International Journal on Applications in Information and Communication Engineering
Volume 1: Issue 12: December 2015, pp 28-30. www.aetsjournal.com ISSN (Online) : 2394-6237

Optimization Of Big Data

Ggs.Gowtham, S. Arun Kumar

Abstract— Big Data is a large and complex data sets with different sources. With a huge development in the field of networking, data storage and data collection, Big Data is now expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents the optimization features of the Big Data revolution, and explains about HACE theorem,. This also involves the mining and security and privacy issues. We also analyze some important issues regarding Big Data.

Keywords— Big Data, Hace Theorem, DataSet, optimization of big data

I. INTRODUCTION

There is a rise in the use of Big Data applications where collection of data is growing tremendously and is beyond the ability of a common software tools. These Big data helps in managing ,and processing the data within the elapsed time. The main challenge for Big Data applications is to get the large amount of data and extract useful information future. In many situations, the extraction process is to be accurate and close to real time because storing all the data is merely not a possible or an easy task. Let us consider flipkart as the best example of the big data. Flipkart has many products which is being collected from different database sets. It has many users accessing the different items from different places at same time. So, whenever the user searches for a particular item it retrieves the item within seconds of time. All this is done using Big data. It stores the input given by user and extracts it from different datasets and provides the information to the user. This is how the big data is being used in every field these days. The optimization of big data involves some of the things like Tracking the real time of a particular item or product, optimized pricing, Automatic product sourcing etc… can be done. Some other examples of big data applications include troma , Google App Engine etc.. Bog data mining is now playing it's part in cloud computing also. This makes the things easier for retrieving the data and with use of big data in cloud there is no problem of storage.

In section2 I have shown the working and functionalities of HACE theorem to model Big Data characteristics and in Section3 I gave a summary of the issues regarding the Big data. Big Data mining. Related and we conclude the paper in Section4.

GGS.Gowtham, M.Tech - CSE Final Year in SRM University, Ramapuram, Chennai, India (email: ggsgowtham17@gmail.com)
Mr.S. Arun Kumar, Assistant Professor – CSE in SRM University,Chennai, India

II. HACE THEOREM

HACE Theorem. Starts with a large-volume of data sets with autonomous sources and distributed control, and explores a complex and huge relationship among the data..

This is the main characteristic of hace theorem thus making it an extremely challenging for discovering useful knowledge from the Big Data. In the above figure, we see that a number of blind men are trying to size a giant elephant (see Fig. 1), which is a Big Data for these people. The main aim of each blind man is to draw a picture of the elephant according to the information he collects during the process. This is done because each person has his own view of assuming the thing and drawing it. Each blind men can assume the thing independently or the same . To make the problem even more complicated, let us assume that 1) the elephant is growing rapidly and its pose changes constantly, and 2) each blind man may have his own information sources that tell him about biased knowledge about the elephant. Retrieving the Big Data in this way is similar to aggregating heterogeneous information from different sources (blind men) to get a possible picture to reveal the correct picture of the elephant in a real-time. Certainly, this task is not that simple as asking each blind man to describe his own feeling about the thing and then getting an expert to draw one single picture with all the views combined.

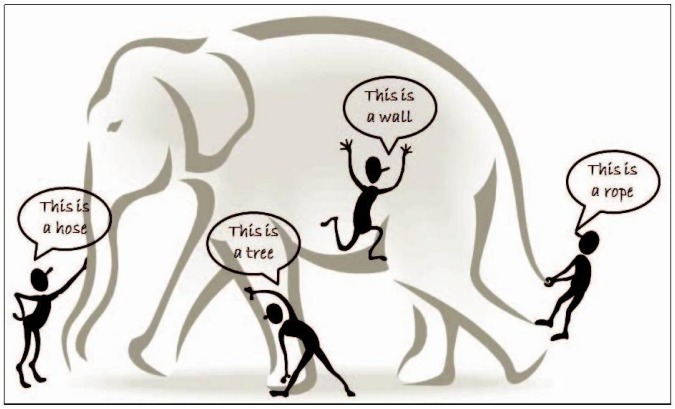

FIGURE 1: HACE THEOREM

All these views combined refers to mining the big data. There are several examples that are related to hace theorem this example is best to get easy understanding of this theorem. This theorem best explains the big data mining. But this theorem faces some issues like privacy, security for the data.

A. Complex and Evolving Relationships

With the increase in use of Big Data, the complexity for the data also increases. In the early stage of data centralized information systems, the focus is on finding best feature values to represent observation. This is similar to using a more number of data fields, such as age, gender, income, education background, and so on, for characterizing an individual. This type of representation treats the individual as an independent entity without considering their relationships, which is one of the most importantfactors.

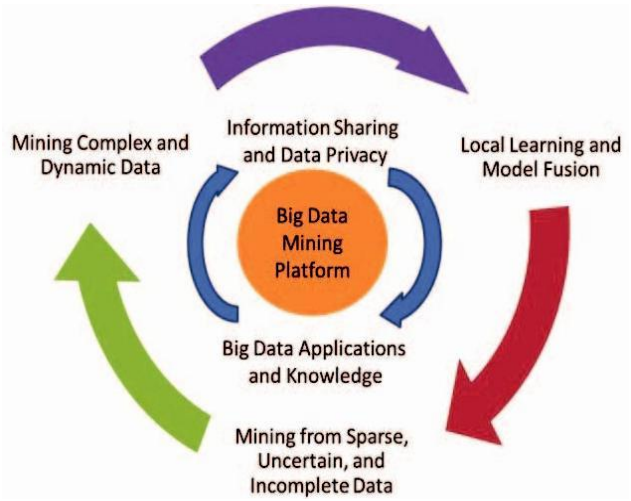

FIGURE 2 : Big Data Processing Framework

III. ISSUES OF BIG DATA MINING

For a database system to handle Big Data, the important thing is to line it up to a large volume of data and for the characteristics of HACE theorem. Big Data processing framework, which includes three tiers on data accessing and computing they are (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III).

The Tier I has an issue that focuses on data accessing and arithmetic computing procedures This happens because Big Data are stored at multiple locations and data sets may have effective access to it, For the tasks that are small scale data mining tasks i.e single desktop computer that contains hard disk and CPU processors, is sufficient to fulfill the goals and mining of data. Many such data mining algorithm are designed for this type of problems. For the tasks that are medium scale data mining tasks, data are a bit large (and possibly distributed) and are not accessible or stored into the main memory. The common solutions for these two are to move to parallel computing or collective mining to sample and aggregate data.

In Big Data mining, the data range is far which cannot be done by a personal computer(PC), a typical Big Data processing will depend on cluster computers which are a high-performance computing , Most of the data mining task being executed by parallel programming tools such as MapReduce on a large number of nodes i.e clusters. In this the role of the software component is to make sure that a data mining process involves a single task. The best example is retrieving the best of the data match with the input query from a database with large amount of records, this task is splitted into many number of small tasks each of which is running on one or many systems.

From the above challenges, many data mining methods have been developed to find from Big Data dynamically changing data volumes. For example, let us consider finding communities and tracing their dynamically evolving relationships are essential for understanding and managing complex systems. Discovering outliers in a social network is the primary step that helps in identifying the spammers accessing it and helps in making the networking safe and better .

Users can now have the solution for the problem simply by purchasing more storage and storage efficiency. However, Big Data complexity is represented in many ways, including complex heterogeneous data types. This shows that the Big Data depends mostly on its complexity.

The emergence of Big Data has also helped in a new computer architectures for real-time data-intensive processing, The main important intensive processing is th open source Apache Hadoop project that runs on an high-performance clusters. The complexity of the Big Data, including transaction and interaction data sets, exceeds a regular technical capability in capturing, managing, and processing these data within reasonable cost and performance.

3.1 Tier I: Big Data Mining Platform

In typical data mining systems, the mining procedures require computing units for analysis and comparisons of the data. It has efficient access to at least two resources 1) data and 2) computing processors. For small data mining tasks, a single desktop computer is sufficient to fulfill the data mining goals. There are many data mining algorithm designed to handle the problems. In a medium scale data mining task the data is typically large and not equal the main memory. The common solution is to depend on parallel computing to carry out the mining process.

For Big Data mining, the data scale is beyond the capacity that a single personal computer (PC) can handle, the Big Data processing framework will depend on a good performance computing, with a data mining task being deployed by running some parallel programming tools, such as Map Reduce that contain large number of computing nodes (i.e., clusters). The role of the software component is to find the best task for performing the data mining, such as finding the best match of a query from a database with billions of records, is split into many small tasks which run on a single or multiple nodes(clusters).

3.2 Tier II: Big Data Semantics and Application Knowledge

Semantics and application knowledge in Big Data refer to user knowledge, and domain information. The two important issues in this are 1) data sharing and privacy 2) domain and application knowledge. The former provides answer to resolve the maintenance of data access, and sharing; whereas the latter focuses on answering questions like "what are the patterns users intend to discover from the data ?"

3.2.2 Domain and Application Knowledge

This section provides an essential information for designing Big Data mining algorithms. The domain knowledge helps in identifying right features for modeling data . The domain and application knowledge can also help us to design business objectives by using the Big Data analytical techniques. For example, In a stock market the data uses these techniques because the data information is bg and huge.al disasters, and so on. The Big Data mining task designs a Big Data mining system for predicting the variations that occur in the market in next couple of minutes. Without correct domain knowledge, it is a clear that finding the relation is such a difficult objective and knowledge is often beyond the mind of some data miners.

A. Existing System

The rise of Big Data applications where data collection has grown and is beyond the ability of commonly used software tools to capture, manage, and process within time. The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the extraction process has to be very effective and close to real time because storing all data is difficult task for any data mining tool. It also provides bigger data. The existing system deals with HACE theorem that explains the analysis of big data.

B. Proposed System

The big data analysis is used for finding solutions to data in large complex datasets. This finding takes a lot of time and most of the time it does not return accurate data and lot of time consumption takes place. Another issue is the privacy of data. It provides all the information of user if multi tasking is not performed correctly. To avoid these problems in this paper we can use ordinal optimization technique and also big data mining.

IV. CONCLUSION

From the real-time applications and key stakeholders and by the national funding companies, managing and mining of Big Data have shown to be a challenging yet very efficient task. The Big Data mining concerns about data sets, These datasets arrangement is explained clearly in the HACE theorem which suggests the key characteristics of the Big Data They are: 1) Heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control. These characteristics suggest that Big Data require a big mind to consolidate data to an extreme extent .

we have analyzed several challenges at the data model, and system levels. To support Big Data mining, high-performance computing platforms are required, which requires a systematic designs of the Big Data. At the data entry level, the information sources and the variety of the data collection environments, results in data with most complicated and diverse conditions, such as missing/uncertain values. In other situations, privacy issues, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is not at all an easy challenge. At the model level, the key challenge is to get global models by combining discovered patterns to form a unique view. This requires a designed algorithm that can easily be used to analyze model correlations between distributed sites, to gain a best model out of the Big Data.

References

[1] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.
[2] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.
[3] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.
[4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, 20-23, 2012.
[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.
[6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.
[7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.
[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.
[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.
[10] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.
[11] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multi-media, (MM '09,) pp. 917-918, 2009.
[12] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data,"
[13] Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.
[14] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms for Influence Maximization in Social Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.
[15] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, "Map-Reduce for Machine Learning on Multicore,"
[16] Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS '06), pp. 281-288, 2006.