Preparing Low Cost Solution Based On Customized Process Of Parallel Clustering Solution

E.Manigandan, Research Scholar
SCSVMV University, Enathur, Kanchipuram
Dr. V.Shanthi, Prof.
Dept. of MCA., St. Joseph's College of Engineering, Chennai
Magesh Kasthuri, Research Scholar
SCSVMV University, Enathur, Kanchipuram

doi: 10.19044/esj.2016.v12n21p159 URL:http://dx.doi.org/10.19044/esj.2016.v12n21p159

Abstract

Big Data analysis is the field of data processing where it involves collections of large volume of data sets which are generally so large and really complex in nature and also there is no unified scientific solution globally for any data analysis due to its nature of difficulties to process them by adopting traditional approaches and technologies. Handling large volume of data and preparing them for deep analysis to evaluate them and prepare required information as required by the mining process is the most complex and sometimes costlier task in real-time. There are many solutions for the data mining process like clustering, special mining, k-means mining to name a few. But the real challenge in data mining process is choosing the correct solution or algorithm to apply for mining the input data and tuning the processing step in such a way that we establish a cost effective solution for the entire mining process. There may be many solutions where mining is efficient but cost of operation is not effective and sometimes it is vice-versa. Hence there is always an ever increasing demand for an efficient solution which is cost effective as well as efficient in data mining technique. The intent of this paper is researching on how we implement a concept called Parallel clustering which gives higher benefit in terms of cost and time in data mining processing without compromising the efficiency and accuracy in expected result. This paper discusses one such custom algorithm and its performance as compared to other solutions.

Keywords: Big Data, Data mining, Clustering, K-means

Introduction

According to Gartner, "Big Data are high-volume, high-velocity,

and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

Due to data volume in data mining, there are real-time challenges which needs to handle all aspects of data manipulation activities like data capturing, data cleansing and filtering, data validation, data storage and retrieval, data search and reporting, data sharing and transporting, data transfer, data analysis, data archiving/purging and visualizing data in various representation of reporting structure defined and expected by the user at the end of processing the data.

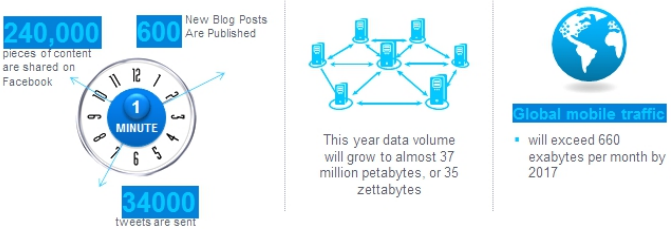

Fig 1: Data growth statistics

In recent times, Big Data attracts significant attention from the IT industry as well as laymen and sales/marketing team due to social media, digital transformation and new trends in digital marketing. This has generated an exponential growth in the popularity of websites from various service providers like Google. For example, Google is estimated to run more than 1 million servers in data centers across the world and to process over 1 billion search requests and approximately 24 PB (petabytes) of user-generated data each day. Amazon as a big retail giant on the other hand handles over 1 million transactions per hour. In the social media space, Facebook, 90 million active users and they generate a massive amount of data in terms of pictures, messages, videos and the like. Even Twitter handles over 0.3 billion tweets on an average per day, which accounts to 4000 tweets per second.

Key issues in Data handling

These giants are struggling with the extremely large volume of data generated due to a huge increase in the number of users and number of data transactions per second. This is exponential growth and handling such Big Data resulting in challenges which are not new [1]: Data Scientists are fighting for decades with the available limitations of the technology in handling such extremely large data sets when applying to various domains like complex physics simulations, genomics, medical science, meteorology and biological and environmental research.

The key issue in data mining in Big data analysis can be classified as

functional issues and technical issues [2]. Below diagram shows the stages and the key functional area of a typical Big data processing system.

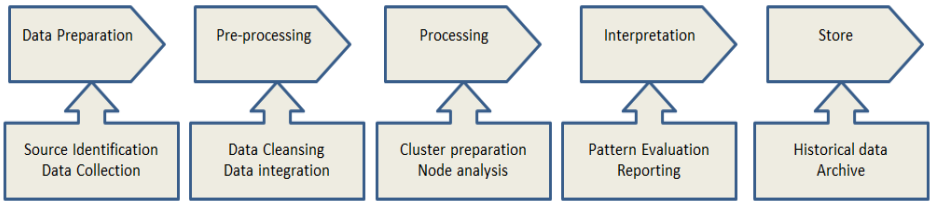

Fig 2: Data processing stages and functions

Functional issues are issues arising in the methodology or process of handling data like mining different kind of knowledge in database, source selection, Interactive mining of information at multiple levels of abstraction, Incepting background knowledge during data selection, query building and filtering options.

Interactive mining

We can even classify another branch of peripheral issues coming out of functional blockers like visualization and presenting reports, pattern evaluation, handling noisy or incomplete data [3]. Technical issues are issues in the structure of the system like efficiency and reliability of the data mining algorithm, handling data processing by parallel or incremental mining algorithm. Apart from this, there are issues arising due to data like handling relational or irrational data and complex data structures, mining information from heterogeneous data source.

This evaluation helps in choosing better data mining solution to produce more accurate results in trend analysis, market evaluation and Return of investment as shown below.

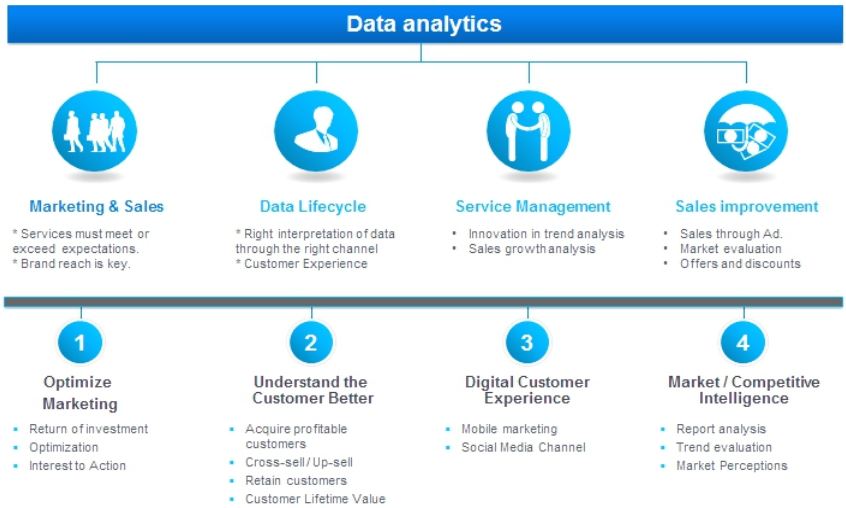

Fig 3: Data analysis solution

Big Data processing is opening up totally new business opportunities like data security. As the world's data volume grows more and more, companies are investing on solutions to handle the challenge of data structuring and its protection from unauthorized access [4]. For example, in a financial industry we can provide a more sophisticated and secured personalized client segmentation. But the challenge is by handling their data to the fullest extent possible, banks are opening the floodgates to new competitors and data security would be a big threat.

Choosing right Mining strategy

The key steps of any Data Mining process for data evaluation and interpretation involves in some key standard steps like Definition of the Business Problem and identifying the key roles of data attributes, Building the Mining Database from the data collected from various source of data sets, Exploring the Data collected from various source and Prepare the Data Modeling structure [5]. This step can go multiple iteration until a concrete set of data model is prepared for the required data processing steps and then user needs to build final Data Model suitable for processing the collected information which can then be used to evaluate the data to prepare reports for pattern analysis.

Spectral clustering is a graph theoretic technique for metric modification such that it gives much more global notion of similarity between data points as compared to other clustering methods such as k-means [6]. It thus represents data in such a way that it is easier to find meaningful clusters on this new representation with inter-connected clusters.

Data mining using Clustering algorithms helps in such a condition where we focus on our analysis area and gather required subset of volumes of data gathered from Lead generation and process them to filter the preference set and produce the required results in terms of reports, diagrams, trend analysis and statistical data points.

Application of Data mining is widely used in Social media and digital marketing area where it requires significant amount of historical information processing which requires higher investment and cost of processing the data volume [7].

We have customized this clustering process by improvising pre-processing steps by targeting Training set to be prepared based on five basic factors of the data set viz attribute type, attribute classification, attribute value range, attribute uniqueness and attribute filtering possibility. This sets the boundary limit of the data processing technique. This customization helps in improvised data model preparation and results in more accurate data evaluation. The steps for a typical customized Clustered data processing model which we named as M-clustering [17] [18] is shown below flow diagram.

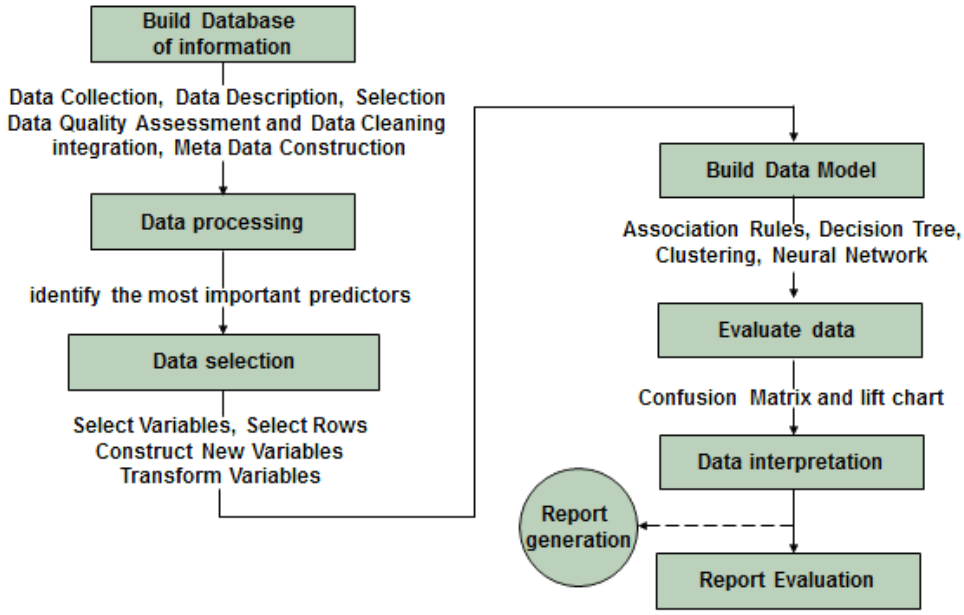

Fig 4: Clustered data evaluation

Though there are predefined spectral clustering technique like spatial clustering or K-means clustering algorithm, we compared it with customized clustering technique and listed below the key differences with standard processing algorithms

Functional aspect	K-means cluster/Spectral cluster	Customized M-Cluster
Clustering Method	Distance based clustering	Tree based clustering (graph)
Data volume handling	Any size of data can be handled	Any size of data can be handled
Data attribute volume	lower number of attributes	large number of attributes
Cluster depth	User specified	Auto calculated
Hierarchical	Depth based	Depth and node based
Cluster assignment	Probabilistic	Ordered node based

Table 1: Evaluation of clustering algorithms

Conclusion

In real-time, Big Data analysis refers to high volume of data to be processed in lesser possible time [12]. In 2001, Gartner Inc. published a key trend analysis research report which exhibits the challenges and open opportunities showing strong and constant or sometimes exponential growth during data processing with increasing data volume and explains a three-dimensional 3V model in the report[10] explaining "Big Data combines high volume, high velocity and/or high variety of information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

There are a lot of researches and investment by various companies

around the world where they practically invested in these areas of data processing and have seen rapid growth in their data volume, from gigabytes (GB) to terabytes (TB) and even to multiple petabytes (PB or 1 million GB) and so on. Therefore, the sheer volume of data which is generated, handled and stored today is the biggest and complex challenge for companies.

This analysis shows that the main focus for a data analytics solution is structuring these massive volumes of data [9]. Customizing Clustering algorithm helps in preparing such structure which is fool proof in producing better trend analysis and report evaluation. According to IDC (a popular the global market research company), the volume of information which was generated and replicated in 2012 exceeded 2.8 zettabytes (ZB) which is equal to 2.8 trillion GB. With such exponential grown and trend, Data Analysts assume that by the year 2020, this numbers will reach an annual growth rate (CAGR) of 42%. This means that it would increase more than 30 times since 2010.

Such data growth and giant data volume is a global phenomenon and we need to focus on key solution to handle such data in more efficient and effective data processing technique which in turn helps in playing a key role for the entire national economy and its citizens, as long as the legal framework is used correctly. This shows that Big Data applications can be used to solve problems that arise when information is distributed to multiple, and variant systems which are not connected by a central element.

One of the costlier tasks in a data mining process is Data cleansing and choosing the right solution/algorithm for data analytics when processing big data analysis. There are several techniques involved in data analytics in a data mining world where the choice of technique depends on various parameters including need for analysis, ROI, data density to name a few.

This article discusses a sophisticated concept for data cleansing which can be considered as pre-processing step for data analytics process. This process is based on relation building based on historical data processing, clustering data model preparation and Data node and tree formation based on a unique and low cost parallel clustering algorithm. This means the purpose of the solution discussed in this paper has customization in the clustering mechanism from improvising the processing logic as compared to special clustering as applied in social networking sites where clustered information provides vital information required for processing customer data.

Also this article highlights the importance and benefits of predicting business challenges and prevents them recurring by use of data mining techniques using a customized clustering algorithm which is detailed in referenced articles [17] [18]. For this approach to succeed it is important to have a strong data classification techniques is needed. Predictive analysis is like

entering a zero surprise zone, where processes function with no surprises and every trigger or alarm is identified well ahead of time. The idea is to collect data by smart mapping (by creating clusters of node and connecting them) in an end to end process and analyzing it to take timely decisions.

References:

Martin, G., & Plaza, A. (2011). Region-based spatial preprocessing for endmember extraction and spectral unmixing. IEEE Geoscience and Remote Sensing Letters, 8(4), 745-749.

Pardos, Z. A., Trivedi, S., Heffernan, N. T., & Sárközy, G. N. (2012, June). Clustered knowledge tracing. In International Conference on Intelligent Tutoring Systems (pp. 405-410). Springer Berlin Heidelberg.

Qin, T., & Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In Advances in Neural Information Processing Systems (pp. 3120-3128).

Dong, J., Xia, W., Chen, Q., Feng, J., Huang, Z., & Yan, S. (2013). Subcategory-aware object classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 827-834).

Nie, F., Zeng, Z., Tsang, I. W., Xu, D., & Zhang, C. (2011). Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. IEEE Transactions on Neural Networks, 22(11), 1796-1808.

Maulik, U., Bandyopadhyay, S., & Mukhopadhyay, A. (2011). Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics. Springer Science & Business Media.

Mirkin, B., & Nascimento, S. (2012). Additive spectral method for fuzzy cluster analysis of similarity data including community structure and affinity matrices. Information Sciences, 183(1), 16-34.

Wang, C., Danilevsky, M., Desai, N., Zhang, Y., Nguyen, P., Taula, T., & Han, J. (2013, August). A phrase mining framework for recursive construction of a topical hierarchy. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 437-445). ACM.

Kumar, R. K., Poonkuzhali, G., & Sudhakar, P. (2012, March). Comparative study on email spam classifier using data mining techniques. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, pp. 14-16).

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In learning analytics (pp. 61-75). Springer New York.

Rangapuram, S. S., & Hein, M. (2012). Constrained 1-Spectral Clustering. In AISTATS (Vol. 30, p. 90).

Tang, J., & Liu, H. (2012, August). Unsupervised feature selection for linked social media data. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 904-912). ACM.

Sim, K., Gopalkrishnan, V., Zimek, A., & Cong, G. (2013). A survey on enhanced subspace clustering. Data mining and knowledge discovery, 26(2), 332-397.
Chitta, R., Jin, R., Havens, T. C., & Jain, A. K. (2011, August). Approximate kernel k-means: Solution to large scale kernel clustering. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 895-903). ACM.
Kang, U., & Faloutsos, C. (2011, December). Beyond'caveman communities': Hubs and spokes for graph compression and mining. In 2011 IEEE 11th International Conference on Data Mining (pp. 300-309). IEEE.
Manigandan.E, Shanthi.V, & Magesh Kasthuri, "Implementing Clustering Algorithms for Data Mining in CRM" in 7th International conference on Advanced Computing and Communication Technologies (ICACCTTM - 2013),ISBN: 978-93-83083-38-1, pp.165-169.
Manigandan.E, Shanthi.V, & Magesh Kasthuri, "Customizing Clustering Algorithm for Data Mining for Lead Generation" in Global Journal for Research Analysis(GRA), Vol.3, Issue 6,2014, ISSN: 2277-8160, pp.40-42