Big Data Analytics

Sharanya S Nair

COMPUTER SCIENCE ENGINEERING

SCHOOL OF ENGINEERING AND TECHNOLOGY

SHARDA UNIVERSITY, GREATER NOIDA

Abstract

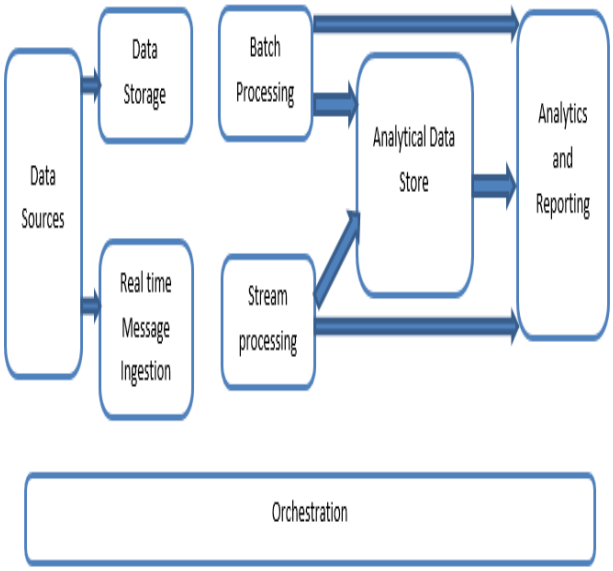
A massive amount of data is being generated and stored in repositories by using cloud computing . for the datathat is being stored in this much amounts like in tera and exabytes we need to take lot of efforts at different level and extract the data with good decision making . It is not possible for the tradition data base system to store that much amount of data and produce data that is wanted in few minutes. Therefore, big data analysis is the current area of research and development presently . The basic objective of this paper is to explore the basic understating of big data, it's architecture the benefits and drawbacks of the same .Additionally, it taks about the future scope of big data and where it's advancement can lead. Keywords—Big data analytics; Hadoop; Massive data;drawbacks, benefits, scopes.

INTRODUCTION

In this world of digitization, we see data is being generated from various sources and the faster transition rates of these technologies has led to the growth of huge data. It has made many evolution breakthroughs in many fields with the collection of such datasets. In general, it refers to the gathering of huge and sophisticated datasets which are difficult to process using traditional management tools or processing applications. They are usually stored in huge amount of data like in petabytes and so and are in structured, semi-structured and unstructured format. Formally, we have three V's in this that are volume, velocity, variety .Volume refers to the large amount of knowledge that are being generated everyday whereas velocity is that the rate of growth and the way fast the info are gathered for being analysis. Thesestructuredsemi-structured and unstructured data provide with variety of information. The fourth V in it refers to veracity which is held for the availability and accountability . In big data, we uses different technologies and techniques mainly in an objective of to process data of high volume,, velocity, variety, veracity. Gandomi and Haider has discussed some of the important methods to extract information from big data. The exact definition for big data hasn't yet be defined as it's problem specific. It is expected that the growth of big data will be in 1000 billion by 2025.big data is mainly pertaining social business and internet of things ,cloud computing and is built on third platform. Large data sets are managed in data warehouses. It is an issue to extract precise data or information from these big repositories.it is not possible to handle the massive datasets successfully using the presented approaches. There is no coordination between the data mining and

analysis tools this is one of the key problems of big data .they are usually arises when we try to perform knowledge discovery. A fundamental problem is how to quantitatively describe the essential characteristics of big data. There is a requirement for epistemological implications in describing data revolution the knowledge of big data tools can further reach to better future in many organisations knowledge abstraction, and guide computing models and algorithms. However, it's to be noted that each one data available within the sort of big data aren't useful for analysis or deciding process. Industry and academia have an interest in disseminating the findings of massive data. This paper is basically a small review of big data and it's architecture build with drawbacks and benefits.

Block diagram of big data analytics



Explanation of Big Data Architecture:

Big Data systems involve more than one workload types and they are broadly classified as follows:

1. Data Sources

The data sources involve all those golden sources from where the data extraction pipeline is built and therefore this can be said to be the starting point of the big data pipeline. For example:
(i) Datastores of applications such as the ones like relational databases

(ii) The files which are produced by a number of applications and are majorly a part of static file systems such as web-based server files generating logs.

(iii) IoT devices and other real time-based data sources.

2. Data Storage

This includes the data which is managed for the batch built operations and is stored in the file stores which are distributed in nature and are also capable of holding large volumes of different format backed big files. It is called the data lake. This generally forms the part where ourHadoop storage such as HDFS, Microsoft Azure, AWS, GCP storages are provided along with blob containers

3. Batch Processing

All the data is segregated into different categories or chunks which makes use of long-running jobs used to filter and aggregate and also prepare data o processed state for

analysis. These jobs usually make use of sources, process them and provide the output of the processed files to the new files. The batch processing is done in various ways by making use of Hive jobs or U-SQL based jobs or by making use of Sqoop or Pig along with the custom map reducer jobs which are generally written in any one of the Java or Scala or any other language such as Python.

4. Real Time-Based Message Ingestion

This includes, in contrast with the batch processing, all those real-time streaming systems which cater to the data being generated sequentially and in a fixed pattern. This is often a simple data mart or store responsible for all the incoming messages which are dropped inside the folder necessarily used for data processing. There are, however, majority of solutions that require the need of a message-based ingestion store which acts as a message buffer and also supports the scale based processing, provides a comparatively reliable delivery along with other messaging queuing semantics. The options include those like Apache Kafka, Apache Flume, Event hubs from Azure, etc.

5. Stream Processing

There is a slight difference between the real-time message ingestion and stream processing. The former takes into consideration the ingested data which is collected at first and then is used as a publish-subscribe kind of a tool. Stream processing, on the other hand, is used to handle all that streaming data which is occurring in windows or streams and then writes the data to the output sink. This includes Apache Spark, Apache Flink, Storm, etc.

6. Analytics-Based Datastore

This is the data store that is used for analytical purposes and therefore the already processed data is then queried and analyzed by using analytics tools that can correspond to the BI solutions. The data can also be presented with the help of a NoSQL data warehouse technology like HBase or any interactive use of hive database which can provide the metadata abstraction in the data store. Tools include Hive, Spark SQL, Hbase, etc.

7. Reporting and Analysis

The insights have to be generated on the processed data and that is effectively done by the reporting and analysis tools which makes use of their embedded technology and solution to generate useful graphs, analysis, and insights helpful to the businesses. Tools include Cognos, Hyperion, etc.

8. Orchestration

Big data-based solutions consist of data related operations that are repetitive in nature and are also encapsulated in the workflows which can transform the source data and also move data across sources as well as sinks and load in stores and push into analytical units. Examples include Sqoop, oozie, data factory, etc.

Benefits or advantages of Big Data:

The benefitsor advantages of big Data:

massive knowledge analysis helps in understanding and targeting customers.

➡It improves health care and public health with availableness of record of patients.

➡It helps in money mercantilism, sports, polling, security/law social control etc.

➡Any one will access Brobdingnagian info via surveys and deliver answers of any question.

Drawbacks or Disadvantages of big data:

The drawbacks or disadvantages of big data:

➡Traditional storage will call for huge value or amount of cash to store big data.

➡Lots of data in the big data storage is unstructured so it becomes hectic .

➡Sometimes big data analysis violates the privacy policies.

➡they can be used to manipulate the data of a person.

➡Increase in the stratification is seen mostly.

➡Big data analysis n brief terms is not useful much as it does not show much detailed results.

➡Big data results are not always accurate they can be deceiving as well.

➡some figures can get miss matched because of the speedy updates.

CONCLUSION

In recent years huge and huge amounts of data is being generated day by day. Analyzing these data is difficult for a person with no knowledge to the current finished paper, we tend and try to survey and understand the various analysis problems, challenges, and tools accustomed analyze these large knowledge.From this survey, we are able to overcome and understood that every big data platform has its own way or mechanism of concentration on a unit. It works in a variety some are designed for execution in real time and some that give good analytic response within fraction of time. Techniques used in them are totally different that give analysis that are applied in mathematics analysis, machine learning, processing, intelligent analysis, cloud computing and data streaming processes. In near future these techniques can be used much efficiently and people can use the data more efficiently and effectively.

Future scopes

These days every other application is generating massive amount of data that is both structured and unstructured .and in the near future big data will be able to stored ten times more data than it is storing right now and a faster rate. Hopefully in the future the hadoop might get more features and the data can be much easily recorded, monitored and combine all kinds of data around us. New tools will be useful and will be easier to extract information like analysing, tracking and auditing, sharing, managing our data. For advancement of any sector whether it is education , marketing , healthcare, telecommunication, sports etc the polishing of big data that will be used to mange everything is necessary.

Bibliography / 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) (2018), 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) (2018), 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 (2017), 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) (2016), pp.2561-2573.

[8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285 (2014), pp.112-137.

[9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics:

current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1 (2014), pp.114-126.

[10] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, A look at challenges and opportunities of big data analytics in healthcare