ISSN 2277-2685
IJESR/December 2014/ Vol-4/Issue-12/972-975
Aparna U.R et al./ International Journal of Engineering & Science Research

AN ANALYSIS THROUGH THE AVAILABLE DATA MINING PROCEDURES WITH BIG DATA

Aparna U.R*[1], Shaiju Paul[1]

[1]Dept. of Computer Science, Jyothi Engineering College, Jyothi Hills, Panjal Road, Vettikkattiri, Cheruthuruthy, Thrissur, Kerala, India.

ABSTRACT

Big Data can be termed as a popular term which is used to represent massive amount of data that is having exponential growth. This data is huge in size and the volume. As the amount of the data increases in the volume, so does the complexity related to the extraction of the relevant information from the data sources. A larger amount of data needs to be searched to get the relevant information. This large amount of data should be properly processed maintaining the features such as security, volume, availability ,nature etc of the data. Data mining can be termed as the process of extraction of relevant data's from a large collection of data. Though there are a large number of methods for the present data mining task to be performed, the big data applications has grown in such a way that the currently available technologies are unfit to process the data as per the needs of the user within the instance of the time.

Keywords: Data Mining, Map-Reduce, Hadoop, Big Data.

1. INTRODUCTION

As the population increases day by day, so does the technology and its requirements. A large collection of data's are produced and these need to be warehoused. In the past, data's were generated from a single source and are delivered to many. But now the situation is different. Each and everyone of us is employed directly or indirectly in the sending as well as receiving of the data from the huge collections of the data. A large amount of data needs to be analyzed properly. This huge collection of the data's beyond our imagination can be termed as the Big Data. This data can be structured, semi-structured or un-structured. Big Data can be represented by using 3 V's-velocity, variety and volume.

Volume-Volume of the data's are increasing day by day due to more participation in the data search process.

Variety-Data's can be of different types. They may be static or streaming. The data's vary in their representation. Some of the data may be audio, while others may be video. The formats and the structure of the data may also be different.

Velocity-Velocity refers to the speed of the data delivery process.

Acquiring of a specific information from this large amount of data is a difficult task. It deals with actions such as extracting data from a particular source, harness the relevant data and to analyze that information. This process of data mining should also consider the important parameters like cost reduction, time complexity, space complexity, optimizations, smarter decision making criteria etc.

2. BIG DATA

Big Data deals with huge data with heterogeneous and diverse dimensionality. Each bit of information varies with the schemata and the protocols specified for their representation. These information may be from autonomous sources .As these data are distributed with decentralized control, they are susceptible to different types of attacks. As the volume of the data increases, the complexity of dealing with the data also increases rapidly. Data required for the entire society changes irrespective of the space and the time. So a proper data

mining is required to provide the needed data to the user while still maintaining the factors dealing with the security, complexity and the performance.

3. METHODS USED

3.1 Data Mining With The Big Data

Big data can be termed as a data repository which needs to handle huge amount-say peta byte or more of the data. It is having data which are heterogeneous and are decentralized. These data need to be managed properly. Proper orchestration of the data should be done in the management process. The data should be readily available to the user ,while maintaining the security and the confidentiality . Several challenges that are related to the big data are meeting the need for the speed, dealing with the outliers, addressing the data quality, and displaying meaningful data after analyzing the data.

Data mining can be termed as the process of extracting information from a huge collection of the data. Extraction of the data from the huge collection of data of heterogeneity, structured, semi-structured and unstructured representation, which are from the autonomous sources with decentralized control while still maintaining the relationship of the data is a difficult task.

Open challenges in the process of data mining include understanding optimal analytical system, monitoring and managing end to end quality of service parameters, provisioning data center cloud resources for real time analytics and ensuring end-to –end security and privacy [1].

3.2 Data Accessing And Computing Procedures

As the population increases day by day, so does the technology. Data grows rapidly overtime to fit into available memory. If we are dealing with small amount of data, then a single desktop computer can meet our demands. But this is not the case when dealing with massive amount of data. Presently available solutions are to perform parallel computing or collective mining [6]. Different data mining algorithms can also be used to make the process easier. Division of the tasks to be performed among several personal computing and aggregating them can provide a better approach in the data mining.

Different programming models are proposed. Of which mostly accepted model is Map Reduce. This programming model is being applied to many machine learning and data mining algorithms [4]. Map Reduce programming model is being applied to work with multi-core processors. In this model, huge amount of the data sets could be sub divided into smaller portions and are assigned to the Mapper nodes. Summation operation on these data can be done to get intermediate results. Finally summation in parallel can be done on the reduce nodes. Further extensions of Map Reduce algorithm can be found in multi-core and multiprocessor systems.

Map Reduce programming model is used as an implementation mechanism in Hadoop. Extensions to these approaches can be seen in integration of R(open soured statistical analysis software) and Hadoop, Weka and Map Reduce, Hadoop ML etc.

But map reduce algorithm posses a lot of disadvantages. There are certain cases when map reduce is not at all a suitable choice. This includes real time processing. It is not always very easy to implement each and everything as Map Reduce program. Map Reduce is not suitable for large number of online transactions etc.

Google I/O 2014 conference reimages the Map Reduce and launches the Data Flow. It is a managed service for creating data pipelines that ingest, transform and analyze the massive amount of the data, upto the Exabyte levels.

Spark is a general purpose engine designed to work under workload faster. It adds some important innovation features like stream processing ,fast fault recovery ,language integrated API, optimized scheduling, data transfer and more. Spark exploits the considerable amount of RAM that is spread across all the nodes in a cluster.

3.3 Security

Big data deals with large amount of data [5]. So a proper management of these data should be done in order to maintain the security. Different security management applications which are readily available are:

1) Usage of Kerberos: This is a popular mechanism that is employed in Hadoop. Here validation for the security is done before entering the node. It also considerers the validation based on the requests.

2) Encryption at the file levels or the operating system levels.

3) Providing certification s to the data.

4) Validation during the deployment process.

5) Usage of communication protocols.

Special communication protocols need to be developed which ensures the security as well as data mining mechanism.

Most of the security tools that are readily available are not achieving the fundamentals to be dealt with in the big data. These architectural specification are not specific to the big data. But many of Big data applications are working on these security measures.

3.4 Framework

A coordinating and structural framework is required to avoid the duplication of the applications and the stored information. The framework should focus on choosing the right data, data modeling and the analytics and data organization and interpretation [2]. A commonly used framework is Hadoop. But Hadoop has missing encryption at the storage and the network layers. The framework is written almost entirely in JAVA, creates implication in various security breaches, its vulnerable in nature and is subjected to many potential scalability issues.

3.5 Domain of The Data

Domain of the data sets provide essential information for the data mining .The big data mining should consider the prediction for the growth of the data. Online streaming feature, synthesizing of the high frequency rules from different data sets, spreading of the behavior in online social network experiment should be clear.

There exist a wide technology gap between the industrial application and the decision makers. Decision makers should get enough information about the technologies in order to extract the data. So unearthing of the information should be done.

3.6 Usage of The Algorithm

As parallel computing is a popular mechanism in the big data mining task, one has to consider the difficulties in aggregating these data sources to get the desired result. Different data mining algorithms may be employed for different data sources. This creates a problem when we are trying to aggregate the data. So information exchange should be done in such a way that all the distributer data sources should be able to achieve the data mining task. These operations should be done without creating loopholes in the integrity considerations of the data.

As big data is a collection of uncertain and sparse data, local learning is a difficult task. So the mining of the complex and the dynamic data should be done by implementing proper data models. The relationship among the data values are another important factor. Some of the information can be of no privacy concern. But others within the same relation may be of crucial status. The syntax and the semantics of different data in different sources are also different. So extraction of the data should be done with giving importance to the application knowledge. Some data may also streaming in nature.

4. CONCLUSION

As data expand in volume and the need for the data abruptly increases, the concept of the big data emerged. The heterogeneous and complex nature of the data increases the complexity of the extraction process. As technologies which promote better data mining are emerging day by day, the streaming data's complexity also increases. The big data's unintended consequences include the managing and the analysis of the data, maintaining the quality of the data from the multiple sources, maintenance of the legality of the collection

process, ensuring of the quality of service parameters, complex mix of the pattern matching during the extraction, cost and the delay time [3]. A proper programming abstraction needs to be implemented for big data analytics. The tools for the processing should be developed in such a way that provide interoperability. The frameworks for the big data analytics should be integrated.

REFERENCES

[1] Ranjan R. Streaming Big Data Processing in Datacenter Clouds. IEEE Cloud ComputIng published by the IEEE Computer society, 2014.

[2] Tekiner F, Keane JA. Big Data Framework, IEEE International Conference on Systems, Man, and Cybernetics, 2013.

[3] Wigan MR, Clarke R. Big Data's Big Unintended Consequences, IEEE Computer Society, 2013.

[4] Padhy RP. Big Data Processing with Hadoop-MapReduce in Cloud Systems. International Journal of Cloud Computing and Services Science (IJ-CLOSER) 2013; 2(1): 16-27.

[5] Schell R. Security-A Big Question for Big Data. IEEE International Conference on Big Data, 2013.

[6] Wu X, Zhu X, Wu G-Q, Ding W, Data Mining with Big Data",IEEE Transactions on knowledge and data Engineering 2014; 26(1).

