International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Special Issue - NGEPT - 2019

Big Data Analytics: Tools and Techniques

Shefali Arora, Assistant Professor, EPCET, Bangalore, India, shefali.cse@gmail.com

Abstract- The big data analytics technology is a combination of various techniques and processing methods. And the property that makes big data effective is their collective use by enterprises to obtain relevant results for real time implementation and strategic management. The radical growth of Information Technology has led to several complimentary conditions in the industry. One of the most persistent and arguably most present outcomes, is the presence of Big Data. The term Big Data is a catch-phrase was coined to describe the presence of Huge amounts of data. The resultant effect of having such a huge amount of Data is Data Analytics. Data Analytics is the process of structuring Big Data. Within Big Data, there are different patterns and correlations that make it possible for data analytics to make better calculated characterization of the data. This makes data analytics one of the most important parts of information technology.

Keywords —Big Data, Big Data techniques, Data Analytics Tool, Security Issue, Privacy Issue

I. INTRODUCTION

In today's world, many organizations are utilizing the innovation to store and examine petabytes of information about their organization, business and their clients. Enormous information is a term connected to the information sets whose size is past the capacity of regularly utilized programming frameworks keeping in mind the end goal to store, oversee and prepare. The enormous information handling and examination have turned out to be basic to the vast majority of the applications like government and endeavors. In the previous couple of years, the aggregate sum of information produced by human has detonated increment 300 circumstances shape exabytes to octabytes. These data are created from various fields like scientific research, government, finance and business, social networks, photography, video, audio mobile phones etc. Data Analytics is the method of structuring Big Data. Within massive information, there are different patterns and correlations that make it possible for data analytics to make better calculated characterization of the data. This makes data analytics one among the foremost necessary components of data technology.

II. BIG DATA OVERVIEW

A. Properties of Big Data

Big Data is an assortment of huge datasets that can't be processed using traditional computing techniques. It is not a technique that can be worked on its own or in isolation; rather it involves many areas of business and technology. The properties of signify Big Data are volume, Variety, Velocity, Variability and Complexity[1] as shown in Fig 1.

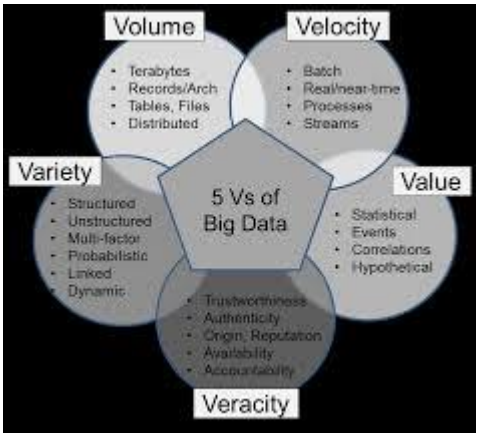

Fig 1. Five V's of BigData

III. BIG DATA TECHNOLOGIES

B. Big Data Key Technologies for Businesses

The big data analytics technology is a combination of several techniques and processing methods. What makes them effective is their collective use by enterprises to obtain relevant results for strategic management and implementation. At this point in the evolution of big data, the challenges for most companies are not related to technology. The biggest impediments to adoption relate to cultural challenges: organizational alignment, resistance or lack of understanding, and change management. Fig 2 shows some Big Data key technologies for businesses[2].

 DOI : 10.18231/2454-9150.2019.0571

Fig 2 Big Data Technologies for Businesses

Predictive Analytics

This is One of the prime tools for businesses to avoid risks in deciding, prognostic analytics will facilitate businesses. Predictive analytics hardware and computer code solutions are often used for discovery, analysis and readying of prognostic situations by process huge knowledge.Such data can help companies to be prepared for what is to come and help solve problems by analyzing and understanding them.

NoSQL Databases

These knowledge bases area unit used for reliable and economical data management across a climbable range of storage nodes. NoSQL knowledge bases store data as computer database tables, JSON docs or key-value pairings.

Knowledge Discovery Tools

These area unit tools that enable businesses to mine huge knowledge (structured and unstructured) that is keep on multiple sources. These sources are often totally different file systems, APIs, DBMS or similar platforms. With search and data discovery tools, businesses can isolate and utilise the information to their benefit.

Stream Analytics

There are situations when the data that an organization needs to process, can be stored on multiple platforms and in multiple formats. Stream analytics computer code is extremely helpful for filtering, aggregation, and analysis of such big data. Stream analytics additionally permits association to external knowledge sources and their integration into the appliance flow.

In-memory Data Fabric

This technology helps in distribution of huge and massive quantities of knowledge across system resources like Dynamic RAM, Flash Storage or Solid State Storage Drives. Which successively allows low latency access and process of huge knowledge on the connected nodes.

Distributed Storage

A way to counter freelance node failures and loss or corruption of huge knowledge sources, distributed file stores contain replicated knowledge. Sometimes the data is also replicated for low latency fast access on large computer networks. These are generally non-relational databases.

Data Virtualization

It allows applications to retrieve data while not implementing technical restrictions such as knowledge formats, the physical location of data, etc. Used by Apache Hadoop and alternative distributed data stores for a period of time or near real-time access to data stored on various platforms, data virtualization is one of the most used big data technologies.

Data Integration

A key operational challenge for many organizations handling big data is to method terabytes (or petabytes) of data in an exceedingly manner which will be helpful for client deliverables. Data integration tools allow businesses to streamline data across a number of big data solutions such as Amazon EMR, Apache Hive, Apache Pig, Apache Spark, Hadoop, MapReduce, MongoDB and Couchbase.

Data Preprocessing

These computer code solutions area unit are used for additional analysis and manipulation of data into a format that is consistent and can be used for further analysis. The knowledge preparation tools accelerate the information sharing process by formatting and cleansing unstructured data sets. A limitation of knowledge preprocessing is that all its tasks cannot be automated and need human oversight, which can be tedious and long.

Data Quality

An important parameter for large data processing is the knowledge quality. The data quality software can conduct cleansing and enrichment of large data sets by utilizing parallel processing. These softwares are widely used for getting consistent and reliable outputs from big data processing.

 DOI : 10.18231/2454-9150.2019.0571

C. Popular Solutions and Techniques for Big Data Analytics

Genetic algorithms

Generic algorithms revolve around using mechanisms that are based on evolution – in this case, the evolution of business.[3] Problems are optimised in a way that evolves solutions. Simply put, genetic algorithms use a process that mimics biological evolution. Few examples where genetic algorithms are used: to schedule which doctors will be working in emergency rooms at any given time, to develop content like jokes from artificially creative sources and to create business processes that mimic the buyer buying process.

A/B split testing

It's a very useful technique for web designers, for example. It can provide you with insight on a way to improve certain features of your product, so that your consumers respond more positively to it. It is a technique that compares a control group with a variety of test groups, in order to get best results. It can be used to determine which treatments are the best, or what kind of layouts, images or colors to use on a website or product package.[4] Since we are talking regarding huge Data and Knowledge, the results you receive are very substantive and statistically vital. Also, if there are more variables that are simultaneously manipulated during testing, it can be often referred to as "A/B/N testing

Association rule learning

It is a good technique for increasing sales, or forming client incentive programs. Association rule learning consist of using various techniques for the purpose of discovering interesting relationships, or to be more precise "association rules".[5] When you have a large database of purchase histories, you can ascertain which of the products are most commonly purchased together. One of the surprising discoveries made using this technique is that shoppers who buy diapers also tend to purchase beer.

Edge analytics

Edge analytics is relatively new and it is still developing, but once it is perfected it will revolutionize the way we process big data. Basically, the information is analyzed as soon as it is collected, so you immediately have a complete analysis. This can be really useful for security cameras, so that irrelevant data is no longer stored, or for navigation devices, etc.

Outsourcing

Outsourcing is another important technique for obtaining the necessary business results. If you are a small business owner, and need a market analysis, hiring more people will only drain your budget. You can realize the people who already have the necessary equipment, and who can analyze the market or media for you. It is a KPO type of outsourcing, so make sure that you find someone who is really competent in the requested field.

IV. BIG DATA TOOLS

Here are some of the top tools used to store and analyze big data:

Apache Hadoop

Apache Hadoop may be a java primarily based free software system framework that may effectively store a great amount of data during a cluster. This framework runs in parallel on a cluster and has a capability to permit United States to process knowledge across all nodes. Hadoop Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute across many nodes in a cluster. This additionally replicates knowledge during a cluster therefore providing high availability everytime.

Microsoft HDInsight

It is a giant knowledge resolution from Microsoft high-powered by Apache Hadoop which is accessible as a service within the cloud. HDInsight uses Windows Azure Blob storage because the default classification system. This also provides high availability with low cost.

NoSQL

While the standard SQL are often effectively accustomed handle huge amount of structured data and knowledge, we need NoSQL (Not Only SQL) to handle unstructured data. NoSQL databases store unstructured data with no particular schema.[6] Each row can have its own set of column values. NoSQL gives better performance in storing massive amount of data. There are several ASCII text file open-source NoSQL DBs available to analyse massive Data.

Hive

This is a distributed data management for Hadoop. This supports SQL-like question possibility HiveSQL (HSQL) to access massive data and knowledge. This can be primarily used for Data mining purpose. This runs on top of Hadoop.

Sqoop

This is a tool that connects Hadoop with various relational databases to transfer data. This can be effectively accoustmed transfer structured data to Hadoop or Hive.

PolyBase

This works on top of SQL Server 2012 Parallel Data Warehouse (PDW) and is used to access massive knowledge and data stored in PDW. PDW is a dataware housing appliance built for processing any volume of relational data and provides an integration with Hadoop allowing us to access non-relational data as well.

Big data in EXCEL

As we know, many people are comfortable in doing analysis in EXCEL, a popular tool from Microsoft, you can also connect data stored in Hadoop using EXCEL 2013. Hortonworks, which is primarily working in providing

 DOI : 10.18231/2454-9150.2019.0571

Enterprise Apache Hadoop, provides an option to access big data stored in their Hadoop platform using EXCEL 2013. Similarly, Microsoft's HDInsight allows us to connect to Big data stored in Azure cloud using a power query option.[7].

Presto

Facebook has developed and recently open-sourced its Query engine (SQL-on-Hadoop) named Presto which is built to handle petabytes of data. Unlike Hive, Presto doesn't depend upon MapReduce technique and might quickly retrieve data and knowledge.

V. CONCLUSION

In conclusion, Big Data is already being used to improve operational efficiency, and the ability to make informed decisions based on the very latest up-to-the-moment information is rapidly becoming the mainstream norm. There's no doubt that Big Data will continue to play an important role in many different industries around the world. It can definitely do wonders for a business organization.

ACKNOWLEDGMENT

I am very grateful to East Point College of Engineering and Technology for organizing National Conference and giving me the chance to publish the paper in International Journal.

REFERENCES

[1] Paul Quinn, "*Big genetic data and its big data protection challenges"*, Elsevier volume 34, Issue 5, October 2018

[2] Xuewei Li, Xueyan Li, "*Big Data and its key Technology in Future"*., IEEE vol. 20, Issue 4, July/Aug. 2018

[3] Munawar Hasan, "*Genetic Algorithm and its Application to Big Data Analysis"*, IJSER, Volume 5, Issue 1, Jan2014 ISSN 2229-5518

[4] Amir Gandomi, Murtaza Haider," *Beyond the hype: Big data concepts, methods, and analytics*", ELSEVIER, International Journal of Information Management 35 , pp. 137–144, 2015.

[5] Bakshi, K.: "*Considerations for Big Data: Architecture and Approaches*". In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)

[6] David Loshin, "*From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph*", eBook ISBN 9780124186644, 30 Aug, 2013

[7] Nada Elgendy and Ahmed Elragal, "Big Data Analytics", German University in Cairo (GUC), Egypt, Aug 2014

[8] Ibrahim Abaker et al, "The rise of big data on cloud computing: Review and open research issues", ELSEVIER, Information Systems 47, pp. 98–115, 2015

 DOI : 10.18231/2454-9150.2019.0571