International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Analysis of Big Data Technology and its Challenges

Dr Ajay Pratap

Assistant Professor, AIIT, Amity University Uttar Pradesh, Lucknow, UP, India

---

Abstract - *In recent decades, the internet application and communication have seen a lot of growth and reputation in the Information Technology sector. These internet based applications and related communication are generating the large size and different variety of difficult multifaceted structure data regularly which is known as big data. This is the era of massive automatic data collection, systematically obtaining various measurements without knowing relevancy of data. For instance, E-commerce transactions may include the data of online buying, selling, investing and exploration based activities which has very complex structure and high in dimensional as well. The traditional data storage and management techniques are not adequate to store and analyses these high volume data. In this survey paper we are discussing the big data architecture for data analysis. Paper also provides the overview of the big data challenges and various technologies to handle the Big Data.*

Key Words: Data Analysis, Big Data, Architecture, High Volume Data, Unstructured Data, Semi-Structured Data

1. INTRODUCTION

In today's era people and system are using the web based application and it causes generation of large size of data in an exponential manner. Exabyte (EB) and Petabytes (PB) are the measuring units for the size of data. This amount of data growth is due to advancements in the fields of communication, digital sensors, computations, and storage of data for business analytics. The Big Data term had been devised by a researcher, Roger Magoulas. In last few decades, data has been grown in massive order in various fields and moreover multiple types of data are also emerging. As per the statists of International Data Corporation (IDC), the overall data volume created in the world was 1.8ZB for year 2011. This volume is approx. nine times larger within next five years. Data or Information will be the fuel for 21st century as all possible domains such as health care, marketing, disease control & prevention, smart city and business intelligence applications are going to generate and use a large amount of data and information.

If we compare big data concept with other traditional datasets and its processes, big data includes semi structured and unstructured data. Big data technology and its analytical processes are used to provide the description about massive datasets generated from various real time systems. Big data is also useful for getting details about new prospects for determining new values, for in-depth understanding of the hidden values, and also some real time situations such as organizing and manipulating big datasets exponentially. The volume of data and information is growing rapidly and also some challenging issues. Big data visualization process is another vital problem in big data analytics. Improvement in the field of Information Technology has generated more and more data on daily basis that causes the problems of gathering and integrating data from distributed data sources. Some other upcoming technologies such as cloud computing, Internet of Things (IoT) and Data Centers also promote the growth of data. Cloud computing technology provides the standard for storing and retrieving the data from the big data assets. IoT technology is based on use of sensors to collect and transmit the data to be stored and processed in the cloud storage.

2. BIG DATA ANALYTICS

Knowledge Discovery in Databases (KDD) refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. Fundamentally, data processing is seen as the collecting, processing, and management of data for producing new information for end users [8]. Big Data analysis is done in four steps which are Acquisition, Assembly, Analyze and Action which are known as 4 A's of big data analysis.

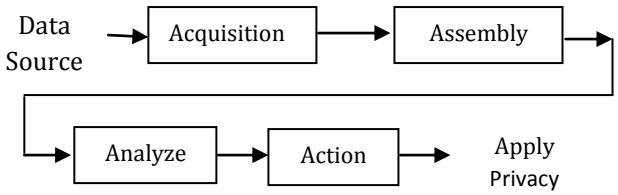

Fig-1: Steps of Data Analysis

2.1 Acquisition Process

In Big Data architecture, the acquisition component has to obtain high speed data from various data sources which deals with various access control protocols also. In some conditions of generation of data are important, and sometimes capturing the metadata and storing them with the corresponding data is also important for further analysis. It is where a filter could be recognized to store only data which could be helpful or underdone data with a lesser degree of uncertainty [9].

2.2 Assembly Process

This is the second step of analysis where the data of structured or semi-structured category have to cleaned, brought into the computable mode. After this they are

integrated and stored at desired location. This is a sort of extraction, transformation and loading (ETL) process done on data. At this point the architecture has to deal with various data formats and must be able to parse them and extract the actual information like named entities, relation between them, etc [9]. Complete cleaning of data in big data environment is not entirely guaranteed.

2.3 Analysis Process

This is the point where we have run queries, perform modeling, and building new algorithms for new cases. Analyzing of data is done by mining technique which requires integrated, cleaned, trustworthy data.

2.4 Action Process

In this step we interpret the results obtained in Analysis step to tale some valuable decisions for the organization or enterprise. At the same time it is also very important to understand and verify outputs obtained at user's end.

2.5 Privacy

Privacy of big data and results obtained by analysis process are very important issue and not putting much effort on privacy may cause many serious problems at the analysis of data and sometimes at the creation of data also. To overcome with these problems the big data architecture must have following properties:

- Big data infrastructure be linear scalable
- Must be able to handle high throughput multi formatted data
- Auto recoverable
- Fault tolerant
- Higher degree of parallelism and distributed data processing

3. BIG DATA ANALYTICS INFRASTRUCTURE

Fig-2 depicts the overall layer based infrastructure of big data analytics. Various layers present in the big data analytics are as follows:

- Data Layers
- Analytics Layer
- Integration Layer
- Decision Layer
- Data Governance Layer

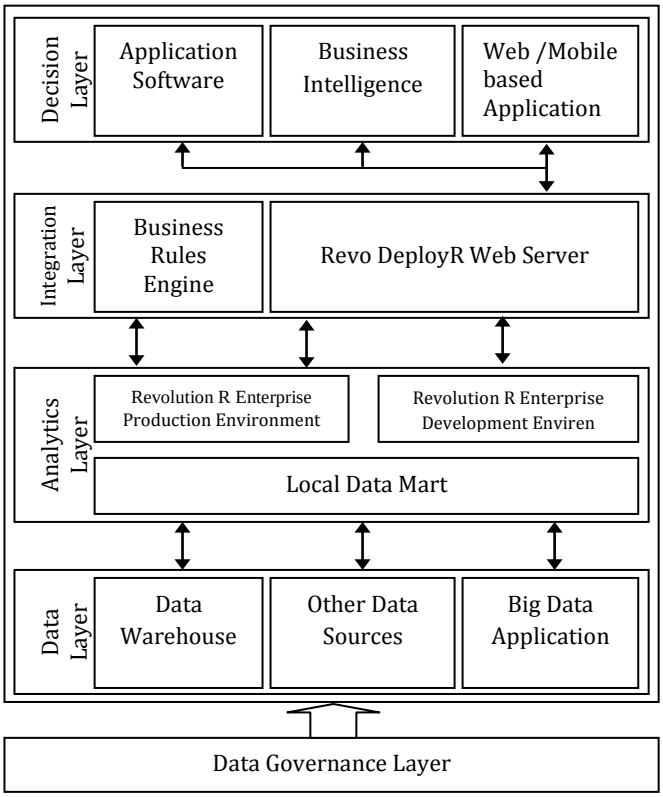

Fig-2: Big Data Infrastructure for Business Analytics

3.1 Data Layer

This layer consists of structured data, semi-structured and unstructured based data. Unstructured data is stored using NoSQL databases such as MongoDB and Cassandra. Example of unstructured data and semi-structured data are data streamed from the web world, social media domain, IoT sensors and other operational systems. We can use software tools such as Hive, HBase, Storm and Spark in this layer.

3.2 Analytics Layer

In this layer, we can implement the dynamic data analytics and then deploy the real time values. This layer has building a model developing environment and it keeps on modifying the local data in regular interval. Layer is responsible for improving the performance of the analytical engine.

3.3 Integration Layer

End user applications and analytical engine are being integrated in this layer. This layer includes rules engine and an API that is being used for dynamic data analytics.

3.4 Decision Layer

This layer includes interface applications for end user such as mobile app or desktop applications for interactive

web applications and business intelligence software. In this layer, people interact with the system to perform some decision based task.

3.4 Data Governance Layer

Data governance is a collection of processes, policies, roles, and standards that ensure the effective and efficient use of information which enables the organization to achieve its goals. This layer includes all key data governance functions, such as business semantics glossary, reference data accelerator and data stewardship manager.

All five layer described above are associated with different sets of end users in real time and enables a crucial phase of real time data analytics implementation.

4. BIG DATA TECHNOLOGIES FOR DATA ANALYTICS

Big data management includes the organization and manipulation of huge volumes of structured data, semi-structured data and unstructured data. Main goal of big data management is to take care of three major factors which are quality of high level data, availability of data for business intelligence and big data analytics applications.

Various tools are available for big data management from data acquisition, data storage to data visualization. In this section we will describe major big data management tools related to data analysis.

4.1 Hadoop

This is an open source platform for handling big data and its analytics. Hadoop is user friendly and provides flexible environment to work with different data sources. Tasks such as gathering various sources of data or accessing of data from a database in order to run process-oriented machine learning process can be done using this platform. This tool also provides different types of applications such as location based data from weather, traffic sensors and social media data.

4.2 Map Reduce

This environment permits larger jobs implementation scalability against group of server. Map Reduce implementation has two major tasks:

(a) The Map Task: It converts input dataset is into a different set of value pairs.

(b) The Reduce Task: It combines several outputs of the Map task to form reduced tuples.

4.3 PIG

It is an analytical tool that attempts to make the Hadoop closer to the developers and business users. PIG permits the query execution over data that is stored on a Hadoop instead of a SQL.

4.4 WibiData

This tool was developed for the enterprises to personalize their customer experiences. It is combination of web analytics with Hadoop. It works as a top layer over the HBase and allows the websites to explore better processes with their user data. It allows real time responses to user, such as recommendations and data related to personalized content, and decisions.

4.5 Hive

Hive permits predictable business applications to run SQL queries against a Hadoop cluster. It provides SQL-like bridge that was developed by Facebook and, then it has been made open source. Hive is a high level perception of the Hadoop and it allows all to make queries against data stored in a Hadoop storage medium easily.

4.6 Rapidminer

It provides an integrated platform for various things such as machine learning, data/text mining and other data analysis task such as predictive analytics and business analytics. Rapidminer is used for both business and commercial applications as well as for education, research, training, application development, and rapid prototyping. It supports all steps of the data mining process which includes dataset preparation, validation, results visualization and optimization.

5. BIG DATA CHALLENGES

There are many crucial challenges need to be focused while handling of Big Data and its analytical process [6]. Some of those challenges are being discussed:

5.1 Storage Related:

As we know that the size of secondary storage device in computing devices is in the range of Terabytes (TB). The amount of data produced via internet is also of very large amount and it is measured in terms of Exabyte (EB) So the traditional relational database management software such as Oracle and MySQL are not able to store or process such kind of Big Data. To give the solution for such issue, databases uses NoSQL based databases such as Cassandra and MongoDB which can handle unstructured and semi-structured data.

5.2 Data Life Cycle Management

This process decides the selection of data for storing purpose which is used for the analytical processes. There are many challenges and one of them is that the existing storage system is not able to support such huge amount of data. Therefore, an effective model is needed that makes the life cycle management system better.

5.3 Data representation

Main goal of data representation is to make data more important for data analytics and user analysis. Many datasets have definite levels of heterogeneity in terms of their structure, semantics, type, organization, granularity and accessibility. Improper data representation technique may reduce the value of the data originality and even disturbs effective data analysis process [23]. Hence effective data representation is highly required for easy analysis process.

5.4 Redundancy Reduction and Data Compression

Redundancy is one of the major problems in the field of database system and analysis. Redundancy reduction and data compression are operational method for decreasing the cost of the system by reducing the data redundancy and data compression.

5.5 Analysis

Big data is generated from various types of online activities and transactions which vary in structure and the volume. Data Analysis is very difficult in case of high volume data. To handle this situation, a special scalable architecture is used to process the data in a distributed environment. Data are distributed into fragments and handled in a number of computers systems in the network and then processed data is combined manner.

5.6 Reporting

Displaying the statistical data in the form of values is known as Reporting. In case of high volume of data traditional reporting tools and methods become challenging to understand and implement. In these situations the statistical reports must be represented in such a manner so that can be easily understood.

5.7 Data Confidentiality

Data confidentiality is another big issue for big data because the service providers and owners of the data are not able to maintain and analyze such huge datasets in effective manner. In this situation they depend on professionals or sometimes third-party tools for the analysis of datasets, which may cause potential safety risks. So, data confidentiality is important issue for the researchers.

6. CONCLUSION

During the composition of this survey paper, various concepts of big data including the concepts of big data analytics, big data analytics techniques, data visualization and big data analysis algorithm have been studied. The literature survey also gives the overview of the possible research opportunities in big data environment. Some of them are as follows:

(i) Applying scheduling methods for handling big data computation in cloud based platform

(ii) Issues related to data privacy and data security in big data environment.

(iii) Reduction of data redundancy and data compression in big data environment.

(iv) Handling independent and identically distributed variables in big data

REFERENCES

[1] H.V. Jagadish, D. Agarwal, P.Bernstein, Challenges and Opportunities in Big Data, The Community Research Association, 2015.

[2] K. Krishnan, Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence, Elsevier Science, 2013.

[3] Vallabh Dhoot, Shubham Gawande, Pooja Kanawade and Akanksha Lekhwani, Efficient Dimensionality Reduction for Big Data Using Clustering Technique, Imperial Journal of Interdisciplinary Research (IJIR), Vol-2, Issue-5, 2016, ISSN: 2454-1362 3.

[4] Gantz J, Reinsel D, Extracting value from chaos.IDC iView, 2011, pp 1–12

[5] Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A survey, Mobile New Applications 2014, 171-209.

[6] Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Quart.2012; 36(4):1165–88.

[7] K. Krishnan, Data warehousing in the age of big data, in the Morgan Kaufmano series on Business Intelligence, Elsevier Science, 2013.

[8] Kitchin R. The real-time city? Big data and smart urbanism. Geo J. 2014, 79(1), pp: 1–14.

[9] Katrina Sin and Loganathan Muthu, Applications of big data in education data mining and learning analytics – A literature Review, ICTACT Journal on soft computing

special issue on soft computing models for big data, July 2015, Vol:05, Iss: 04, pp: 1035-1049

[10] Mayer-Schonberger V, Cukier K, Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.

[11] Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A Survey , Computer Science Review, 2015, Vol: 17, pp: 71-80

[12] K. Davis, D. Patterson, "Ethics of Big Data: Balancing Risk and innovation", O'Reilly Media, 2012.

[13] Mike Barlow, Real-Time Big Data Analytics: Emerging Architecture, ISBN: 978-1-449-36421-2, 2013

[14] Shuhui Jiang, Xueming Qian, Tao Mei, Yun Fu, Personalized Travel Sequence recommendation on Multisource Big Social Media, 2016, IEEE Transactions on Big Data,Vol.2, Issue:1 2.

[15] http://www.techrepublic.com/blog/big-data-analytics/10- emerging-technologies-for-big-data