Facebook is a data warehouse built on top of Hadoop and provides a simple language known as HiveQL similar to SQL for querying, data summarization and analysis. Hive makes querying faster through indexing [13].

5.2 Data Integration Components of Hadoop Ecosystem- Sqoop and Flume

Sqoop: Sqoop component is used for importing data from external sources into related Hadoop components like HDFS, HBase or Hive. It can also be used for exporting data from Hadoop to other external structured data stores. Sqoop parallelizes data transfer, mitigates excessive loads, allows data imports, efficient data analysis and copies data quickly.

Flume: Flume component is used to gather and aggregate large amounts of data. Apache Flume is used for collecting data from its origin and sending it back to the resting location (HDFS).

5.3 Data Storage Component of Hadoop Ecosystem –HBase

HBase: HBase is a column-oriented database that uses HDFS for underlying storage of data. HBase supports random reads and also batch computations using MapReduce. With HBase NoSQL database enterprise can create large tables with millions of rows and columns on hardware machine. The best practice to use HBase is when there is a requirement for random ‘read or write’ access to big datasets.

5.4 Monitoring, Management and Orchestration Components of Hadoop Ecosystem- Oozie and Zookeeper

Oozie: It is a workflow scheduler where the workflows are expressed as Directed Acyclic Graphs. Oozie runs in a Java servlet container Tomcat and makes use of a database to store all the running workflow instances. The workflows in Oozie are executed based on data and time dependencies.

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

Zookeeper: Zookeeper is the king of coordination and provides simple, fast, reliable and ordered operational services for a Hadoop cluster. Zookeeper is responsible for synchronization service, distributed configuration service and for providing a naming registry for distributed systems.

6. Challenges

Challenges are actual implementation hurdles which need to be considered immediately. Therefore during implementation we have to handle these challenges, to avoid system failure and generation of abnormal results. Difficulties lie in data capture, storage, searching, sharing, analysis, and visualization. If we cannot conquer those challenges, Big Data will become a gold ore but we do not have the capabilities to explore it, especially when information surpass our capability to harness.

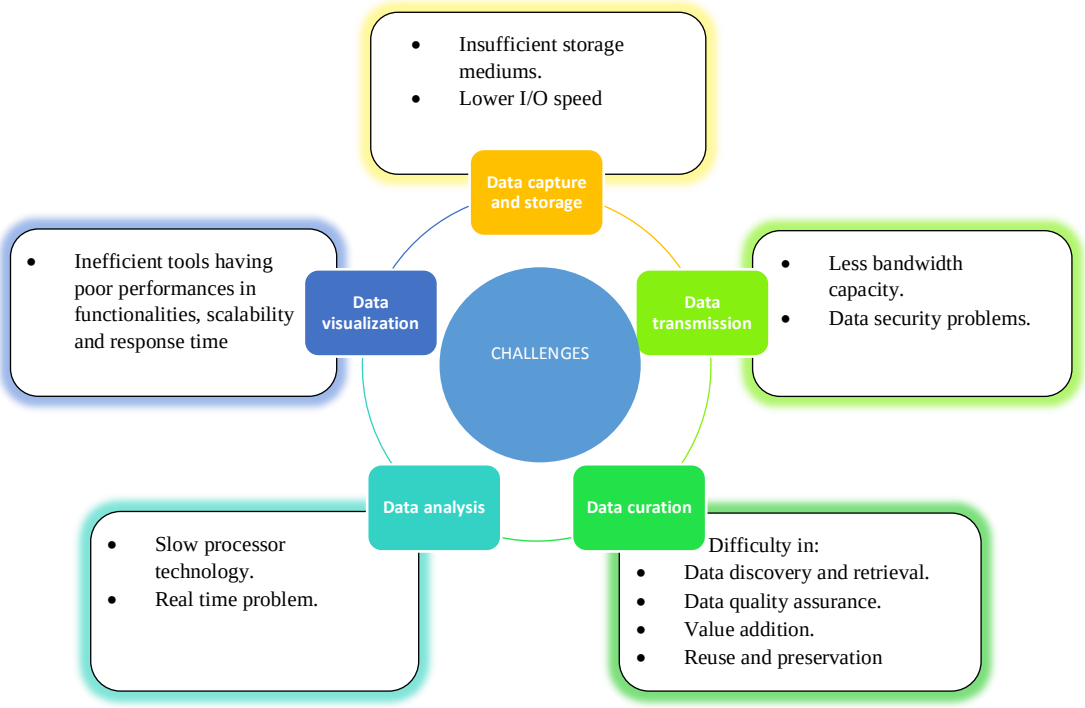

Figure 4: Identification of various phases of big data analytics and their challenges

Data capture and storage

In technical perspective the world's information storage capacity has roughly doubled around every 3 years since the 1980s [14]. It's obvious that we need great innovations to fulfill the requirement of more storage mediums and higher I/O speed to meet the challenges. Firstly, the accessibility of Big Data is on the top priority of the knowledge discovery process. Current storage technologies are incapable of performing better for both the sequential and random I/O simultaneously, which forces us to rethink how to design storage subsystems for Big Data processing systems. However, the existing storage architectures have severe drawbacks and limitations when it comes to large-scale distributed systems. Aggressive concurrency and per server throughput are the essential requirements for the applications on highly scalable computing clusters, and today's storage systems lack the both.

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

Data transmission

While dealing with large volume of communication, the network bandwidth capacity is treated as the bottleneck in cloud and distributed systems. Apart from that, cloud storage also creates data security problems as the requirements of data integrity checking.

Data curation

Data curation is focused at data discovery and retrieval, data quality assurance, value addition, reuse and preservation over time. But the size of big data keeps increasing exponentially and current capability to work with is only in the relatively lower levels of petabytes, exabytes and zettabytes of data.

Data analysis

As volume is the first characteristic of Big Data, so it is clear that scalability is the biggest and most important challenge while dealing with the Big Data analysis tasks. It has been observed that the data size is scaling much faster than CPU speeds, so there is a natural dramatic shift in processor technology. Although the clock cycle frequency of processors is doubling following Moore's Law, the clock speeds still highly lag behind. Alternatively, processors are being embedded with increasing numbers of cores which leads to the development of parallel computing. For those real-time Big Data applications, like navigation, social networks, finance, biomedicine, astronomy, intelligent transport systems, and internet of thing, timeliness is at the top priority. We can't guarantee the timeliness of response when the volume of data will be processed is very large. It is still a big challenge for stream processing involved by Big Data. Big Data not only have produced many challenges and changed the directions of the development of the hardware, but also in software architectures. For Big Data related applications, data security problems are more stubborn for several reasons. Firstly, the size of Big Data is extremely large, channeling the protection approaches. Secondly, it also leads to much heavier workload of the security.

Data visualization

The aim of data visualization is to represent knowledge more intuitively and effectively by using different graphs [14]. To convey information easily by providing knowledge hidden in the complex and large-scale data sets, both aesthetic form and functionality are necessary. For Big Data applications, it is particularly difficult to conduct data visualization because of the large size and high dimension of Big Data [15]. However, current Big Data visualization tools mostly have poor performances in functionalities, scalability and response time. New framework for modeling uncertainty and characterizing the evolution of the uncertainty information are highly necessary through analytical processes.

The shortage of talent will be a significant barrier in capturing values from Big Data. It usually takes many years to train Big Data analysts that must have intrinsic mathematical abilities and related professional knowledge. The same situation also happens in other nations, no matter developed countries or developing countries around the world. It can be predicted that there will be another hot competition about human resources in Big Data developments. They have enough confidence that we have the capabilities to overcome all the obstacles as new techniques and technologies are developed. There are many critiques and negative opinions from the pessimists. Some researchers think Big Data will lead to the end of theory, and doubt whether it can help us to make better decisions. Whatever, the mainstream perspectives are most positive, so a large number of Big Data techniques and technologies have been developed or under developing [15].

From the above analysis we reach at the point that a lot of technologies have been developed and are yet to be developed. As the volume of data is increasing day by day exponentially and this also can't be ceased due to the ever increasing sources of Big data, so we can say, the researchers will get ample of opportunities to explore more in this emerging area of Big Data in order to counteract the challenges.

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

7. Conclusion and Future Work

As Big data is an emerging research area, there is the potential for making promising research work in many application domains through better analysis of the large volumes of data. This survey paper gives a profound idea on various application domains which belong to primitive and data product based applications, the design of Hadoop architecture with horizontal scaling and the challenges in various phases of big data analytics.

The identified challenges would be explored for a specific application domain using some big data tools to have an implemented proof of the proposed challenges leading towards the design and development of solutions.

References

[1] Rumbold J. M. M., and Pierscionek B.K., 2017, What Are Data? A Categorization of the Data Sensitivity Spectrum, *Big Data Research*.

[2] Honest N., and Patel A., 2016, A Survey of big data analytics, *International Journal of Information Sciences and Techniques*, 6(1/2), pp.35-43.

[3] Zikopoulos P., and Eaton C., 2011, Understanding big data: Analytics for enterprise class hadoop and streaming data (McGraw-Hill Osborne Media).

[4] De Mauro A., Greco M., and Grimaldi M., 2015. What is big data? A consensual definition and a review of key research topics, Proc. AIP Conf., pp.97-104.

[5] Lakshmi C., and Nagendra Kumar V. V., 2016, Survey Paper on Big Data, *International Journal of Advanced Research in Computer Science and Software Engineering*, *6*(8), pp.368-381.

[6] Koti S., and Seeri S.V., 2017, A Survey on Big Data Issues and Challenges, *IOSR Journal of Computer Engineering*, 19(5), pp.75-78.

[7] https://www.qubole.com/blog/big-data-evolution/

[8] Owais S. S., and Hussein N. S., 2016, Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data, *International Journal of Advanced Computer Science and Applications*, *7*(3), pp.254-258.

[9] https://www.simplilearn.com/big-data-applications-in-industries-article

[10] https://data-flair.training/blogs/big-data-applications-various-domains/

[11] Rajeshwari D., 2015, State of the art of big data analytics: A survey, *International Journal of Computer Applications*, 120(22), pp.39-46.

[12] Oussous A., Benjelloun F. Z., Lahcen A. A., and Belfkih S., 2017, Big Data technologies: A survey, *Journal of King Saud University-Computer and Information Sciences*

[13] https://www.dezyre.com/article/hadoop-ecosystem-components-and-its-architecture/114

[14] Acharjya D. P., and Ahmed P. K., 2016, A survey on big data analytics: challenges, open research issues and tools, *International Journal of Advanced Computer Science and Applications*, 7(2), pp.511-518.

[15] Chen C. L. P., and Zhang C. Y., 2014, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, *Information Sciences, 275*, pp.314-347.

