International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 10 | Oct 2019 www.irjet.net p-ISSN: 2395-0072

A Review on the Importance, Tools, Research Area and Issues in Big Data

Mr.Gajanan Babhulkar[1], Deepali D. Rane[2]

[1,2]Assistant Professor, IT, D.Y. Patil College of Engineering, Pune, Mahararashtra, India

---***---

Abstract – *A huge amount of petabytes of data is producing each day from smart devices (i.e. IoT), social media, cloud Computing etc. To extract the knowledge from this repository requires a lot of efforts which is time consuming. The output come from this is a data which is very useful for the business. Therefore, Big Data is become popular in the field of research. The basic aim of this paper is to examine the importance, challenges, tools and research areas for big data. This paper evaluates and explores a new era for research activities.*

Key Words: Big Data Analytics, Tools, Challenges, Research Areas.

1. INTRODUCTION

IoT, Social media, cloud computing, Web services and databases are the major source of digital data generation and led to growth of big data. Big Data is a complex because it is the collection of large and variety of data sets. The traditional tools are unable to process it efficiently. Data is available in structured format, unstructured format and semi structured format in petabytes and beyond. There are ten V's which makes data as a big data. These are Volume, Velocity, Variety, Value, Veracity, Variability, Validity, Vulnerability, Volatility and Visualization. Volume refers to the huge amounts of data generated from various sources. Velocity refers to the speed at which huge amounts of data are being generated, collected and analyzed. Variety refers the different types of data. Value refers the worth of the data being extracted. Veracity is the quality or trustworthiness of the data. Variability indicates dynamic behavior of the data. Validity refers to how accurate and correct the data is for its intended use. Visualization tactics include applications that can display real-time changes and more illustrative graphics, thus going beyond pie, bar and other charts. Volatility refers to how long is data valid and how long should it be stored. Vulnerability indicates the loophole to hamper security.

The figure-1 refers to the definition of big data. However exact definition of big data is not defined but all these 10 V's defined in big data. This will help us to get and obtain enhanced decision making, discover and optimize while being innovative and cost – effective.

In 2020, at a CAGR of 10%, it will leap beyond $76 billion. While IDC is far more optimistic; it projects that revenues from big data and business analytics will surpass $210 billion at a CAGR of 11.9% in two years. Hence, according to Grand View Research, the big data market will boast a size of $123.2 billion by 2025[1]. The market for Big Data is estimated to cross $100 billion by the year 2020, which is now roughly at $25 billion. Again, Big Data analytics solutions have no indications of slowing down and are in much demand. It is expected to touch the $40.6 billion mark by the year 2023. Besides, the technology will have a significant effect on the world economy together with the industrial internet improving global GDP by $10 to $15 trillion in the next 10 to 15 years. Therefore, you can see that the aggressive growth of Big Data solutions will continue in the future. [2] We can say that, the big Data has a bright future in case of economy, accuracy and job market. It is the booster for next generation of Information Technology.

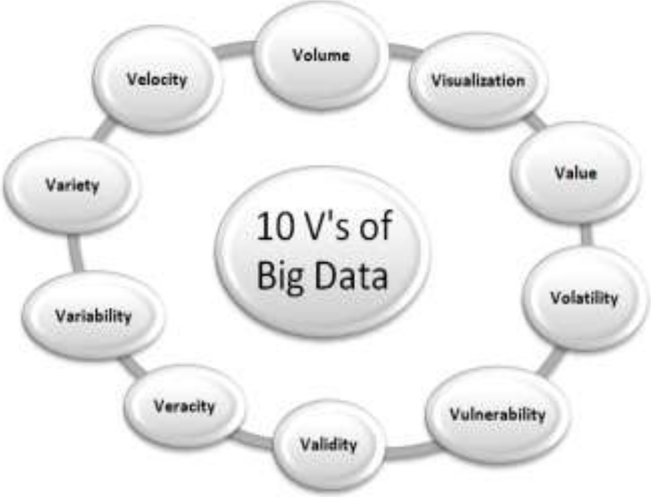

Fig -1: 10 V's of Big Data

Generally Data Warehouse have been used to manage the large datasets .In this case extracting the precise knowledge from the available big data with accuracy, integrity and proper use of tools is a foremost issue. The key problem in the analysis of the big data is the lack of coordination between database system and analysis tool. The study on complexity theory of big data will help to understand essential characteristics and formation of complex pattern in big data, gets better and accurate knowledge abstraction [3]. However, it is observed that all the data available in the form of big data are not useful for analysis or decision making process.

1.1 Challenges in Big Data

Now a day's Big Data has been accumulated in healthcare, manufacturing, media & entertainment, IoT and government sectors [5]. The social media is the biggest source to use Big Data. There is a continuous increase in the number of people interacting with brands on social media platforms. This makes it essential for you to be data savvy to remain competitive and stay relevant in the social media landscape. Seeing the massive amount of data produced by these platforms, it becomes crucial that you make use of big data in your social media marketing campaigns.

Big data will allow you to analyze the behavior of buyers and target an exact group of people. By giving you in-depth insights, it will assist you in fine-tuning your social media messages and choosing the right platform to communicate them to buyers. The more information you get about consumers, the better you will be able to target them through your social media campaigns. There are numbers of opportunities in big data but opportunities always follow some challenges.

To handle these challenges effectively we need to know various computational method, security issues and computational complexities. Different methodology is required as per the data. although Big Data is built up to be as a the "Holy Grail" for healthcare, small data techniques using traditional statistical methods are, in many cases, more accurate and can lead to more improved healthcare outcomes than Big Data methods. Big Data for healthcare may cause more problems for the healthcare industry than solutions, and in short, when it comes to the use of data in healthcare, "size isn't everything."[4].On the basis of observation, challenges are categorized i) Security ii) Knowledge discovery with accuracy iii)Data storage and processing iv)Uncertainty of Data Management v) Talent Gap.

A) Security in Big Data

Big data security is the collective term for all the measures and tools used to guard both the data and analytics processes from attacks, theft, or other malicious activities that could harm or negatively affect them. The first challenge is incoming data, which could be corrupted or intercepted in transit. The second is data in storage, which can be stolen or held hostage while resting on cloud or on-premise servers. The last is data that is being outputted, which seems unimportant but could provide an access point for hackers or other malicious parties.

B) Knowledge Discovery with Accuracy

Big data has been deemed effective for decision-making as it improves ample business processes from marketing, analytical Customer Relationship Management (aCRM) to analytically Supply Chain Management (aSCM). This is the next arena for the 21st century. According to IBM, about 80% of organizational data is unstructured, meaning that there is a significant prospect to leverage the analysis of unstructured data. If such an opportunity is unlocked, then such a potential signifies the next challenge in big data for firms which use big data to extract valuable information for making informed decisions for a competitive advantage. For instance, companies like Amazon, Google or eBay used text analytics to analyze vast knowledge, communicate with customers and enhance operations. This is just one example where research is scant to assess the role of text analytics as an enabler of knowledge management. There is an opportunity in big data discipline to discover hidden knowledge so new knowledge can be generated. Furthermore, knowledge has become even more important for firms to acquire information from a wider array of sources, and such knowledge needs to be managed effectively so to assist firms to meet their ample challenges and to be better positioned for attaining lasting competitive advantage. At this stage, knowledge management becomes critical for enhancing firms decision-making power. The proliferation of knowledge has given rise to the concept of big data. In this regard, many firms that embraced big data have outperformed those who did not utilize big data in these business functions. Firms that outperformed did so since their speed and accuracy of decision-making led them to outperform others. 10 billion cell-phones that will come to be used by 2020, 294 billion emails that will be sent daily, and trillion of sensors which are due for collaborative monitoring and tracking and as a result populate the Internet of Things (IOT) with real-time data. Big data and big data analytics will enlighten hidden data patterns and such a wonder will bring imperative potentials to knowledge management.

C) Data Storage and Processing

To store data large infrastructure has to be set up. It is important that companies gain proper insights from big data analytics and it is important that the correct department has access to this information. A major challenge in big data analytics is bridging this gap in an effective fashion. But storage is not a major issue because cloud hosting services are there.

D) Talent Gap

On one side, there are very few experts available in this field. Because Big data is a complex field and people who understand the complexity and intricate nature of this field are few. The second side, because big data is continuously expanding, there are new companies and technologies that are being developed every day. A big challenge for companies is to find out which technology works bests for them without the introduction of new risks and problems.

2. OPEN SOURCE TOOLS FOR BIG DATA PROCESSING

There are numbers of open source tools. Based on popularity and usability, the open source tools are as follows:

Apache Hadoop is the most prominent and used tool in big data industry with its enormous capability of large-scale processing data. This is 100% open source framework and runs on commodity hardware in an existing data centre. Features of Hadoop are Authentication improvements when using HTTP proxy server, Specification for Hadoop

Compatible File system effort, Support for POSIX-style file system extended attributes, it offers robust ecosystem that is well suited to meet the analytical needs of developer, it brings Flexibility in Data Processing and it allows for faster data Processing.

The Apache Spark is an open source big data tool which is fills the gaps of Apache Hadoop concerning data processing. Spark can handle both batch data and real-time data. As Spark does in-memory data processing, it processes data much faster than traditional disk processing. This is indeed a plus point for data analysts handling certain types of data to achieve the faster outcome. Apache Spark is flexible to work with HDFS as well as with other data stores. For example, with OpenStack Swift or Apache Cassandra. It's also quite easy to run Spark on a single local system to make development and testing easier. Spark is an alternative to Hadoop's MapReduce. Spark can run jobs 100 times faster than Hadoop's MapReduce.

Apache Storm is a distributed real-time framework for reliably processing the unbounded data stream. The framework supports any programming language. The unique features of Apache Storm are it benchmarked as processing one million 100 byte messages per second per node, it uses parallel calculations that run across a cluster of machines, it will automatically restart in case a node dies. The worker will be restarted on another node, Storm guarantees that each unit of data will be processed at least once or exactly once and once deployed Storm is surely easiest tool for Big data analysis

Apache Cassandra is a distributed type database to manage a large set of data across the servers. This is one of the best big data tools that mainly process structured data sets. It provides highly available service with no single point of failure. Additionally, it has certain capabilities which no other relational database and any NoSQL database can provide. Apache Cassandra architecture does not follow master-slave architecture, and all nodes play the same role. It can handle numerous concurrent users across data centres. Hence, adding a new node is no matter in the existing cluster even at its up time. The features of Cassandra are support for replicating across multiple data centers by providing lower latency for users, data is automatically replicated to multiple nodes for fault-tolerance, it is most suitable for applications that can't afford to lose data, even when an entire data centre is down and Cassandra offers support contracts and services are available from third parties.

RapidMiner follows a client/server model where the server could be located on-premise, or in a cloud infrastructure. It is written in Java and provides a GUI to design and execute workflows. It can provide 99% of an advanced analytical solution. It allows multiple data management methods, GUI or batch processing, integrates with in-house databases, interactive and shareable dashboards, Big Data predictive analytics, remote analysis processing, data filtering, merging, joining and aggregating.

MongoDB is an open source NoSQL database which is cross-platform compatible with many built-in features. It is ideal for the business that needs fast and real-time data for instant decisions. It is ideal for the users who want data-driven experiences. It runs on MEAN software stack, NET applications and Java platform.

Neo4j: Hadoop may not be a right choice for all big data related problems. For example, when you need to deal with large volume of network data or graph related issue like social networking or demographic pattern, a graph database may be a perfect choice.Neo4j is one of the big data tools that is widely used graph database in big data industry. It follows the fundamental structure of graph database which is interconnected node-relationship of data. It maintains a key-value pattern in data storing.

Apache SAMOA is among well known big data tools used for distributed streaming algorithms for big data mining. Not only data mining it is also used for other machine learning tasks such as classification, clustering, regression and programming abstractions for new algorithms. Due to following reasons, Samoa has got immense importance as the open source big data tool in the industry:

- You can program once and run it everywhere
- Its existing infrastructure is reusable. Hence, you can avoid deploying cycles.
- No system downtime
- No need for complex backup or update process

Cloudera is the easiest, fastest and highly secure modern big data platform. It allows anyone to get any data across any environment within single, scalable platform.

Pentaho provides big data tools to extract, prepare and blend data. It offers visualizations and analytics that change the way to run any business. This Big data tool allows turning big data into big insights. It allows data access and integration for effective data visualization, It empowers users to architect big data at the source and stream them for accurate analytics, seamlessly switch or combine data processing with in-cluster execution to get maximum processing, allow checking data with easy access to analytics, including charts, visualizations, and reporting and supports wide spectrum of big data sources by offering unique capabilities.

Statwing is an easy-to-use statistical tool. It was built by and for big data analysts. Its modern interface chooses statistical tests automatically. Statwing helps to clean data, explore relationships, and create charts in minutes. It allows creating histograms, scatter plots, heat maps, and bar charts that export to Excel or PowerPoint. It also translates results into plain English, so analysts unfamiliar with statistical analysis

R is one of the most comprehensive statistical analysis packages. It is open-source, free, multi-paradigm and dynamic software environment. It is written in C, FORTRAN and R programming languages. It is broadly used by statisticians and data miners. Its use cases include data analysis, data manipulation, calculation, and graphical display.

Lumify is a free and open source tool for big data fusion/integration, analytics, and visualization. Its primary features include full-text search, 2D and 3D graph visualizations, automatic layouts, link analysis between graph entities, integration with mapping systems, geospatial analysis, and multimedia analysis, real-time collaboration through a set of projects or workspaces.

Datawrapper is an open source platform for data visualization that aids its users to generate simple, precise and embeddable charts very quickly. Its major customers are newsrooms that are spread all over the world. Some of the names include The Times, Fortune, Mother Jones, Bloomberg, Twitter etc.

KNIME stands for Konstanz Information Miner which is an open source tool that is used for Enterprise reporting, integration, research, CRM, data mining, data analytics, text mining, and business intelligence. It supports Linux, OS X, and Windows operating systems. It can be considered as a good alternative to SAS. Some of the top companies using Knime include Comcast, Johnson & Johnson, Canadian Tire, etc.

Qubole data service is an independent and all-inclusive Big data platform that manages, learns and optimizes on its own from your usage. This lets the data team concentrate on business outcomes instead of managing the platform. Out of the many, few famous names that use Qubole include Warner music group, Adobe, and Gannett.

Tableau is a software solution for business intelligence and analytics which present a variety of integrated products that aid the world's largest organizations in visualizing and understanding their data. The software contains three main products i.e. Tableau Desktop (for the analyst), Tableau Server (for the enterprise) and Tableau Online (to the cloud). Also, Tableau Reader and Tableau Public are the two more products that have been recently added. Tableau is capable of handling all data sizes and is easy to get to for technical and non-technical customer base and it gives you real-time customized dashboards. It is a great tool for data visualization and exploration.

Apache Hive is a Data warehouse system which is built to work on Hadoop. It is used to querying and managing large datasets residing in distributed storage. Before becoming an open source project of Apache Hadoop, Hive was originated in Facebook. It provides a mechanism to project structure onto the data in Hadoop and to query that data using a SQL-like language called HiveQL (HQL).Hive is used because the tables in Hive are similar to tables in a relational database. If you are familiar with SQL, it's a cakewalk. Many users can simultaneously query the data using Hive-QL.

3. RESEARCH AREAS

Healthcare: Making use of the petabytes of patient's data, the organization can extract meaningful information and then build applications that can predict the patient's deteriorating condition in advance.

Traffic control: Traffic congestion is a major challenge for many cities globally. Effective use of data and sensors will be key to managing traffic better as cities become increasingly densely populated.

Manufacturing: Analyzing big data in the manufacturing industry can reduce component defects, improve product quality, increase efficiency, and save time and money.

Search Quality: Every time we are extracting information from Google, we are simultaneously generating data for it. Google stores this data and uses it to improve its search quality.

Retail: Retail has some of the tightest margins, and is one of the greatest beneficiaries of big data. The beauty of using big data in retail is to understand consumer behaviour. Amazon's recommendation engine provides suggestion based on the browsing history of the consumer.

Telecom Sector: Telecom sectors collects information analyzes it and provides solutions to different problems. By using Big Data applications, telecom companies have been able to significantly reduce data packet loss, which occurs when networks are overloaded, and thus, providing a seamless connection to their customers.

Education is another extremely hot topic across the country. What can be done to improve education with the help of Big Data? There are a lot of different things to be done and up-to-date. Big data helps governments to understand more about educational needs on a local and federal level in order to ensure that the youth of the nation are getting the best possible education in order to serve the country in the future.

Transportation: Every day millions of Indians are on the road driving. There are so many different nuances to driver safety, from roads to police officers, weather conditions and vehicle safety that it's impossible to control everything that might cause an accident. However, with big data governments can better oversee transportation to ensure better roads, safer roadways, better routes and new routes.

Agriculture: We can keep track land and livestock that exists in our country and across the globe. All the different crops that are grown, the animals that are held and so many other complicated issues come together in the agriculture world to form a very difficult job for the government. It's hard to monitor because of the vast numbers. Big data is changing the ways governments manage and support the farmers and their resources. Its ability to gather huge amounts of information and analyze them quickly makes all the difference.

Poverty: There is so much poverty in the world. It's extremely difficult to combat and we've been trying to do so for thousands of years. Big data gives governments tools to discover more effective and innovative ideas on how to decrease poverty across the world. It's easier to pinpoint areas with the greatest need and how those needs can be met. Big data technology is vitally important for governments across the world. It can't solve every problem, but it's a step in the right direction. It's giving leaders the tools necessary to enact important changes that will be of benefit for citizens now and in the future.

4. CONCLUSION

In recent years variety of data is generated at a dramatic pace. Analyzing these data is challenging for a common man. To end of this paper, we survey the various tools, challenges and research issue to analyze big data. We also discussed the various research opportunities in which big data play an important role. From this survey, it is observed that every big data platform has its individual importance. Some of them are used for batch processing, real time processing, visualization or data processing and storing. We believe that in future researchers will pay more attention to the real life problems which is discussed with the help of big data.

REFERENCES

[1] https://www.outsource2india.com

[2] https://tweakyourbiz.com/growth/productivit/big-data-growth

[3] 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.

[4] Househ MS, Aldosari B, Alanzi A, College of Public Health and Health Informatics, King Saud bin Abdulaziz University, for Health Sciences, Riyadh, Saudi Arabia, Big Data , Big Problems : A Health Care Perspectives, NCBI

[5] Mr. G. P. Babhulkar and P. Pimple, A Review on Globalized Internet of Things(IoT) and Applications, IJIRCCE, volume 5, issue 10,October 2017.