Journal of Engineering and Applied Sciences 13 (9): 2575-2579, 2018
ISSN: 1816-949X

Divergence of Managing Scalability and Unstructured Data in Big Data Analytics

[1]D. Viji, [1]R. Lavanya, [2]D. Hemavathi, [1]P. Saranya
[1]Department of Computer Science and Engineering,
[2]Department of Information Technology, SRM University, Chennai, India

Abstract: The promise of data-driven decision-making is now being recognized broadly and there is growing enthusiasm for the notion of "Big data. Heterogeneity, scale, timeliness, complexity and privacy problems with big data hinder advancement at all stages of the channel that can form assessment from information. The promising visualization of big data is to facilitate organizations will be capable to produce and tie together every byte of related information and use it to make the preeminent decisions. Big data technologies not only support the ability to collect large amounts but more importantly, the ability to understand and take advantage of its full value. Traditional data mining techniques are deals with structured, homogeneous and small dataset.But today =s perspective major characteristics of big data are heterogeneity. In big data mining heterogeneity data set have to accept and deal with following types of data like structured, semi structured even though, fully unstructured data simultaneously. In this study, an interesting idea given about partitioning to handle the heterogeneity data. First it helps to determine whether the given dataset is fully heterogeneity or not. Then the given dataset is accordingly partitioned into several homogenous subsets. Finally, a specialized model for each subset is constructed and narrated with various features.

Key words: Big data, heterogeneity, scalability, complexity, analytics, heterogeneity

INTRODUCTION

Big data is a collection of large and complex data sets, so, it becomes very difficult to analyze in traditional data processing system (Agrawal *et al*., 2012). What is big data how it is differ from the small data: small data from within one organization sometimes stored in file. But big data spread throughout electronic space like social networks data, Facebook, Twitter, Skype, etc. These social network website are generating a huge amount of data day by day. Small data structure contains highly structured data. Mostly big data's are unstructured or semi structured data (Berkovich and Liao, 2012; Fan, 2013). Big data are nothing but collection of small datasets. So, large and complex to maintain as well as analyze those data.

MATERIALS AND METHODS

Dimensions of big data: Normally big data are charact erized by 3 V's: Volume, Velocity, Variety. Now newly two attributes are found to know better about big data. We will now discuss about all five of these attribute to know better about big data that is illustrated in Fig. 1.

Volume: Volume is nothing but considered the size of the data. Big data volume much larger than the traditional type of data or small data. Big data volume much larger than terabyte, petebytes and these data are generated at like, Facebook, Twitter and Skype (Gu *et al*., 2014).

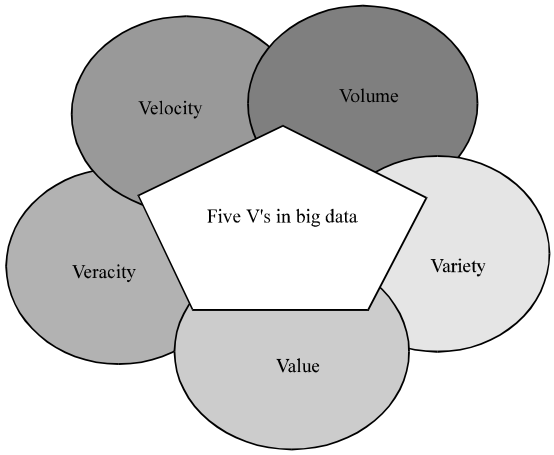

Fig. 1: Dimensions of big data

Velocity: Velocity refers to the speed of the data processing. Data is being generated at very high speed since data streams are arriving continuously. In big data handling the data streams coming at high velocity is a very big challenge (Prakash and Hanumanthappa, 2014).

Corresponding Author: D.Viji, Department of Computer Science and Engineering, SRM University, Chennai, India

Variety: Big data collection of data from various sources that include both structured and unstructured data. What is structured, unstructured and semi structured data? How they differ from each other? Structured data refers to data that has fixed length and format for the data. This type of data well defined how the data will be stored and accessed. Structured data are managed by traditional relational database and spreadsheets. Semi-structured data lies between structured and unstructured type of data. Semi structured data also describe as self describing structure, since, it does not form formal structure of data models. But it contains tags such as meta data. Unstructured data usually refers information that does not follow any structure of data. It is quite opposite of structured data. Unstructured data often include text, audios, videos, images, etc. About 80-90% of the data in the any organization is unstructured data only. They are highly dynamic and do not have any particular format. Amount of unstructured data growing very faster than structured data (Tsai *et al*., 2015).

Veracity: Big data are collecting data from the several sources and maintained those data. But those data may contain the noisy. Big data and analytics technology now allows us to work with these types of data so, we need to clean and eliminate that noise and maintained the quality of data for the further use. The volumes often make up for the lack of quality or accuracy (Lee *et al*., 2014).

Value: Value is the most important V of the big data. Now a day's big data has a cost, since, those data can be sold and hiring money to the organizations. It is important that businesses make a business case for any attempt to collect and leverage big data (Sharma *et al*., 2015).

Challenges in big data analysis: Big data has attributes like volume (huge size), verity (heterogeneity) and velocity (growth speed) which make it valuable but also creating big dispute in this field. Such as capturing, storing, managing, analyzing and understanding useful information out of unstructured data. Converting unstructured data into structured format is a major challenging process in big data mining. Now a days, big data analysis not only considered the size of the data as well as need to be considered the various characteristics like heterogeneous, unstructured data, incomplete, noisy and erroneous. The following challenges are raised under these characteristics (Fan *et al*., 2014).

High dimensionality brings noise accumulation problem and incidental homogeneity high dimensionality and large sample size are combined and creates issues such as heavy computational cost and algorithmic instability Small data are aggregated from the various sources at different time points and then it becomes as large samples in big data. This large samples are creates problem of heterogeneity

Big data is a combination of unstructured and semi-structured data. So, the traditional systems are not able to analyzing data in huge size, high speed and heterogeneity type of data. It is very important as well as challenging task to analyze amorphous data properly and fetching essence of information from that (Jiang *et al*., 2012). Big data are coming from various sources, e.g., social media, smart phones, sensors, traffic updates, etc. And in different formats like text, audios, videos, web logs, etc.

Major dispute in big data extract the meaning full information from the diverse kind of data. Various sources for rising big data with large size and high dimensional include heterogeneity character for example genomics this contain large set of microarray dataset and gene expressions. Similar to the field of genomics an important asset in neuroscience is to aggregate datasets from multiple sources (Subramaniyaswamy *et al*., 2015; Soibelman *et al*., 2008). Another familiar area social network data analysis for example, massive amount of social network data is being produced by Facebook, YouTube, LinkedIn and Twitter.

RESULTS AND DISCUSSION

Heterogeneity/amorphous data in big data analytics: Traditional data mining techniques are deals with structured, homogeneous and small dataset. But today's perspective major characteristics of big data is heterogeneity, i.e., variety of data these data are generated by different sources. Mining from such huge and variety data set is quite difficult. For example, such as academic social network consisting of papers, workshop, conference, universities and companies all contains links such as work-at written by presented by etc.

In big data mining heterogeneity data set have to accept and deal with following types of data like structured, semi structured, even though fully unstructured data simultaneously. Obviously structured data can fit into traditional data base system as well as semi structured data also partially fit in but unstructured data can't fit into the traditional data base system. We need to construct specialized, more complex and multi model system for unstructured data.

To handle heterogeneity data: In this study, Vucetic and Obradovic (2000) an interesting idea given that is partitioning to handle the heterogeneity data. First we have to determine the given dataset was fully heterogeneity or not. Then, the given dataset partitioned

into several homogenous subset. Then we need to construct specialized model for each subset. This partitioning concept very helpful in the process of knowledge discovery from the heterogeneous data very fast and effective manner (Beyer and Laney, 2012; Sun *et al*., 2012). Mining process in the big data also start with data selection from different sources, then data filtering, cleaning, reduction and transformation process will done. Before going to start up these process we need to done pre-processing steps. Pre processing is an important task for heterogeneity data set.

Scalability: Big data volume very huge as well as size also increasing rapidly now a days, so the traditional software tolls can't manage these type of data. Data analysis, retrieval and organization also challenging task due to scalability and complexity of data. Google created a programming model named MapReduce and facilitated by the GFS (Google File System) for reduce this scalability (Che *et al*., 2013) which is a distributed file system where the information is capable of being effortlessly divided over thousands of nodes in a group.

To process a large number data entries in parallel Hadoop MapReduce framework released by Yahoo and other big companies. It utilizes the Hadoop distributed File System (HDFS) B which is available directly through Google's GFS. Users follow the concept of the MapReduce framework to work out a huge number of data access simultaneously and it defines the following two functions, map and reduce.

Hadoop: Hadoop is a Java based software framework for distributed storage and processing of very large data set. Uses commodity hardware. It is based on simple MapReduce programming. Main advantage of Hadoop is easy to handle machine failure. cost also very less. Highly scalable. Core Hadoop has the following two components: Hadoop Distributed File System (HDFS) and MapReduce.

Hadoop Distributed File System (HDFS): HDFS is a distributed file system designed for large datasets to provide high-throughput access which are redundantly stored across multiple machines. One large data file containing billions of records and user wants to access this file frequently for that many queries are submitted simultaneously (e.g., the Google search engine), the traditional file system is not applicable due to the I/O limit. HDFS gives the solution to this problem by partitioning a large file into small blocks or sub sets and store them in different machines. Each machine is called a DataNode. In the Hadoop distributed file system user can able to access the data via. a 'write once and read many's' approach. And also, this system allowed many clients to modified

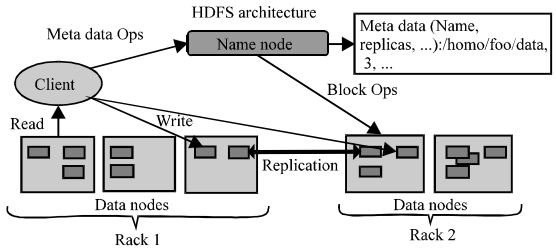

Fig. 2: Hadoop distributed file system architecture

the metadata structures simultaneously. The metadata structures is nothing but the file names and directories. A single machine maintain all the meta data structures that is called as NameNode.

NameNode machine allows fast access to the metadata, since, NameNode only can tracks file names, and the locations of each block of each file. If a client want to do any modification or access a data file have to contact the NameNode, since, NameNode only can retrieves a locations of the data file. The retrived locations identify the DataNodes which hold each block. After identify the data nodes clients can read file data directly from the DataNode servers, possibly in parallel. Under the process of data transfer NameNode is not directly involved, keeping its working load to a minimum. To avoid loss of data HDFS has maintain a builtin redundancy and replication feature. By default each DataNode has three copies. whenever a new DataNode is included in the cluster automatically the HDFS balances its load and it is illustrated in Fig. 2.

If the NameNode itself crashed to avoid loss of the important metadata of the file system we need to safely store the NameNode information by creating multiple redundant systems.

MapReduce: MapReduce is a organized processing model which is worn to processing large datasets in a parallel fashion. The following an example will clearly explain how MapReduce works. For example, user given the following sample input sequence like (DOG CAT RAT CAR CAR RAT DOG CAR CAT) and the task is to write a program that counts the number of words in the sequence. The simplest idea is to read each word from the sequence, add it into a hash table with key as the word and set value to its number of occurrences. If the utterance is not in the hash table so far, then add the word as a fresh type to the hash and allot value to 1. If the utterance is previously in the hash table at that time the assessment augmented by 1. This procedure runs in a serial manner for the entire

Table 1: Comparison of various frameworks of big data

Variables	HDFS	MapReduce	Spark	Storm
Language support	Java	Java, php	Java, Phython, R, Scala	Any programming language
Data processing	Distributed processing	Batch processing	Batch, streaming processing	Micro batching, streaming
Performance	Fault tolerance High throughput access. Load balance	Fault tolerance	Fault tolerance Good performance	Fault tolerance Good performance
Ease of development	Simple and easy to develop	Written in Java easy to develop	Difficult to implement in Java	Easy to program, supports interactive mode, compared with Hadoop
Features	Scalability flexible High availability	Scalability flexible cost effective	Scalability 100 times faster than MapReduce. Handle real time streaming data	Focused on stream processing Operate data in motion data
Limitations	Rough manner Programming model is very restrictive. Cluster management is hard	Not suitable for real time process. Not always easy to implement each and everything in program. Not suitable lot of data to be shuffle over network	Still working out bug, as it matures Operates on data in rest	Core storm does not offer ordering guaranties of message. Duplicates may occur

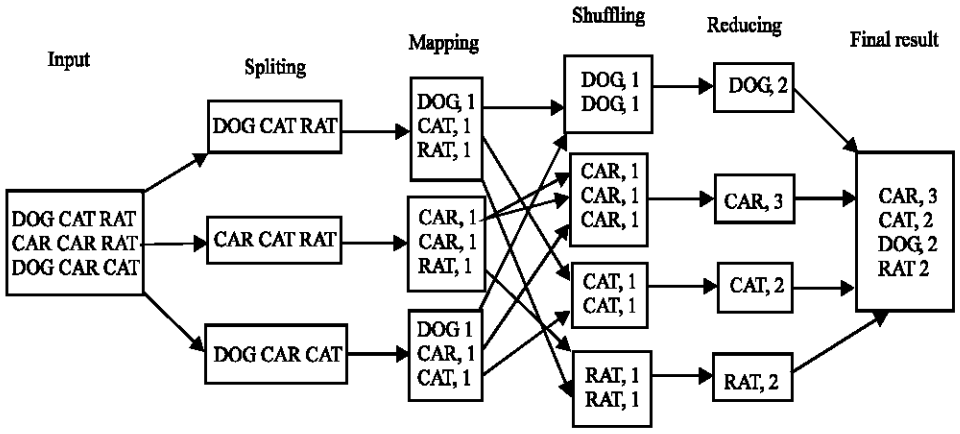

Fig. 3: MapReduce process for word count process

series. The time complexity is majorly based on the length of the sequence. So, dispute will arise to handling count the number of symbols in the whole genomes of many biological subjects. This type of large amount of information definitely takes more time consumption.

The MapReduce paradigm for the word counting task is illustrated in Fig. 3. Mapper executes the assignment of sorting and categorizing and performs the task of summarizing the result. Sorting and grouping two intermediate stages in between the map and reduce stages. First stage of mapReduce was dividing the original sequence into several files (but in this case only one input file). We further split each file into several sub sequences (e.g., three sub sequences in this case) and 'map' the number of each word in each subsequence. The output of the mapper gathering all pairs of the same key value, then finally, the reduce function combine and gives the desired output: #CAR = 3, #CAT = 2, #DOG = 2, #RAT = 2. Thus, the big data features and their applications in detail in the Table 1.

CONCLUSION

In this study, more than a few big data tools were elucidated along with their features of several tasks. Big data provide vastly effective supporting processes for collection of data sets which is too complex and large. In this assessment an assortment of big data tools are exemplified and as well as proved to handle divergence in scalability issues in big data.

This study enhances the possibility of attaining unstructured data features and their applications in detail. Nevertheless, several procedural disputes portrayed in this study ought to be concentrated to this prospective can be recognized completely.

The confront comprises not only the noticeable concerns ofsize, although, heterogeneity, short of formation, fault-management, confidentiality, correctness, attribution and revelation at all phases of the investigation channel from data attainment to outcome analysis.

REFERENCES

Agrawal, D., P. Bernstein, E. Bertino, S. Davidson and U. Dayal *et al*., 2012. Challenges and opportunities with big data a community white paper developed by leading researchers across the United States. Computing Research Association, Washington, USA.

Berkovich, S. and D. Liao, 2012. On clusterization of big data streams. Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications, July 1-3, 2012, ACM, Washington, USA., ISBN:978-1-4503-1113-7, pp: 1-26.

Beyer, M.A. and D. Laney, 2012. The importance of big data: A definition. Gartner Research Company, Stamford, Connecticut.

Che, D., M. Safran and Z. Peng, 2013. From big data to big data mining: Challenges, issues and opportunities. Proceedings of the International Conference on Database Systems for Advanced Applications, April 22-25, 2013, Springer, Berlin, Germany, pp: 1-15.

Fan, J., 2013. Features of Big data and Sparsest Solution in High Confidence Set. In: Past, Present and Future of Statistical Science, Lin, X., ?C. Genest and ?D.L. Banks (Eds.). Princeton University, Princeton, New Jersey, pp: 507-523.

Fan, J., F. Han and H. Liu, 2014. Challenges of big data analysis. National Sci. Rev., 1: 293-314.

Gu, R., X. Yang, J. Yan, Y. Sun and B. Wang *et al*., 2014. SHadoop: Improving MapReduce performance by optimizing job execution mechanism in Hadoop clusters. J. Parallel Distrib. Comput., 74: 2166-2179.

Jiang, K., L. Liu, R. Xiao and N. Yu, 2012. Mining local specialties for travelers by leveraging structured and unstructured data. Adv. Multimedia, Vol. 2012, 10.1155/2012/987124

Lee, D., J.S. Kim and S. Maeng, 2014. Large-scale incremental processing with MapReduce. Future Generation Comput. Syst., 36: 66-79.

Prakash, B.R. and D.M. Hanumanthappa, 2014. Issues and challenges in the era of big data mining. Intl. J. Emerging Trends Technol. Comput. Sci., 3: 321-325.

Sharma, S., 2015. Rise of big data and related issues. Proceedings of the 2015 Annual IEEE India Conference on INDICON, December 17-20, 2015, IEEE, New Delhi, India, ISBN:978-1-4673-7399-9, pp: 1-6.

Soibelman, L., J. Wu, C. Caldas, I. Brilakis and K.Y. Lin, 2008. Management and analysis of unstructured construction data types. Adv. Eng. Inf., 22: 15-27.

Subramaniyaswamy, V., V. Vijayakumar, R. Logesh and V. Indragandhi, 2015. Unstructured data analysis on big data using MapReduce. Procedia Comput. Sci., 50: 456-465.

Sun, Y., J. Han, X. Yan and P.S. Yu, 2012. Mining knowledge from interconnected data: A heterogeneous information network analysis approach. Proc. VLDB. Endowment, 5: 2022-2023.

Tsai, C.W., C.F. Lai, H.C. Chao and A.V. Vasilakos, 2015. Big data analytics: A survey. J. Big data, 2: 1-32.

Vucetic, S. and Z. Obradovic, 2000. Discovering homogeneous regions in spatial data through competition. Proceedings of the 17th International Conference on Machine Learning, June 29-July 02, 2000, Morgan Kaufmann Publishers, San Francisco, California, USA., pp: 1095-1102.