A REVIEW PAPER: BIG DATA PROCESSING AND APPLICATIONS

Vinita Tapaskar[1], Raghavendra Rao B[2]
[1]Assistant Professor, The Oxford College of Science, Bangalore
[2]Assistant Professor, Sri Sairam College of Engineering, Bangalore

Abstract: With the development of Social network and all Internet of Things technology there is volume's growth of data on Internet. This high speed growing data is unstructured in the form of blogs, posts , tweets , news articles, video , audio etc. This all is termed as Big Data. Big data is said to be the massive volume of information that is difficult to be processed using traditional database techniques. Big data can be of both types structured or unstructured. The growth of big data is not stoppable due to social network. The size of data available online merely is vast, with large data getting added every second. Every day 2.5 quintillion bytes of data are generated online. Big data holds tremendous use to improve our lives. Analyisted big data can be used to provide weather predictions, useful product recommendation, can suggest suitable medical treatment etc. Big data analytics has proven to be a boom for industry as it helps to extract useful patterns and unknown correlations of for identifying consumer market, various client preferences, different buying attributes and lot of other information from complex data sources. This paper targets to provide a detailed review on big data processing technique and comparative assessment of latest tools and frameworks used for big data analytics.

Keywords: Big Data, Data Analysis, Hadoop, Applications of Big Data, Analysis Process.

I] Introduction:

Big Data are high volume, high velocity, or high-variety data that requires unique forms of processing to enable enhanced decision making, insight discovery, and process optimization [1][2].

The term 'big data' is explain itself − a collection of large data sets that normal computing techniques cannot process. The term not only refers to the data, but also to the various frameworks, tools, and techniques involved. Technological advancement and the usage of new channels of communication like social networking and new, stronger devices have presented a challenge to industry that we have to find other ways to handle this large volume of data. All this big data is useful when processed. All this data analysis in meaning manner can provide a greater insight and can help in better decision making. The big data has features shown in Figure 1 which makes in different and complex for processing.

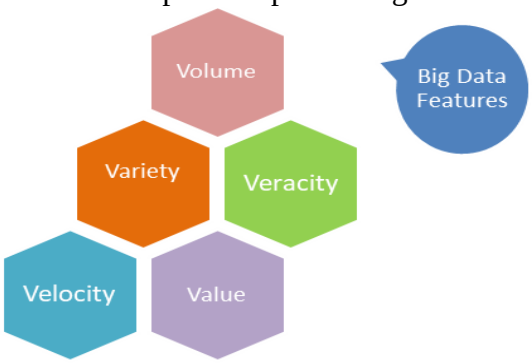

Figure 1: Features of Big Data

Big data analytics is used to extract valuable information from the raw material (data) taken from variety sources. That data helps us get hidden patterns, unknown correlations, market trends, and a lot more, depending analysis done. The primary intention of big data analytics is to provide valuable suggestion to make better decisions for the future[3]

II] Methodology of Big Data Analytics'

1. Data identification and collection- This phase, identify wide variety of data sources . Added data resources mean large chances of finding hidden correlations and patterns.

Tools are needed to capture terms, data and information from this heterogeneous data sources which are primary source or secondary sources.

2. Data storage- The captured structured and unstructured data need to be stored in databases or data warehouse. NoSQL databases are needed to accommodate Big Data as data is mostly unstructured and RDBMS fail to process this unstructured data in a optimized way. Various frameworks and databases have been developed that allow analytics tools to fetch and process data from these storages.
3. Data filtering and noise elimination- In this phase from gathered data we remove of replicated, corrupt, null and irrelevant data objects. Filtered data provide more accurate and meaning full result. But this filtered and removed data might be of some importance in another context or analysis so, it is advised sets in compressed form to save storage space
4. Data classification and extraction- This phase is responsible for extracting distinct data and converting it into a common data format that the underlying analytics tool can use for its purpose. The data to be converted to standard format so that it can be processed. This involves extracting significant fields or texts to lessen the volume of data to be submitted to analytics engine.
5. Data cleansing, validation and aggregation- In this stage we apply validation rules based on the business case to confirm the need and relevance of data extracted for analysis. Although it may be difficult sometimes to apply validation constraints to the extracted data due to complexity and volume. Aggregation helps to combining multiple data sets into less numbers based on common fields. This simplifies further data processing.
6. Data analysis and processing- This stage carries out actual data mining and analysis to identify unique and hidden patterns for making business decisions. Data analytics technique may differ based upon the scenario i.e. exploratory, confirmatory, predictive, prescriptive, diagnostic or descriptive [6].
7. Data visualization- This phase involves representation of analysis results into visual or graphical form that makes it easier to understand for the end user.
8. Final analysis result - This is the last step of the Big Data analytics where the final results of the analysis are provided to all business stakeholders who will take action based on patterns and trends identified.

III] Types of Big Data Analysis

1. Descriptive Analytics

This summarizes past data into a form that people can easily read. This helps in creating reports, like a revenue, profit and loss statements, sales report, and so on. Also, it helps in the matrix formation of social media data.

Use Case: The Chemical Company analyzed its past data to increase facility utilization across its office and lab space. Using descriptive analytics, the company was able to identify underutilized space.

2. **Diagnostic Analytics**

This type of analysis is done to understand the reason that caused a problem. Techniques like drill-down approach, deep data mining, and data recovery are examples. Organizations use diagnostic analytics because they provide an in-depth insight into a particular problem.

Use Case: An marketing and product based company report shows that their sales have gone down, although customers are adding products to their carts. This can be due to various reasons like the form didn't load correctly, the shipping fee is too high, or there are not enough payment options available. This is where you can use diagnostic analytics to find the reason.

3. Predictive Analytics

This type of analytics sees into the historical and current data to make predictions of the future. Predictive analytics uses data mining, AI, and machine learning to analyze current data and make predictions about the future. It works on predicting customer needs, market trends etc.

Use Case: Online payment service provider determines what kind of precautions they have to take to protect their clients against fraudulent transactions. Using predictive analytics, the company uses all the past payment data and

user behavior data and builds an algorithm that predicts fraudulent activities.

4. Prescriptive Analytics

This type of analytics provides the solution to a particular problem specified. Perspective analytics works with both descriptive and predictive analytics. Most of the time, it relies on Artificial Intelligence and machine learning.

Use Case: Prescriptive analytics can be used to maximize an airline's profit. This type of analytics is used to build an algorithm that will automatically adjust the flight fares based on numerous factors, including customer demand, weather, destination, holiday seasons, and oil prices.

III] Challenges of Big Data

Traditional data management and analysis are based on relational database management systems (RDBMSs). But such systems are only applicable to structured data, rather than semi-structured or unstructured data. In addition, RDBMSs increasingly utilize expensive hardware and suited for structured data. Traditional RDBMSs also cannot handle the high volume and heterogeneity of big data. Below is the list of problems in the development of big data applications, and the key challenges faced during processing the big data.

- Data representation: Data representation is important to enhance the meaningfulness of data for computerized system analysis and user interpretation. Proper data representation reflects data structure, class, and type in addition to integrated technologies to facilitate efficient operations on variety of datasets.
- Redundancy reduction and data compression: This challenge is to reduce redundancy and storage cost for the organization. The reduction should be lossless data are not affected.
- Data life cycle management: The present storage system cannot support huge data. The hidden values of big data generally depend on data freshness; therefore, a data importance principle that is related to analytical value should be developed to decide which pieces of data should be stored and which ones should be removed. The data has it life and storage should be carefully used and decided.
- Analytical mechanism: The big data analytical system processes masses of heterogeneous data within a limited time. Non-relational databases those are NO-SQL databases have shown their unique advantages in processing unstructured data and have become main stream in big data analysis. Even so, non-relational databases still encounter performance- and application-related problems. A solution that facilitates compromise between RDBMSs and non-relational databases must be obtained.
- Data confidentiality: Most current big data service providers cannot effectively maintain and analyze large datasets due to limited capacity of infrastructure and human power. These users must rely on professionals or tools instead, thus increasing the potential safety risks. Therefore, big data analysis may be done by a third party only when proper preventive measures are taken to protect sensitive data. Due to involvement of third party there is a threat to data security.
- Energy management: With the increase in data volume and analytical demands, the processing, storage, and transmission of big data inevitably consume increasing amounts of processing and memory resource. Therefore, system consumption control and management mechanisms must be established for big data while ensuring expandability and accessibility.
- Expendability and scalability: The analytical system of big data must support current and future datasets. The analytical algorithm must be capable of processing expanding and increasingly complex datasets not current but also future sets..
- Cooperation: Big Data analysis is an interdisciplinary research field that requires experts of different fields to cooperate in deriving the potential of big data analysis solution. Thus, comprehensive network architecture must be established for big data to provide scientists and engineers in various fields with access to different kinds of data, as well as to maximize the expertise of such individuals in collaborating to meet analytical objectives.

IV] Tools used for Big Data Analysis

This study shows the top tools used for big data analytics. Figure 2 shows the popular tools used [3][4][5] for big data analysis .

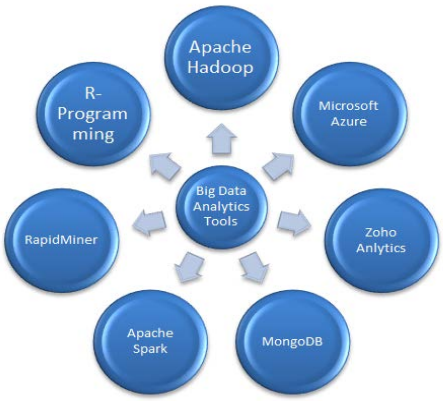

Figure 2: Tools of Big Data Analysis

1. R Programming

R programming is one of the finest big data analytics tools. It is a important statistics programming language that can be used for statistical analysis, scientific computing, data visualization. The R programming language can also extend itself to perform various big data analytics operations and provides patterns.

- It includes a set of operations for working with arrays, and matrices.
- Effective and efficient storage facility and data handling.
- It offers a total, integrated set of big data tools for data analysis.
- It includes graphical data analysis tools that may be seen on screen or printed.

2. Apache Hadoop

Apache Hadoop is the top big data analytics tool which is open source. It is a software framework used to store data and run applications on clustering of commodity servers. It is the framework that consists of a software ecosystem which is called as HADOOP ecosystem.

- **Distributed Processing & Storage:** The framework offers a lot of flexibility and manages distributed processing and storage on its own. Processing logic should be only written by the user.
- **Highly and easily scalable:** Vertical and horizontal scalability are both available in HADOOP. In horizontal scalability it allows more nodes to be added to the system on the fly as data volume and processing demands rise without affecting existing systems or applications.
- **Cost-effective:** Hadoop delivers cost efficiency by introducing massively parallel computation to commodity servers. It always results in a significant drop in the cost per terabyte of storage. The commodity servers are increased and decreased as per need.

3. MongoDB

MongoDB is one of the popular document data store software in the world. It is based storage of unstructured data with higher volume of data than RDBMS-based database software has failed to do. MongoDB is robust, and it is one of the best big data analytics tools.

- **High Performance:** Due to distinctiveness such as scalability, replication, indexing, and others, MongoDB has a very high speed compared to other databases.
- **Replication:** MongoDB enables high availability and redundancy by creating numerous copies of the data and sending these copies to a separate server. This ensures that in case one server fails, the data can be accessed from another server. This effectively assure availability and reliability of data.
- **Indexing:** Every field in the documents in the MongoDB database is indexed with main and secondary indices. This makes it easier and faster to obtain or search data from the volume of data. If the data isn't indexed, the database will have to search each document individually for the query. Indexing makes the search of data faster which is necessary in big data processing.

4. RapidMiner

RapidMiner is one of the platforms for analysts to integrate data preparation, machine learning, analytical model deployment, etc. It is the best big data analytics tools free that can be used for data analytics and text mining.

- RapidMiner can connect to various Hadoop clusters, which includes Cloud era

Distribution, MapR Hadoop, Apache Hadoop with Hive,

- It support different data sources like Excel, , Oracle, Microsoft SQL, Ingres, Sybase, MySQL, SPSS, Postgres, dBase, Text files
- Several data management approaches are available. It includes data loading, modeling, transformation, and visualization.

5. Apache Spark

It is one of the best and most powerful big data analytics tools which is open source. It can process a high volume data sets with the help of its data processing framework. It is pretty easy to distribute data processing tasks across multiple computers with its with other distributed computing tools.

- Spark code can be reused and may be used to connect streaming data with historical data, batch processing, and conduct ad-hoc queries on streaming data.
- Spark allows Hadoop applications to run quicker and use less storage. Spark reduces the number of disk read/write operations required for intermediate results. It keeps data in memory and only conducts disk IO Operations when needed. DAG, query optimizer and a highly efficient physical execution engine use by Spark to carry out this.
- Hadoop may used as an input data basis or a destination for Spark. Apache Spark is well connected with Hadoop's HDFS file system and supports various file formats.

6. Microsoft Azura

Microsoft Azure is one of the well known big data analytics tools. Microsoft Azure is called as Windows Azure. It is a public cloud computing platform that Microsoft handles, and it is the leading platform that provides a different range of services, which include computing, analytics, storage, and networking.

- **Scalability has been improved:** Microsoft Azure can be scaled up or scaled down fast to meet your demands based on requirements. This makes it a practical choice for numerous enterprises with varying sizes.
- **Strong Analytical Support:** Microsoft Azure has built-in data analysis and critical service.
- **System of storing that is unique:** Azure offers more delivery points and data centres. That is why it can provide a better user experience and deliver content to your business environment more quickly.

➢ **Zoha Analytics**

Zoho Analytics is one of the most reliable big data analytics tools. It is a Business Intelligence tool that works seamlessly for data analytics and helps us to visually analyze the data to get a better understanding of the raw data.

- **Geography Visualization:** Interactive map charts allow sales professionals to compare geographical performance quickly and simply. Comparisons can do between countries, states, local areas, and other areas.
- **Connects to various data connectors:** Connection is easier between files and feeds, CRM systems, cloud storage, various databases, Google Analytics, social media, financial platforms, e-commerce platforms, Human Resource.
- **White Labelling:** Individual reports or dashboards can be set in using this technique. And the solution is white-labelled to simplify integrating into websites and apps.

V] Applications of Big Data Analytics

Here are some of the sectors where Big Data is actively used:

➢ Ecommerce - Predicting customer trends and optimizing prices are a few of the ways e-commerce uses Big Data analytics

➢ Marketing - Big Data analytics helps to drive to know the customer needs based on seasons and time period

➢ Education - Used to develop new and improve existing courses based on market requirements

➢ Healthcare - With the help of a patient's medical history, Big Data analytics is used to predict how likely patients will have health issues

➢ Media and entertainment - Used to understand the demand of shows, movies, songs and deliver a personalized recommendation list to each users

- Banking - Customer income and spending patterns help to predict the possibility of choosing various banking offers, like loans and credit cards

- Telecommunications - Used to forecast network capacity and improve customer experience
- Government - Big Data analytics helps governments in law enforcement.

Conclusion:

The rate of development of information processing tools is reasonably much slower than the rate of development of information. Currently available tools in the market do not address all the problems of Big Data analytics. Even the most high-tech tools and techniques can't justify real-time analysis in true sense. Though they have fairly increased the ease of handling diverse data sets and reduced the time of data processing. There are still some unsolved issues related to effective storage, searching, analysis, sharing and security. This gives a way for future improvements and developments of Big Data analytics tools.

References:

[1] S. Mujawar, S. Kulkami, "Big Data: Tools and Applications", International Journal of Computer Applications, vol. 115, No. 23, pp. 7-11, 2015.

[2] T. Erl, W. Khattak, and P. Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques, Prentice Hall, India, pp. 65-88, 2015.

[3] N. Khan et. al, "Big Data: Survey, Technologies, Opportunities, and Challenges", The Scientific World Journal, vol.2014, Issue.4, pp.1-18, 2014.

[4] Online source, [Available] Top 7 Big Data Analytics Tools To Use In The Year 2022 (statanalytica.com)

[5] Online source, [Available] Big Data Analytics: Types, Tools and Applications [Updated] (simplilearn.com)

[6] S Kaushal, J.K. Bajwa, "Analytical Review of User Perceived Testing Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, issue 10, 2012.

[7] S. M. Ali et.al, "Big Data Visualization: Tools and Challenges", 2nd International Conference on Contemporary Computing and Informatics,2016.

[8] Firas D. Ahmed, MazlinaBinti Abdul MajidAge, Aws NaserJaber, MohdSharifuddin Ahmad Agent Based Big Data Analytics in Retailing: A Case Study, Aug 2015 Conference Paper