|| Volume 2 || Issue 12 || JULY 2017 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

# REVIEW ON BIG DATA ANALYTICS

**Vinaya Keskar[1], Dr. Ajay Kumar[2]**

*Assistant Professor, ATSS's College of Business Studies and Computer Applications, Pune.* [1]
*Director, JSPM's Jaywant Technical Campus, Pune*[2]
[1]*vasanti.keskar@gmail.com*
[2]*ajay19_61@rediffmail.com*

*Abstract*— **It must be simpler for the decision-maker to access vast volumes of accessible data sets in the information age. The use of standard methods and techniques is highly demanding because big data applies to large and high-performance datasets. The essential and critical datasets need to be studied and valued because of the subsequent growth in these broad files, handling and extraction solutions. Besides, decision-makers must collect valuable information from this diversified information, which continuously varies from regular sales to customer experiences. Only using data analysis can these useful insights be provided. The implementation of Big Data methodologies is incremental. This paper aims to analyse a few of the various approaches and existing tools available for Big Data, the advantages offered by the use of data processing in different areas of comparison, and the potential of predictive analytics in those areas.**

***Keywords:*** *Data Analysis, Big Data, Architecture, High Volume Data, Unstructured Data, Semi-Structured Data*

## I INTRODUCTION

In today's age, web-based applications are used by people and systems that exponentially produce extensive data. The measurement units for data size are exabyte (EB) and petabytes (PB). This increase is attributable to advances in connectivity, digital sensors, computations and business analytics data storage. A scholar, Roger Magoulas, invented the Big Data concept. Data have expanded massively in many fields in recent decades as well as various forms of data. According to the International Data Corporation (IDC) statistics, the global data volume generated in the globe is 1.8 ZB in 2011. Over the next five years, this amount will be about nine times greater. Information and data will become the fuel for the 21st century, with vast quantities of data and information produced and used in all feasible areas such as healthcare, advertising, epidemic prevention and management, smart cities and business intelligence. When we equate Big Data principles with other conventional databases and their processes, Big data contains semi-structured and unstructured data. Big data technologies and analysis methods represent large data sets from different systems in real-time. Big data are also helpful in gaining knowledge on potential opportunities to assess new values, grasp the hidden values, and exponentially organize and exploit big data sets in real-time.

Data and information volumes are rising increasingly and are also presenting challenges. A significant challenge in big data analysis is the big data visualization process. Increasingly, data collection from the decentralized data sources has been generated daily by developments in information technology. Other emerging technology, such as cloud computing, IoT and data centres, will also support data development. The standard for storing and extracting data from the big data resources is cloud computing technology. IoT technology for the processing and transmission of data to be collected and analyzed in cloud storage is relying on sensors.

## II BIG DATA ANALYTICS

The common word "Big Data" has long been implemented to databases growing in size and challenging to use traditional database administration systems. Those data sets exceed the capacity in an acceptable period to capture, collect, maintain and process data in commonly accessible computer devices and storage facilities.

Big data ranges from just a few terabytes (TB) to a few hundred petabytes (PB) of vast data in a single location are increasing dramatically. Nevertheless, the capturing, save, search, share, record and evaluate are some significant challenges. Today, organizations analyze vast volumes of data that are too organized to disclose the data they did not learn before.

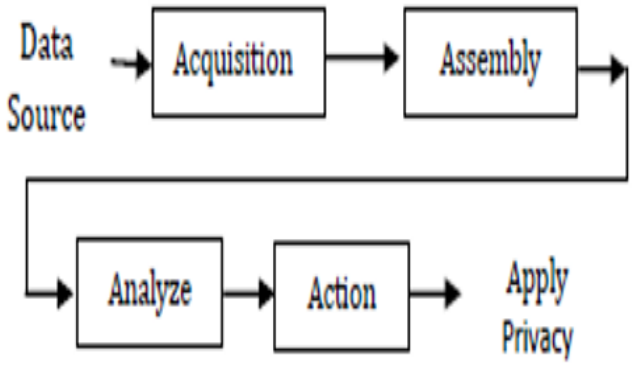



***Figure 1.1. Steps of Data Analysis***

Consequently, big data is the location of advanced processing on Big Data Sets. The more significant, however, the more difficult it is to manage the data collection. We will begin with an overview of the attributes and value of big data in this section. Business gains are usually accomplished by the processing of detailed and too complicated data requiring real-time technology and different criteria for advanced data systems, computing software and tactics.

### 2.1. Taxonomy of Big Data

The Big Data was listed in six groups: semantic, numerical, storage, big data management, big data mining, big machine learning and security & privacy, according to figure 1. The big data is classified into three significant categories: The extensive data storage framework is divided into four groups, i.e. storage architecture, storage implementation, storage structure and storage devices.

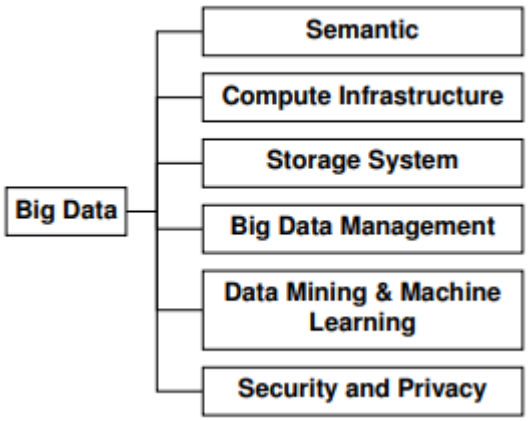
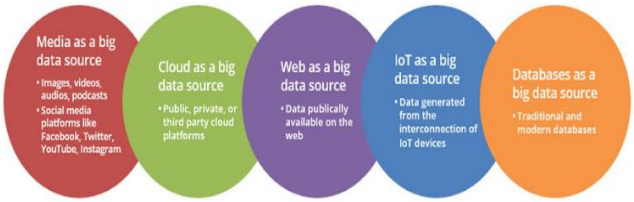



Figure 1.2. Taxonomy of Big Data Technology

### 2.3. Sources of Big Data




**1. Big data source media**

The Media offers essential ideas into customer data preferences and evolving patterns as the most common source of data. Offer companies a simple description of their intended groups, draw trends and observations and improve decision-making because they are self-divulging and cross all physical and demographic hurdles. Media involves public media and interactive platforms such as Google, Facebook, Twitter, YouTube, Instagram, and generic media such as photos, videos, audios, and podcasts, which offer a quantity of information on every dimension of interactions between individuals.

**2. Cloud as a Big Data Source**

By moving their data onto the cloud, businesses today are advancing conventional data sources. Cloud storage adapts organised and unstructured data and delivers customer knowledge and on-demand perspectives in real-time. Cloud computing has its simplicity and scalability as a crucial factor. The cloud offers an economically feasible data source via networks and servers, as big data can be stored and collected in public or private clouds.

**3. The Web as a Big Data Source**

Big data is widely available and conveniently accessible via the public web. Site or 'Internet' data is widely accessible both to individuals and businesses. Besides, web services like Wikipedia provide us with free and fast information. The website's colossal scale guarantees its varied usability. It is beneficial for start-ups and SMEs because they don't have to wait until they can build their own Big Data infrastructure.

**4. IOT AS A SOURCE OF BIG DATA**

A valuable source of Big Data is content produced by machines or data provided by IoT. This information is produced generally through electronic device-connected sensors. The capability to provide reliable, real-time information depends on the ability of the sensors. In addition to computers and smartphones, IoT is now gaining momentum and includes extensive data generated from any device which may emit data. Data from medical equipment, car processes, video games, metres, cameras, home appliances and the like can now be accessed from IoT.

**5. Databases as a Big Data Source**

Today, businesses tend to use conventional and digital databases to access the appropriate big data. This combination leads to creating a hybrid data model, requiring little investments and IT infrastructure costs. Besides, these databases are used for many objectives in the area of business intelligence. These databases will then yield insights to boost the profits of the organisation. Data sources, including MS Access, DB2, Oracle, SQL, Amazon Quick, etc., are common in popular databases.

It is challenging and can be stressful and time-consuming to collect and interpret data from vast big data sources. These complexities can be overcome when organisations consider all required big data issues, consider the appropriate data sources and use them well for their organisational objectives.

## III. 3 V'S OF BIG DATA

Innovative technology and strategies are also required to help people and institutions integrate, interpret, display, and consume big data's rising flood. We all have learned about the 3V's of large data volumes, variation and speed, but also about other V's, with which IT, businesses and data scientists, particularly Big Data Veracity, have to deal.

- **Data value**: The volume of data determines an organisation's quantity of data, and does not have to own it as long as it can access it. As the amount of information rate is proportional to age, sort, wealth, and quantity, different data records' value will decline.
- **Variety of data**: Data is a measurement of the wealth of data representation – text, video images, audio, etc. It is probably the most significant barrier in analytical terms to the successful use of large quantities of data. Incompatible data formats, non-aligned data structures, and incoherent data semantics are significant problems that could lead to analytical expansion.
- **Data speed**: Data speed tests data generation speed, streaming speed and aggregation. The speed and wealth of data used during various business transactions have grown significantly. The control of data speed is more than a problem of bandwidth. It is also an ingesting problem.
- **Data truthfulness**: the validity of the data applies to data bias, noise and anomalies. Is the data stored and collected important to the analysis of the problem? The most challenging problem compared with such aspects as volume and speed is the validity of data analysis. Big data means the processes in which massive data sets ("big data") are obtained, organised and analysed to identify trends and other valuable information. Big data analytics will only

- allow you to understand the data and help you classify the most relevant data for business and strategic business decisions. Big data analysts need information from the data analysis.

## IV. BIG DATA TECHNOLOGIES FOR DATA ANALYTICS

Big data management covers structured data, semi structured and unstructured data organisation and management of large volumes. Large-scale data management's key objective is to cover three essential elements: high-level data quality, data availability for business intelligence, and large-scale data analysis. Different methods from data collection to data storage to visualisation are available for big data management.

**1. Hadoop**

This is an open-sourced forum for the management and study of Big Data. Hadoop offers an infrastructure that is simple to use and versatile to work with various data sources. This framework can perform tasks such as collecting different sources of information or accessing information from the database to run process-based machine learning processes. This tool also offers numerous applications such as weather location, traffic sensors and social media info.

**2. Map Reduce**

This environment allows for improved usability of work implementation against a server party. Map Reduce's implementation has two main tasks: (a) The map task: it transforms the data set into a different set of value pairs. (b) Reduce task: combines a variety of Map work outcomes with decreased tuples.

**3. PIG**

It is a tool to evaluate Hadoop closer to developers and business users. PIG allows the processing of queries via data stored on a Hadoop from SQL.

**4. Wibi Data**

This platform is designed to customise user experiences by companies. It's a mix of Hadoop and web analytics. It acts as a top layer over the HBase and helps the websites to explore better user data processes. It enables users to respond to their content and decision in real-time, such as suggestions and data.

**5. Hive**

Hive allows predictable business apps to operate a Hadoop cluster for SQL queries. It offers a Facebook-like SQL-like bridge and then it is open source. Hive is a high-level Hadoop perception which enables everyone to quickly query data from the storage of Hadoop data.

**6. Rapidminer**

It offers a built-in platform for various items such as machine learning, data/text extraction and other data analytics functions, such as prediction and market analysis. Rapidminer is used for software and business applications and training, testing, training, development of applications and quick prototyping. It promotes all steps in data mining, including the planning, validation, viewing and optimisation of datasets.

## V. CHATATERISTICS OF BIG DATA

Doug Laney, an analyst for the Gartner Company, listed 3 'V's - Variety, pace and volume in Big Data in 2001. These characteristics are adequate to know what big data are, isolated enough.

Its Big Data name itself is linked to an immense scale. Size is one of the main elements to determine the data value. The fact that the data can be calculated in Big Data is also dependent on the amount of the data. 'Size' is also another attribute which must be considered.

Variability refers to both heterogeneous and organised sources and their presence. Excel sheets and the files of corporations were the only data points used during the early days, which were considered by most applications. Data are often considered in analytical applications in different ways, including email transfers, images were taken, videos captured, monitoring systems and mp3 formats. Speed essentially refers to the rate at which data are produced in real-time. In a broader sense, the transition speed is required; incoming data sets are linked at different speeds and operational bursts.

**1. Tools and Methods Used**

The need for faster and more efficient ways to process this data has evolved with technological advancements and increasing data flowing in and out of organisations daily. Stacks of knowledge are no longer adequate to take successful decisions in good time. The data sets can no longer be conveniently evaluated using traditional data collection and analysis techniques and infrastructures. The need for emerging technology and approaches in this field, parallel with this, is the infrastructure required to store and processes these data sets. With the emergence of comprehensive data, the method also incorporates techniques into and approaches to decision making, from knowledge collection to interpretation, leading to the conclusion of B-DAD decisions. The system structures the various methods for storing, handling and handling big data, data analysis techniques and processes, mapping and estimating the various data analytics changes in three main aspects: the data and analysis processing, big data and architecture, and the regular Big Data analysis. All areas will be further explored in this portion. As big data continues to grow as such an important research area and current innovations and techniques are continuously introduced, this section does not include all possibilities. It is based on a coherent concept, rather than a description of all the possible possibilities technologies.

**2. Management and Data Storage**

In the face of vast volumes of data, companies must deal with one of the first things where or how these data are obtained. Historically, data marts, relation databases, and warehouses provide organized data collection and extraction approaches. Extract, Convert and Load data is transmitted from operational data stores into storage software to access external sources to meet technical needs and then store and store information. Thus, before data collection and advanced analytical operations are available, the data is processed, updated and recorded.

However, big data environments need skills to solve models like Agile, Magnetic, Deep (MAD), which differ from those in an (EDW) context. Second, traditional approaches for EDW preclude

the introduction of new data sources before cleaning up and incorporation. Big data systems would inevitably take complex reactions due to data's iniquity and pull all sorts of data, whatever the standard. Furthermore, the growing number of source numbers of diverse data and the sophistication of such data analyses will allow analysts to produce the data quickly and modify it, given a large amount of data storage. This ultimately comprises an agile infrastructure which is synchronized with rapid system creation in physical and logical content. Finally, considering the complex statistical methods used in the current data analyses and that experts must still explore massive data sets by finding up or down, an extensive database also needs in-depth data and end up as an advanced runtime algorithm.

Too many methods have been used to provide high query efficiency. The application scalability of memory base for the large data ranges from massive parallel processing (MPP) and distributed network databases.

Databases including Not Only SQL have been built to store and management unorganized or non-relational data. In comparison to database systems, No SQL databases vary in data processing and storage. These databases rely mainly on data storage, which is scalable and needs to be implemented on the application layer rather than written in various languages in databases.

On the other side of the coin, the secret in-memory repositories manage the useful data in storage memory, delete disc input and output, and allow real-time database responses. The whole database can be integrated into silicone-based main memory, rather than using mechanical hard discs. Furthermore, in-memory databases are currently being used for advanced research on large volumes of knowledge to speed up the entry and scoring of expository. This offers adaptability to enormous knowledge and speed for disclosure. Alternatively, the basis for success in Big Data Analytics is implementing the MapReduce model mentioned in the next section, which provides consistency and reliability. In distributed blocks, the information is recorded in many data nodes, as is the name node. The name node serves as the supervisor between the data node and the client, directing the client to the actual data node that contains the necessary data.

### 3. Processing of Big Data

After massive data storage, computational processing occurs. The big data processing requires essential criteria under the last portion. Fast loading of data is the first requirement. However, as file and network or Internet traffic correlate mainly with query results, the loading time for data has decreased during the data preparation. The second condition is that queries be processed quickly. In the reaction time, most queries are critical for fulfilling the high workload and real-time demand requirements. This allows the data structure to sustain high query speeds, even as the number of queries is increasingly rising.

Map Reduce is a tool for programming and Java-based software configuration that uses distributed computing. Two essential tasks are included in the MapReduce algorithm: map and reduction. Map generates a data set and transforms it into any other data set where unique features are separated into tuples.m second, reduce the task of extracting functions from the map and converting these other unique knowledge lists into a tinier subset. The reduction task is often performed after the map task, as indicated by the name series MapReduce.

The MapReduce functionality is based on two different nodes within Hadoop: the Work or the Job Tracker, and the Tracker Task nodes. The Work Tracker node is responsible for providing the functions, such as the map function, and the reducer functions respectively available task tracked for monitoring the results.

## VI. BIG DATA CHALLENGES

Many crucial challenges need to be focused while handling Big Data and its analytical process.

Some of those challenges are being discussed:

**1. Storage Related:**

As we know, the size of a secondary storage unit is in the Terabytes range (TB). There is also an enormous amount of data generated through the internet. Exabyte (EB) is weighed such that conventional relational database management applications like Oracle and My SQL cannot store or process such big data. To address this, databases are used to manage unstructured and semi-structured data using No SQL-based databases, including Cassandra and Mongo DB.

**2. Management Data Life Cycle**

This method specifies the collection of the data used for the storage process. The current storage system cannot accommodate so large several data, and one of these is difficulties. There are several. A useful model is, therefore required, which improves the framework for life cycle management.

**3. Data representation**

The main objective of data representation is to improve the criticality of data processing and user analysis. Many datasets have certain systemic, semantics, form, organizational, granular and accessible levels. The improper technique for data representation can reduce the importance of data originality and even interrupt an efficient process of data analysis [23]. For an easy analysis method, therefore, efficient data representation is highly needed.

**4. Redundancy Reduction and Data Compression**

Redundancy is one of the big database system and research issues. Redundancy and data compression are organizational approaches for cost savings by data redundancy and compression.

**5. Analysis**

Big data are generated from different types of online activities and transactions that differ in structure, and in the case of high volume data, the analysis is complicated. To manage this situation in a distributed environment, a single, scalable architecture is used. Data is scattered to pieces and processed across several computer systems in the network, which then incorporates processed data.

## VII. CONCLUSION

During this study, we discussed the revolutionary topics of big data that have become very much involved in interpreting possibilities and advantages beyond precedent. Weighty differences are created globally via high-speed data in the era of information we live in.

However, the details and trends of the hidden patterns derived and used are inherent in them. Big data analytics can also make the most market change and optimize decision-making by using different statistical approaches for Big Data and presenting strong information and useful knowledge. In the age of data overspill, we think Big Data Analytics is essential and gives unanticipated insights. This helps decision-makers in various fields. Big data analytics can provide a basis for technological, mathematical and humanistic developments when successfully utilized and applied.

**REFERENCES:**

1. H.V. Jagadish, D. Agarwal, P.Bernstein, Challenges and Opportunities in Big Data, The Community Research Association, 2015.
2. K. Krishnan, Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence, Elsevier Science, 2013.
3. Vallabh Dhoot, Shubham Gawande, Pooja Kanawade and Akanksha Lekhwani, Efficient Dimensionality Reduction for Big Data Using Clustering Technique, Imperial Journal of Interdisciplinary Research (IJIR),Vol-2, Issue-5, 2016, ISSN: 2454-1362 3.
4. Gantz J, Reinsel D, Extracting value from chaos.IDCi View, 2011, pp 1–12
5. Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A survey, Mobile New Applications 2014, 171-209.
6. Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Quart.2012; 36(4):1165–88.
7. K. Krishnan, Data warehousing in the age of big data, in the Morgan Kaufmano series on Business Intelligence, Elsevier Science, 2013.
8. Kitchin R. The real-time city? Big data and smart urbanism. Geo J. 2014, 79(1), pp: 1–14.
9. Katrina Sin and Loganathan Muthu, Applications of big data in education data mining and learning analytics –A literature Review, ICTACT Journal on soft computing special issue on soft computing models for big data, July 2015, Vol:05, Iss: 04, pp: 1035-1049
10. Mayer-Schonberger V, Cukier K, Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.
11. Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A Survey, Computer Science Review, 2015, Vol: 17, pp: 71-80
12. K. Davis, D. Patterson, "Ethics of Big Data: Balancing Risk and innovation", O'Reilly Media, 2012.
13. Mike Barlow, Real-Time Big Data Analytics: Emerging Architecture, ISBN: 978-1-449-36421-2, 2013
14. Shuhui Jiang, Xueming Qian, Tao Mei, Yun Fu, Personalized Travel Sequence recommendation on Multisource Big Social Media, 2016, IEEE Transactions on Big Data,Vol.2, Issue:1 2.
15. http://www.techrepublic.com/blog/big-dataanalytics/10-emerging-technologies-for-big-data