IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BIG DATA

THE REVOLUTIONARY CONCEPT IN DATA ANALYTICS

[1]Tamanna Gajanan Shenoy

[1]Undergraduate Engineering Student

[1]Computer Science and Engineering (IoT & Cybersecurity with Blockchain Technology),

[1]Lokmanya Tilak College of Engineering, Koparkhairne, Navi Mumbai, Maharashtra, India

Abstract: Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations is named big data analytics. This useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. This paper presents an overview of big data's content, scope, samples, methods, advantages and tools used for big data.

Index Terms **-** big data, volume, variety, velocity, and big data analytics.

I. INTRODUCTION

Big Data is a collection of data that is huge in volume yet growing exponentially with time. It is data so large a size and complex that none of the traditional data management tools can store it or process it efficiently. Big data is also data but in huge sizes.

Big data has become the revolution of Information Technology which is transforming industries around the world. Big data is a combination of technology and data that integrates reports and accesses all available data filtering, reporting, and correlating insights achievable with previous data technologies. Current usage of the term big data tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from big data, and seldom to a particular size of data set. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

For example-

- The New York Stock Exchange is an example of Big Data that generates about one terabyte of new trade data per day.
- The statistics show that 500+terabytes of new data get ingested into the databases of social media sites Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments, etc.
- A single Jet engine can generate 10+terabytes of data in 30 minutes of flight time. With many thousand flights per day, the generation of data reaches up to many Petabytes.

II. TYPES OF BIG DATA

1. **Structured:** Any data that can be stored, accessed, and processed in the form of fixed format is termed 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when the size of such data grows to a huge extent, typical sizes being in the range of multiple zettabytes.

2. **Unstructured:** Any data with an unknown form or structure is classified as unstructured data. In addition to the size being huge, unstructured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos, etc. Now day organizations have wealth of data available to them but unfortunately, they don't know how to derive value from it since this data is in its raw form or unstructured format.

3. **Semi-structured:** Semi-structured data can contain both forms of data. We can see semi-structured data as structured in form but it is actually not defined with e.g. a table definition in relational DBMS. An Example of semi-structured data is data represented in an XML file.

III. CHARACTERISTICS OF BIG DATA

1. **Volume:** The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. The size of big data is usually larger than terabytes and petabytes. This can be data of unknown value, such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment.

2. **Variety:** Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining, and analyzing data. The type and nature of the data. . Unstructured and semi- structured data types such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

3. **Velocity:** Here Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous. The speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

4. **Value:** The worth of information that can be achieved by the processing and analysis of large datasets. Value also can be measured by an assessment of the other qualities of big data. The value may also represent the profitability of information that is retrieved from the analysis of big data.

5. **Variability:** This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively. The characteristics of the changing formats, structure, or sources of big data. Big data can include structured, unstructured, or combinations of structured and unstructured data. Big data analysis may integrate raw data from multiple sources. The processing of raw data may also involve transformations of unstructured data to structured data.

IV. How Big Data Analytics Works

Big data analytics refers to collecting, processing, cleaning, and analyzing large datasets to help organizations operationalize their big data.

1. **Collect Data**

   Data collection looks different for every organization. With today's technology, organizations can gather both structured and unstructured data from a variety of sources — from cloud storage to mobile applications to in-store IoT sensors and beyond. Some data will be stored in data warehouses where business intelligence tools and solutions can access it easily. Raw or unstructured data that is too diverse or complex for a warehouse may be assigned metadata and stored in a data lake.

2. **Process Data**

   Once data is collected and stored, it must be organized properly to get accurate results on analytical queries, especially when it's large and unstructured. Available data is growing exponentially, making data processing a challenge for organizations. One processing option is batch processing, which looks at large data blocks over time. Batch processing is useful when there is a longer turnaround time between collecting and analyzing data. Stream processing looks at small batches of data at once, shortening the delay time between collection and analysis for quicker decision-making. Stream processing is more complex and often more expensive.

3. **Clean Data**

   Data big or small requires scrubbing to improve data quality and get stronger results; all data must be formatted correctly, and any duplicative or irrelevant data must be eliminated or accounted for. Dirty data can obscure and mislead, creating flawed insights.

4. **Analyze Data**

   Getting big data into a usable state takes time. Once it's ready, advanced analytics processes can turn big data into big insights. Some of these big data analysis methods include:

- **Data mining** sorts through large datasets to identify patterns and relationships by identifying anomalies and creating data clusters.
- **Predictive analytics** uses an organization's historical data to make predictions about the future, identifying upcoming risks and opportunities.
- **Deep learning** imitates human learning patterns by using artificial intelligence and machine learning to layer algorithms and find patterns in the most complex and abstract data.

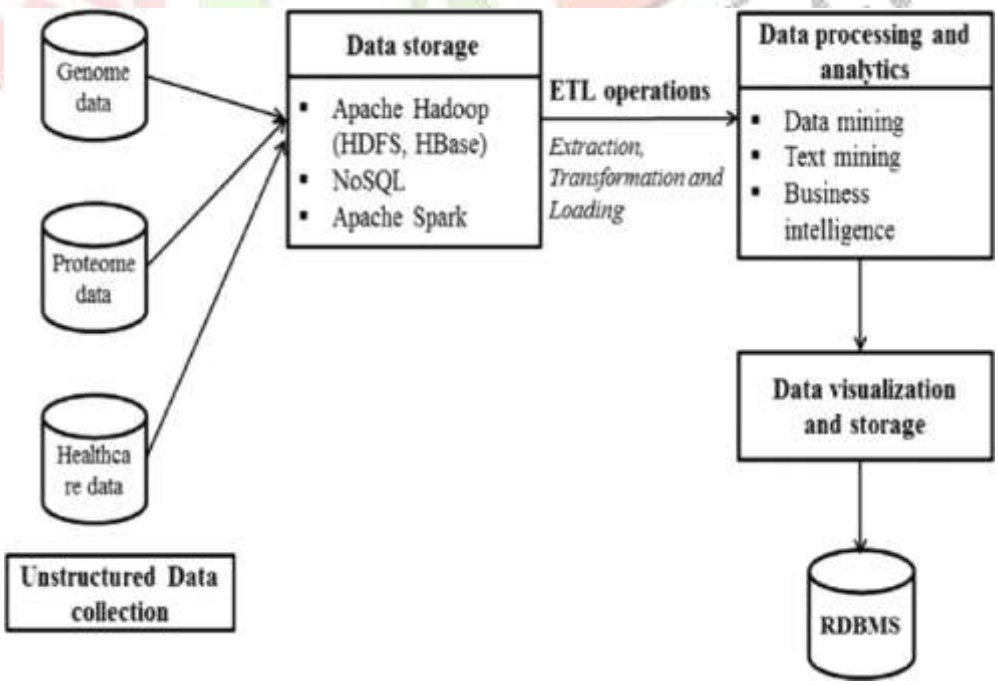

Fig. Big Data Workflow Architecture in Bioinformatics

V. TOOLS FOR BIG DATA ANALYTICS

1. **Hadoop** is an open-source framework that efficiently stores and processes big datasets on clusters of commodity hardware. This framework is free and can handle large amounts of structured and unstructured data, making it a valuable mainstay for any big data operation.

2. **NoSQL** databases are non-relational data management systems that do not require a fixed scheme, making them a great option for big, raw, unstructured data. NoSQL stands for "not only SQL," and these databases can handle a variety of data models.

3. **MapReduce** is an essential component to the Hadoop framework serving two functions. The first is mapping, which filters data to various nodes within the cluster. The second is reducing, which organizes and reduces the results from each node to answer a query.

4. **YARN** stands for "Yet Another Resource Negotiator." It is another component of second-generation Hadoop. The cluster management technology helps with job scheduling and resource management in the cluster.

5. **Spark** is an open source cluster computing framework that uses implicit data parallelism and fault tolerance to provide an interface for programming entire clusters. Spark can handle both batch and stream processing for fast computation.

6. **Tableau** is an end-to-end data analytics platform that allows you to prep, analyze, collaborate, and share your big data insights. Tableau excels in self-service visual analysis, allowing people to ask new questions of governed big data and easily share those insights across the organization.

VI. ADVANTAGES OF BIG DATA

The ability to process Big Data in DBMS brings in multiple benefits, such as-

1. **Businesses can utilize outside intelligence while taking decisions:** Most businesses are primarily aimed at improving their decision-making by investing in big data. As more detail is accessible in a functional way, it is easier to see what consumers want or don't want. Access to social data from search engines and sites like Facebook, and Twitter is enabling organizations to fine-tune their business strategies.

2. **Improved customer service:** Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

3. **Detects fraud:** Another important advantage company's find with big data is that it can help identify fraud. That benefit is one that the financial services industry most often mentions, but any company can take advantage of those opportunities. AI and machine learning will detect anomalies or transaction patterns for individual accounts that aren't part of the daily routine

4. **Security:** Analysis of the data in real time allows you to spot anomalies in expected patterns almost instantly. This allows you to identify and, in fact, fix any problems that may have occurred, resulting in a better customer experience.

VII. DISADVANTAGES OF BIG DATA

1. **Cost of Implementation:** Many of the big data resources available today depend solely on open source technologies. This fact ensures that tech costs are practically gone from this attempt to collect information, but it also poses a problem with hardware, repair, and staffing issues.

2. **Lack of talent:** Big data analytics isn't an asset that the average IT personnel can look at to glean useful information for decisions. Companies need information scientists who know how to glean results from this approach.

3. **Security risks:** Most of the information that companies collect in a data lake includes sensitive info that requires a specific level of protection. Having access to these analytics can make an organization become an attractive target for a potential cyberattack.

VIII. APPLICATIONS OF BIG DATA

1. Banking & Securities

- **Fraud Detection:** By applying analytics and machine learning, we shall be able to define normal activity based on a customer's history and distinguish it from an unusual behavior indicating fraud. Immediate actions can be taken such as blocking the transaction, which shall improve profitability.

- **Customer Segmentation :** Big data enables the banks to group their customers into distinct segments, which are then, defined by data sets that may include customer demographics, daily transactions, interactions with online and telephone customer service systems, and external data, such as the value of their residences.

2. Communications, Media and Services

- **Predicting the Audience's Interest:** These days traditional content delivery has been replaced with services like pay per view, on demand, live streaming and much more. In the process of content delivery across these forms, broadcasters also collect a vast amount of user data which can give an in-depth understanding of behaviour and preferences.

- **Effective Advertisement Targeting:** Big Data takes the guesswork out of programmatic advertising, which has been done in a random manner hoping the customers shall like what is being shown to them. It helps advertisers and businesses pinpoint the exact preferences of customers. It also gives a better understanding of what type of content viewers watch at what time and duration resulting in improved efficiency of ad targeting.

3. Transportation

- **Road Safety:** Road accidents are unpredictable and the analysis of the factors which are responsible for road accidents is important to prevent the same instances in the future. Thus, big data analysis has been used widely to overcome the number of accidents all along maintaining the infrastructure.

- **Traffic Management:** To curb heavy traffic, many cities have traffic systems that have integrated analytics to identify traffic problems quickly.

IX. KEY CHALLENGES

Despite having so many positive aspects of Big Data analytics, it's still not up to the mark in some areas as it is still in an evolving phase of technology. Some of the key challenges faced by and should be taken care of in Big Data analytics right now are:

- **High Solution Cost -** The building up and setup of these big data analytics infrastructures have very high financial demands. And maintaining ROI (return on Investment) is fairly a difficult task for small-scale projects.

- **Unavailability of skilled manpower -** As this field is still in its evolving phase, the specialized and skilled personnel for this are very less as compared to demand.

- **Difficulty in data integration -** As the data can and is coming from various sources so some of this data is structured and some of it is unstructured so integrating this data becomes a fairly difficult task.

- **Handling Incoming data -** As the data nowadays coming is of huge intensity and requires both speed and storage infrastructure, which sometimes becomes a challenge for data analysis.

X. CONCLUSION

In conclusion, the review of big data in this article has provided an in-depth overview of its content, scope, samples, methods, advantages, and challenges. While the literature offers a wealth of data, tools, and techniques, there are still numerous areas that require further consideration, discussion, improvement, development, and analysis. One critical issue that stands out is the privacy and security concerns surrounding big data. As we move forward, it is evident that addressing these concerns will be paramount in ensuring the responsible and ethical use of big data. The future discussions and developments in this field will undoubtedly focus on enhancing privacy measures and strengthening security protocols to safeguard the integrity and confidentiality of data. It is imperative that researchers, practitioners, and policymakers collaborate to address these challenges and pave the way for a more secure and privacy-conscious big data landscape.

REFERENCES

[1] J. Panneerselvam, L. Liu, and R. Hill, "An introduction to big data," in Application of Big Data for National Security. Elsevier, 2015, pp.3-13.

[2] T. Mahmood and U. Afzal, "Security analytics: Big Data Analytics for Cybersecurity: A review of trends, techniques and tools," in 2013 2nd National Conference on Information Assurance (NCIA), Dec 2013, pp.129-134.

[3] A. Vailaya, "What's All the Buzz Around "Big Data?"", IEEE Women in Engineering Magazine, December 2012.

[4] Big data ppt (slideshare.net)

[5] S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011

[6] K. Smith and M. Johnson, "Understanding the Impact of Big Data in Healthcare," Journal of Health Information Management, vol. 25, no. 3, pp. 45-52, 2018.

[7] R. Gupta and S. Sharma, "Big Data Applications in Financial Services: A Comprehensive Review," International Journal of Information Management, vol. 36, no. 5, pp. 667-679, 2016.

[8] L. Chen and H. Wang, "Big Data Analytics in Smart Cities: Opportunities and Challenges," IEEE Transactions on Industrial Informatics, vol. 14, no. 6, pp. 2538-2546, 2018.

[9] A. Patel and S. Desai, "Big Data and Machine Learning for Predictive Maintenance in Manufacturing," Procedia Computer Science, vol. 132, pp. 137-144, 2018.

[10] M. Lee and J. Kim, "Big Data Analytics for Customer Relationship Management: A Review and Future Directions," Expert Systems with Applications, vol. 47, pp. 233-246, 2016