International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 1, Jan-Feb-2020, ISSN (Online): 2395-566X

A Survey on Big Data Analytics: Challenges Of Big Data Integration, Open Research Issues And Tools

GeyaSudha Singanamalla, Sree Devi E
Dept. Master of Computer Applications (MCA)
Sree Vidyanikethan Institute of Management, SVIM, Tirupati, India
sudhasinganamalla@gmail.com, sreedevi.mca15@gmail.com

Abstract **- Data has evolved in the last 5 years like never before. Lots of data is being generated each day in every sector such as social media, E-Commerce, Tech giants Organizations have slowly started realizing that they would be interested in working on all the data. Organizations are interested in doing precise analysis and they want to work on different formats of data such as structured, unstructured and semi structured data. Organization are trying to gain the inside data or to find the hidden treasure in so called Big Data. Now, any Organization would want to have a solution which allows them to store huge amount of data, capture it, process it, analyze it and also look into the data to give more value to the data. So, Big Data analysis is a current area of research and development. The basic objective of this survey paper is to give an idea on the potential impact of big data challenges, open research issues and various tools associated with big data. As a result, this article provides a platform to explore and provides the organizations and the researchers to develop the solution to big data based on the challenges and open research issues.**

Keywords- **Big data analytics, Massive data, Structured data, Unstructured data, Semi structured data.**

I. INTRODUCTION

Big Data is also the data generated from various sources but with a huge size. It is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it and process it efficiently. It provides evolutionary breakthroughs in many organizations and in many fields with collection of large datasets.

Big Data is referred as the datasets which cannot be recognized, obtained, managed, analyzed and processed by present tools. Big Data has been defined in many ways by different analysts of Big Data such as research scholars, data analysts and technical practitioners [1].Apache Hadoop refers Big Data as “Big Data is a dataset which could not be captured, managed, and processed by general computers within an acceptable scope”.
Basically, big data is available in three formats. They are:

- Structured
- Unstructured
- Semi structured

Structured Data: Structured data is defined as the data which can be stored, accessed and processed in a fixed formed. The structured data contains data in the form of tables. Let us have an example of employees in an organization as shown in Table I.
Table I: Structured Data

EMP_ID	E_NAME	DEPT	SALARY
101	Sumana	Finance	65000
102	Geya	Admin	50000
103	Vedha	Bank	60000
104	Sasi	RTO	40000

Unstructured Data: Unstructured data is defined as the data which has no structure and this data is the combination of text, images, videos, audios and graphs. Unfortunately, the organizations are unable to derive the value of it since this data is in raw format or unstructured format.

Semi Structured Data: Semi Structured data is defined as the data which has both Structured and Unstructured formats. The Semi Structured data is in the form of XML data. We can also take the example as the table definition in the Relational Database Management System.
Big Data can be characterized from 3Vs to 4Vs. 3Vs refers to Volume, Velocity and Variety. The fourth V refers the veracity of the data. [2] The following Fig.1. refers the characteristics of Big Data.

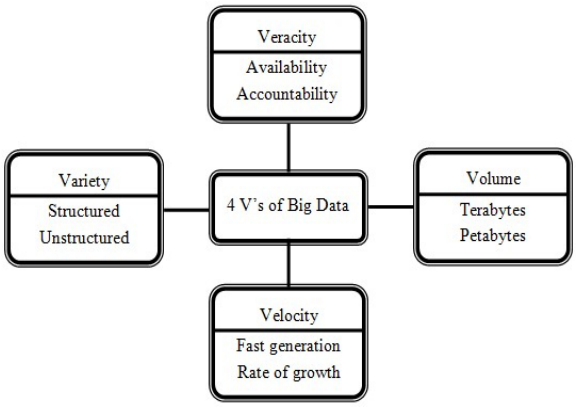

Fig.1. Characteristics of Big Data.

Volume: Volume is one characteristic which we need to be considered while dealing with Big Data. It represents the Size of the data which is enormous.
Velocity: Velocity represents the speed of the generation of the data inorder to reach and process the demands of real potential in the data.
Variety: Variety represents the nature of the data, whether the data is in structured, unstructured or semi structured format. The data may also be in the form of text, images, audio, video, pdf etc.,
Veracity: Veracity is also one characteristic which we need to be considered while dealing with big data. It represents the availability and accountability of the big data.

It is expected that the growth of big data is estimated to reach 40 zettabytes of data by 2020. The market of big data analytics expected to reach $103 billion by 2023. As, internet users are becoming more they generate 2.5 quintillion of data each day. The technology has been increasing big data becomes a vital role in industries, cloud computing, internet of things and social business [3]. Generally, for managing the large datasets Data warehouses have been used.

But extracting the precise knowledge from those datasets became a major issue. Most of the approaches using in these days are not sufficient to deal with the big data. The major problem in the analysis of big data is the lack of coordination between the database system and as well as with the tools such as data mining and statistical analysis. The study on intricacy theory of big data will help to understand the essential characteristics and formation of complex patterns in big data, simplify its representation, gets better knowledge abstraction and to know the computing models and algorithms on big data [4].

There is a need for methodological analysis in elaborating the big data revolution. [5]. Many researchers carried various researches on big data and its trends.[6], [7]. This paper mainly focuses on the challenges of big data and its techniques. Additionally, we also focus on the open research issues and tools. So, to explain these topics, the paper is divided into the following sections. Section 2 deals with the challenges of big data integration, section 3 elaborates about the open research issues that help us to process big data and extract useful knowledge from it. Section 4 deals with the tools and techniques of big data. Finally, the conclusion remarks are provided in Section 5 to summarize outcomes.

II. CHALLENGES IN BIG DATA INTEGRATION

Traditional data processing applications are inadequate for large and complex datasets. So, Big Data plays a major role to deal the complex and large datasets. The integration of such huge datasets is quite complex. During the integration, one has to face several challenges such as analysis, capture, data sharing, data searching, visualization, information privacy and storage. The traditional relational database model was not efficient in order to handle the data.

The key elements of the big data platform are to handle the data in new ways as compared to the traditional relational database [8]. The big data platform will provide more accuracy in managing the data and leads to more confident in decision making. In this paper we can discuss integration of big data and the challenges that can be faced during the big data process.
Six Challenges in the integration of Big Data:

The integration and handling of Big Data is very complex. So, one has to face the six challenges during its integration such as uncertainty in managing the data, talent gap in big data, getting raw data into big data structure, concur across data sources, getting useful information out of the big data, volume, skill, availability, solution cost etc., [9].

1. Uncertainty in managing the data

The conventional way of big data management is the use of a wide range of innovative data management tools and frameworks which are dedicated to support operational and analytical processing. The abnormal way of managing the data is by using the NoSQL (not only SQL). The NoSQL frameworks are used to differentiate the traditional relational database management systems which are designed to increase performance demands of big data applications such as managing a large amount of data and quick response times. There are different approaches of NoSQL to deal the big data are hierarchical object representation (JSON, XML and BSON) and the concept of key value storage. The wide range of NoSQL tools, developers and the status of the market are creating uncertainty with the data management.

2. Big Data Talent Gap

The traditional approach was not suitable to gain the respect from media and the analysts with the content on the analysis of big data. So, the new tools had been evolved in this sector which can range from traditional approach. The traditional relational database tools with some alternative data layouts designed to maximize access speed of the analysis of the data. Apart from the big data management aspects, the typical expert has also gained the experience through tool implementation and its use as a programming model.

3. Extracting data into Big Data structure

The intent of managing the big data contains analyzing and processing a massive amount of data. Most of the people have raised their expectations and considers analyzing the huge data sets for big data platform without getting any awareness about the complexity behind the transmission, accessing, and delivery of the data. They get the data from different resources and then load the data into the big data platform. The detailed aspects of transmitting the data, accessing the data and loading the data are only the part of the challenge. But, the conventional relational data set of transforming and extracting the data to big data platform is limited.

4. Synchronization across Data Sources

While importing the data into big data platform we may come to a conclusion that the data copies are migrated. Such data copies from a wide range of sources on different rates and schedules can rapidly get out of the synchronization with the originating system. Therefore, the data coming from one source is not out of date as compared to the data coming from another source. The classical data management and data warehouses, the sequence of data transformation, extraction and migrations all arise the situation in which there are risks for data to become unsynchronized.

5. Extracting Information from data in Big Data Integration

The use cases for Big Data practically involves the availability of data, managing the existing storage of the data and allowing the end-users to access the data, employing the business intelligence tools for the purpose of the discovery of the data. The business intelligence tools connect with different big data platforms and provide the transparency to the data consumers in order to eliminate the custom coding. If the number of data consumers increases, then one can provide a need to support an increasing collection of simultaneous user accesses. The increment of demand will create the problem to different aspects of business process cycles and becomes a challenge in big data integration to ensure the right-time data availability to the data consumers.

6. Miscellaneous Challenges

There are some other challenges that occur while integrating the Big Data. Some of them are integration of data, skill availability, solution cost, the volume of data, the speed of data, veracity and validity of the data. It is also a big challenge of big data to process the large volume of data at a reasonable speed so that the information may available to the data consumers whenever they need it. The validation of data set is also fulfilled while transferring the data from one source to another as well as the data consumers.These are the challenges one can face and one must be considered and should be taken care of while integrating the Big Data or managing any big data platform.

III. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science are becoming the research major point in industries and academia. The main aim of Data science is researching the big data and knowledge extraction from data. The applications of data science and big data include information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing. Effective integration of analysis and technologies will result in predicting the future drift of events. The main focus of this section is to discuss about the open research issues of big data analytics. The issues pertaining to big data analysis are classified into three categories. They are Internet of things (IoT), cloud computing, bio inspired computing and quantum computing. However, it is not limited to these issues. More research issues related to health care big data can be found in 'HusingKuo et al'. paper [9].

1. IoT for Big Data Analytics

The art of businesses, cultural revolutions, global interrelations and an unbelievable number of personal characteristics are reconstructed by the Internet. Now-a-days, machines are controlling the innumerable autonomous gadgets via internet and create Internet of Things (IoT). Therefore, the appliances are becoming the user of internet, just like humans with the web browsers.

As, Internet of Things has an imperative societal and economic impact for the future construction of information, network and communication technology, it is attracting the attention of recent researchers for its most promising opportunities and challenges. The new regulation of future will be eventually, everything will be conducted and intelligently controlled. The concept of IoT is becoming more popular to the realistic world due to the development of mobile devices, embedded and more communication technologies, cloud computing, and data analytics. IoT nonce challenges in combinations of volume, velocity and variety. In boarder sense, similar to

the internet, Internet of Things enables the devices to exist in a myriad of places and facilities applications ranging from trivial to the crucial. Several assorted technologies such as big data and computational intelligence can be incorporated together to improve the data management and knowledge discovery of large-scale automation applications. Much research in this concept has been carried out by Mishra, Lin and Chang.

The biggest challenge that big data professional are facing is the knowledge acquisition from IoT data. It plays a major role to develop infrastructure and to analyze the IoT data. An IoT device generates continuous streams of data. By using machine learning techniques, the researchers can develop tools to extract meaningful information from these data. Understanding of generating continuous streams of data from IoT devices and analyzing those data to get meaningful information became a big challenging issue and it leads to big data analytics. The solution to handle big data from IoT devices is machine learning algorithms and the computational intelligence techniques. Decisive technologies that are associated with IoT are also discussed in many research papers [10]. Fig.2. shows an overview of IoT big data and Knowledge discovery process.

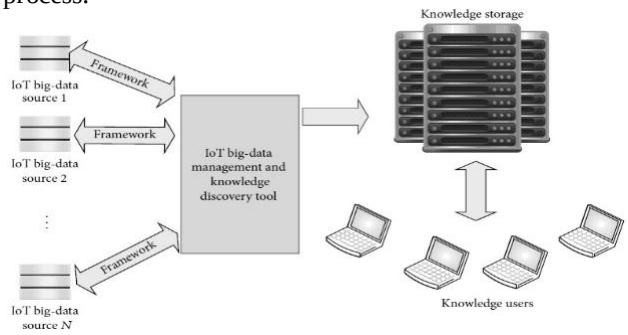

Fig.2. IoT Big Data Knowledge Discovery.

2. Cloud Computing for Big Data Analytics

Virtualization technologies have developed as supercomputing, more accessible and affordable. With the flexibility of specification details such as number of processors, disk space, memory and operating system, the computing infrastructures that are hidden in virtualization software make systems to behave like a true computer. The use of these virtual computers is known as cloud computing which has been a robust big data technique.

Cloud computing technologies and big data have been developed with the importance of developing a scalable and on demand availability of resources and data. Cloud computing computes large amount of data by on demand access to configurable computing resources through virtualization techniques. Cloud computing provides the benefits to the cloud consumers by offering the resources according to their demands and also provide the facility on rent basis. Cloud computing helps to develop the business models for all varieties of applications with infrastructure and tools. Research issues and open challenges of cloud computing and big data are discussed with the detailed process by many researchers which highlights the challenges in data management, data variety and velocity, data storage, information processing and resource management [11], [12].

3. Bio-inspired Computing for Big Data Analytics

A technique which is inspired by nature to address complex real-world problems is Bio-inspired computing. Biological systems are syllogized without a central control. The optimal data service solution is found on considering cost of data management and service maintenance by a search of bio-inspired cost minimization mechanism. These techniques are developed by biological molecules such as DNA and proteins for conducting computational calculations which involves storing, retrieving, and processing of data.

Bio-inspired computing techniques serve as a key role in data analysis and its applications in big data. Because of the optimization applications, these algorithms help in performing data mining for large datasets. The most advantage is its simplicity and their rapid convergence to optimal solution while solving service provision problems [13]. Some of the applications of bio inspired computing were discussed in detail by Cheng et al. From the discussions, we can observe that the bio inspired computing models provide smarter interactions, inevitable data losses and help in handling ambiguities. Hence, it is concluded that in future bio-inspired computing may help in handling big data to a large extent.

4. Quantum Computing for Big Data Analytics

A computer which has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously is called a Quantum Computer. A real quantum computer can solve the problems that are exceptionally difficult to solve with the recent computers, of course today's big data problems. In traditional computers, the data is stored in the form of long strings of bits which encode either a zero or a one.

Diversely, a quantum computer uses quantum bits or qubits. The main difference between qubit and bit is that a qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states. Therefore, it can be capitalized on the phenomenon of super position and entanglement because qubits behaves quantum. Many Big Data problems can be solved much faster by larger scale quantum computers when compared with classical computers. Hence, it is a challenge for this generation to build a quantum computer and facilitate quantum computing to solve big data problems.

IV. TOOLS FOR BIG DATA PROCESSING

To achieve the competitive edge in the market almost every organization extensively handles big data. The open source big data tools for data analysis and data processing are the most useful choice of organizations in order to consider the cost and other benefits. The top open source project and big data bandwagon roller in the industry is Hadoop. There are plenty of other vendors who follow the open source path of Hadoop. Big data tools are mainly used for finding the large data sets, what type of analysis we are going to perform on the data sets, what is the expected output etc.,

Big data open source tools are categorized into on the data stores, development platforms, development tools, analytics and reporting tools. Organizations are rapidly developing the new solutions to achieve the competitive advantage in big data market, we need to concentrate more on big data tools which are driving the big data industry. For example, the two big data platforms which supports interactive analysis of big data are Dremel and Apache Drill. An astonishing list big data techniques and tools are also discussed by many researchers [6], [14]. The system of big data project by Huang et al is clearly elaborated in the Fig.3.

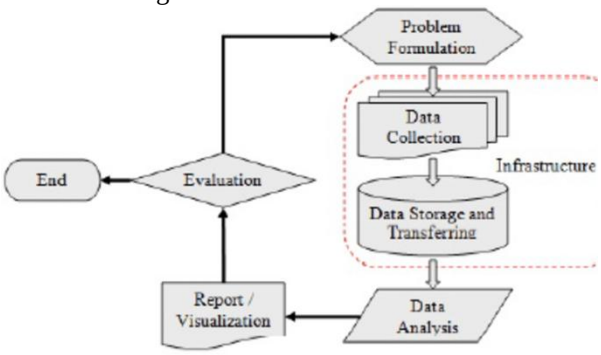

Fig.3. Work flow of Big Data Project.

1. **Apache Hadoop-** For the large-scale processing of Big Data, Apache Hadoop is the most suitable and useful open source tool in big data industry with its colossal capability. Apache Hadoop is the 100% frame work which runs on commodity hardware in the existing data center. It can also run on cloud infrastructure. The four main parts in Hadoop is the following.

1.1 Hadoop Distributed File System- HDFS is adistributed file system compatible with very high scale bandwidth.

1.2 Map Reduce- Map Reduce is a programming model which is used for processing Big Data.

1.3 YARN- Yarn is a platform which is used for scheduling and managing the Hadoop's resources in Hadoop Infrastructure.

1.4 Libraries- Libraries are used to help other modules to work with Hadoop.

2. Apache Spark

The alternative of Apache Hadoop is Apache Spark. It is the successor of Hadoop in many aspects. To address the shortcomings of Hadoop Spark was used. The batch data and real time data is processed by spark and it can operate 100 times faster than MapReduce. Spark also provides the capabilities of in-memory data processing. Spark also works with HDFS, Open Stack and Apache Cassandra both in cloud and on-prem, adding another layer of versatility to big data operations for your business.

3. Apache Storm

A real time frame work for data stream processing that supports any programming language is Storm. It is the apache product which balances the work load between multiple nodes based on topology configuration and works well with Hadoop HDFS. The benefits of Apache storm are the following:

- Built-in-fault-tolerance
- Great horizontal scalability
- Clojure-written
- Auto-restart on crashes
- Output files are in JSON format
- Works with Direct Acyclic Graph (DAG) topology.

4. Apache Cassandra

One of the pillars in Face Books massive data is Apache Cassandra. It is used to process structured data sets distributed among huge number of nodes across the globe. The unique capabilities of Apache Cassandra are as follows:

- Simplicity of operations due to simple query language used.
- Great linear scalability
- High fault tolerance
- Constant replication across nodes
- Built-in-high-availability
- Simple removal and adding of nodes from a running cluster.

5. Mongo DB

The open source NoSQL database with rich features, which is compatible with many programming languages, is MongoDB. The most important MongoDB features are the following:

- Cloud-native deployment and great flexibility of configuration
- Significant cost savings, as dynamic schemes enable data processing on the go.
- Data partitioning across multiple nodes and data centers.

6. R Programming

For enabling the wide scale statistical analysis and data visualization R is used along with JuPyteR stack (Julia, Python, and R). The major benefits while using R are as follows:

- R is highly portable

- R supports Apache Hadoop and Spark
- R can run inside the SQL
- R easily changes from a single test machine to vast Hadoop data lakes
- R runs on both Windows and Linux servers.

V. CONCLUSION

Now-a-days, data are generated as a dramatic footstep. To analyze those data is a big challenge for a common man. In this paper, we made a survey on various research issues of big data challenges, data integration, and tools used to analyze those data. With this survey, we understood that every platform of big data has its own focus. Some of them are used for real-time processing and some other are used for batch processing. The various types of techniques used for big data analytics include machine learning, data mining, statistical analysis, intelligent analysis, cloud computing, quantum computing and data stream processing. From this survey, we believe that in future researchers will spend more time and attention to these techniques to solve big data problems efficiently and effectively.

REFERENCES

[1]. M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.
[2]. A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2015), pp.137-144.
[3]. C. Lynch, Big data: How do your data grow?, Nature, 455 (2008), pp.28-29.
[4]. X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, Big Data Research, 2(2) (2015), pp.59-64.
[5]. R. Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1) (2014), pp.1-12.
[6]. C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275 (2014), pp.314-347.
[7]. K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.
[8]. Dong, X.L. and Srivastava, D., 2013, April. Big data integration. In 2013 IEEE 29th international conference on data engineering (ICDE) (pp. 1245-1248). IEEE.
[9]. Chen J, Chen Y, Du X, Li C, Lu J, Zhao S, Zhou X. Big data challenge: a data management perspective. Frontiers of Computer Science. 2013 Apr 1;7(2):157-64.
[10]. X.Y. Chen and Z. G.Jin, Research on key technology and applications for internet of things, Physics Procedia, 33, (2012), pp. 561-566.
[11]. M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R.Buyya, Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79 (2015), pp.3-15.
[12]. I. A. T. Hashem, I. Yaqoob, N. BadrulAnuar, S. Mokhtar, A. Gani and S. Ullah Khan, The rise of big data on cloud computing: Review and open research issues, Information Systems, 47 (2014), pp. 98-115.
[13]. L. Wang and J. Shen, Bioinspired cost-effective access to big data, International Symposium for Next Generation Infrastructure, 2013, pp.17.
[14]. M. Herland, T. M. Khoshgoftaar and R. Wald, A review of data mining using big data in health informatics, Journal of Big Data, 1(2) (2014),pp. 1-35.
[15]. Z. Pawlak, Rough sets, International Journal of Computer Information Science, 11 (1982), pp.341-356.