Analysis of E-Commerce Big Data using Clustering and CloudSim Load Balancing

Neha Jain
Dept. Of Computer Science and Engineering
Samrat Ashok Technological Institute
Vidisha (M.P), India

Anil Suryavanshi
Asst. prof. Dept. Of Computer Science and Engineering
Samrat Ashok Technological Institute
Vidisha (M.P), India

ABSTRACT
In this paper an efficient technique is implemented for the analysis of E-Commerce based Applications over Big Data. The Proposed Methodology implemented here is based on the concept of providing Extracting Feature Vectors from the E-Commerce Data and Load balancing of Data using CloudSim based Load balancing and finally Clustered the Data. The Proposed Methodology implemented provides efficient Accuracy & Processing Time as compared to the existing methodology implemented for the analysis of E-Commerce Data.

Keywords
Big-Data, E-Commerce Data, Hadoop, CloudSim, Clustering, Load Balancing, Feature Vectors.

1. INTRODUCTION
Here is an increasing interest in big data technology, driven by the constant growth of information in electronic -commerce companies. It gathers customer data like names, Addresses, preferences, etc. are important firm's assets, but their utilization becomes a more complex and expensive task day by day. On one hand The data volume is so large that its storage and processing using standard programs and devices is very difficult. On the other hand, a significant part of these data is unstructured and their collecting is not easy, but they are exceedingly valuable (for example, likes, tweets, clicks). Nowadays, electronic commerce companies need to analyze a large volume and a great variety of data on products, customers, transactions, and deliveries to increase conversion rate. Companies are aware that scale matters and take into consideration many other implications related to successful implementation of

Personalization of products and services according to customers' needs, improve the process of integration between partners' value chains and finally, dramatically reduce costs [1]. Keeping in inspection the above declared issues, there is required for proposing a system architecture that receives both real-time over and above offline data processing. Such effects for scientific recognizing of transformation of the remote sensed data are significant task [2].

The expression big data refer to the huge quantity of digital information companies and governments bring together about us and our neighboring. This data is not only produced by conventional data exchange and software use through desktop computers, mobile phones and so on, but also from the numerous of sensors of different kinds embedded in different situations, whether in city streets that was cameras, microphones or jet engines that were temperature sensors and the soon-to-propagate Internet of Things, where virtually every electrical device will connect to the Internet and generate data. On a daily basis, we create 2.5 quintillion bytes of data--so much that 90% of the data in the world today has been formed in the last two years on your own. The problems of computing, protection, privacy, storing and methods are all exaggerated by the velocity, volume and selection of big data, for example huge amount of cloud infrastructures range of data sources and formats streaming environment of data acquisition and high volume inter cloud movement. The six-dimensional classification is shown in Figure 1. These six dimensions happen for the key characteristics that are required to find a big data infrastructure. We will explain each of the dimensions in the rest of the document.

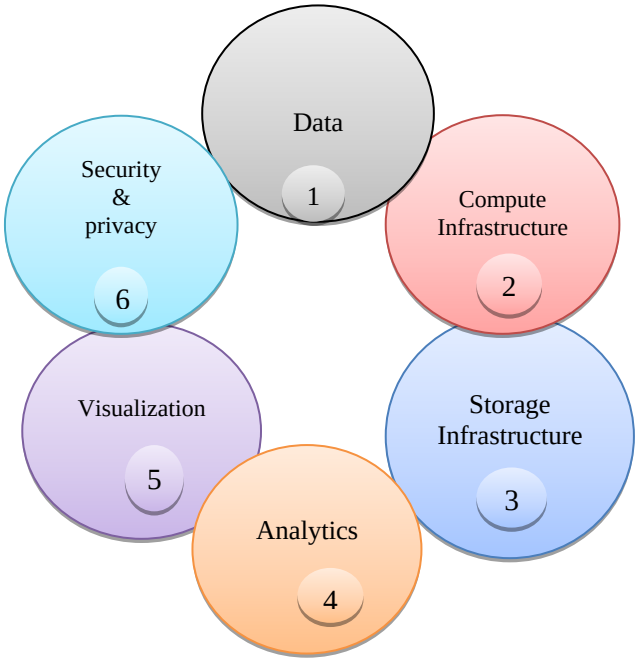

Figure 1: Big Data 6-D Classification

Big Data are frequently produced by online transaction, video/audio, email, and numeral of clicks, scientific data, remote access sensory data, mobile phones, records, posts, social network data and their applications. These data are collected in databases that produce extremely and develop into difficult to imprison, shop, store, control, share, develop, examine, and visualize through characteristic database software tools. With the development in Big Data sensing and computer technology transforms the method remote data

composed, procedures, examined and controlled [3]–[5]. Mainly most recently considered sensors used in the earth and planetary observatory scheme are producing a continuous stream of data. Additionally, common of work have been done in the different areas of remote sensing satellite image data, for example, change detection, gradient-based edge detection, region similarity based edge detection and intensity gradient technique for well-organized intra forecast. The digital world, generating the highest amount of the data continuously, current technology and the tools to store and analyze the large amount of data not an easy task, since it is not able to extract the needed data sets. So there is a need of an architecture that can analyze both the offline data as well as real time data sets. There is an influential benefit in the business enterprise by obtaining the required information from the big data than sample data sets. Day by day the data becoming very large by social networking, online streaming, system logs, mails and remote data, it will be very difficult to compute the massive amount of data. The main problem is how to store the large amount of data, i.e. big data and what data is to keep and what data is to be discarded; extracting the useful data from the big data is the challenging task [6].

Most of the data are generated by the streaming data. In the data stream model, the data will arrive at a very high speed and the algorithms have to process them. This data stream causes several challenges in the design of the data mining algorithms. First, the algorithm has to make use of less number of resources. Second, it can deal with data that can change over time. Resources are managed in an efficient and low cost way, by the green computing [7]. Green computing is the procedure or learns to utilize the computing resources in a well-organized method. Due to the above mentioned heterogeneity and high dimensionality of big data in remote sensing, they also appearance significant computational and statistical challenges associated with dealing out scalability, noise accumulation, spurious correlation, subsidiary endogeneity and measurement errors [8], [9], [10]. These challenges need a novel method for computational and statistical methods to facilitate a deal with big data analysis and processing. The analysis and processing methods are data driven and can advantage from theories and techniques from the areas of statistics, machine learning, pattern recognition, artificial intelligence, data mining, etc. Domain knowledge is an additional essential characteristic that should be closely related to data analysis.

2. LITERATURE SURVEY

In this paper author presents [11] an unsupervised change detection approach for synthetic aperture radar images based on a fuzzy active contour model and a genetic algorithm. The fundamental proposal of this approach is to partition the discrepancy image which is manufactured from multi temporal satellite images into distorted and natural regions. Fuzzy method is a suitable move toward to examine the difference image where regions are not always statistically consistent. Since interval type-2 fuzzy sets are compatible for representation various uncertainties in evaluating to conventional fuzzy sets they are combined with an active contour method for appropriately modeling uncertainties in the discrepancy image. The interval type-2 fuzzy active contour representation is considered to make available beginning study of the difference image by producing transitional transform detection masks. Each transitional alter detection mask has a cost importance. A genetic algorithm is utilized to get the concluding change detection mask with the smallest cost value by developing the realization of transitional change detection masks. An experimental result shows [11] that transform detection, consequences obtained by the enhanced fuzzy active contour model shows less error than earlier approaches. In this paper [12], the author has proposed real-time Big Data analytical architecture for remote sensing satellite application which is utilized to observe real time data with offline data. Primarily, the data are remotely pre-processed which is then understandable by the machines. Subsequently, this valuable data is broadcasted to the Earth Base Station for additional data processing. Earth Base Station completes two types of processing, for example processing of real-time and offline data. In case of the offline data the data are transmitted to offline data-storage mechanism. The integration of offline data-storage device helps in presently practice of the data, whereas the real-time data is directly transmitted to the filtration and the load balancer server where filtration algorithm is employed which removes the valuable information from the Big Data. The proposed design contains three main units, such as 1) remote sensing Big Data acquisition unit (RSDU); 2) data processing unit (DPU); and 3) data analysis decision unit (DADU). First, RSDU acquires data from the satellite and sends this data to the Base Station where primary processing takes place. Second, DPU plays an essential responsibility in structural design for efficient processing of real-time Big Data by providing filtration, load balancing and parallel processing. Third, DADU is the upper layer unit of the proposed design, which is liable for collection, storage of the effects and creation of decision based on the effects acknowledged from DPU. The proposed architecture [12] has the ability of dividing, load balancing and parallel processing of only valuable information. Thus, it affects in efficiently examine real-time remote sensing Big Data using earth observatory scheme. Additionally, the proposed design has the ability of storing incoming unprocessed data to present an offline analysis on fundamentally stored dumps when essential. Finally, a comprehensive study of remotely sensed earth observatory Big Data for land and sea area are offered.

In this paper [13], author has to present a determining analysis on modern advances in methods for hyper spectral image processing, investigation which can effectively agree with the dimensionality difficulty and take into account both the spectral and spatial characteristics of the data. Here the author has focus is on the design of methods competently to agreement with the higher - dimensional environment of the data to deal with required for knowledge-based expansions able to develop a priori data about the spatial understanding of the objects in the prospect with the purpose of balance spectral data. The performance of the above methods is estimated in different investigation circumstances. To assure time-critical limitations in precise applications, they also extend well-organized parallel implementations of some of the discussed algorithms. As mutual these components make available an outstanding snapshot of the modern in those regions and present a considerate perception on potential prospective and promising tests in proposing [13] of robust hyper-spectral imaging algorithms.

Martnez et al [14], based on the foundations of the Lambda architecture, proposed another architecture called SOLID (Service-On-Line-Index-Data architecture): architecture for real-time management of big semantic data which separates the difficulty of big semantic data storage and indexing from real-time data acquisition and utilization. In this architecture,

historical data and real-time data are both stored. It ensures efficient volume management and high processing velocity.

In this paper [15]As McGuire observes, social networking has presented business entities with avenues of customer engagement and helps in providing market opportunities as well as insights for lead generation. The development of social media in marketing platforms has been increasingly used by organizations in building social signals that are very crucial in many SEO digital marketing campaigns. Apparently, the emergence of various media platforms has offered internet markers a broader range of marketing opportunities. The availability of a large quantity of data from online customer interactions could be helpful for online sales personnel in making sales that are more effective. The author identifies a good example, as being a shopping cart abandonment. This provides information concerning the products intended for the purpose of purchasing and the sales personnel can conduct a follow up by using a cell phone in getting the correct information. In addition, sales representatives in identifying the patterns of customers could use real-time updates. This will subsequently help them in selling products at the most appropriate times with the most optimized pricing options. Additionally, it offers opportunities in cross selling and up selling. Streamlined data provided options in market segmentation for peak sales forecasting sales accurately, deploying sales resources. Pipelines and sales forecasts had been based on historical trends. Nonetheless, Big Data analytics could be helpful in precise and relevant information and therefore, the high hypothetical numbers no longer drive sales personnel.

3. PROPOSED METHODOLOGY

The Proposed Methodology implemented here works in following stages:

1. Import an input E-Commerce Big Data.
2. Raw Data Acquisition.
3. Preprocessing of Data.
4. Load Balancing using fuzzy rule based load balancing inspired by ACO.
5. Cluster Data using K-means Clustering.

3.1 Ant colony optimization:

ACO is an evolutionary metaheuristic algorithm based on a graph representation that has been applied successfully to

Solve various hard combinatorial optimization problems. The main idea of ACO is to model the problem as the search for a minimum cost path in a graph. Artificial ants walk through this graph, looking for good paths. Each ant has a rather simple behavior so that it will typically only find rather poor quality paths on its own. Better paths are found as the emergent result of the global cooperation among ants in the colony.

3.1.1 ACO Using Fuzzy rule based Load Balancing:

1. Obtain a problem & represent it as a graph so that it is covered by ants.
2. Assign a heuristic preference to each choice that the ant has to take in each step to generate the solution.
3. Initialize the pheromone value.
4. Define fitness function.

Do for each ant

Calculate the fitness value of the ant fa

/*updating ants best fitness value so far*/

If fa is better than abest then set current value as the new abest

/*updating population best fitness value so far*/

Set gbest to the best fitness value of all ants

5. Repeat until the termination criteria is not met.

3.2 K-Means Clustering

K-means is unique of the meekest unverified education procedures that resolve the well identified bunching problematic. The technique shadows a modest and easy way to organize a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main impression is to outline k centroids, one for each group. These centroids would be positioned in a astute way since of diverse position grounds different result. So, the superior preference is to position them as much as promising far away from each other. Around has been introduced.

Finally, this procedure aims at reducing an objective function, in this case a squared error meaning. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where,

$\left\| X_I^{(J)} - C_J \right\|^2$ **is a selected aloofness degree among a** statistics opinion and the group Centre C_J, is an needle of the aloofness of the *n* information opinions from their individual cluster middles.

The procedure is collected of the subsequent stepladders:

1. Dwelling K arguments into the planetary characterized by the matters that are actually grouped. These arguments epitomize initial assembly centroids.
2. Disperse each article to the assemblage that has the bordering centroid.
3. When all matters have been allotted, recalculate the sites of the K centroids.

Duplication Steps 2 and 3 pending the centroids no lengthier move. These harvests a departure of the substances into assemblages from which the metric to be curtailed can be considered.

4. RESULT ANALYSIS

The Table shown below is the analysis of accuracy and processing time on the basis of their accuracy and processing time. The Proposed Methodology implemented provides efficient Accuracy & Processing Time as compared to the existing methodology implemented for the analysis of E-Commerce Data.

Table 1. Analysis of Processing Time

# of values in Data	Existing Work (in ms)	Proposed Work (in ms)
10000	600	550
20000	800	740
50000	1100	1000
100000	1300	1220
200000	1550	1470
500000	1730	1680
1000000	1900	1800

Table 2. Analysis of Average Accuracy

# of products	Existing Work (In %)	Proposed Work (in %)
500	87.34	91.35
1000	88.12	92.22
1500	88.43	92.65
2000	89.17	92.8
2500	89.32	93.1
3000	90.1	93.43

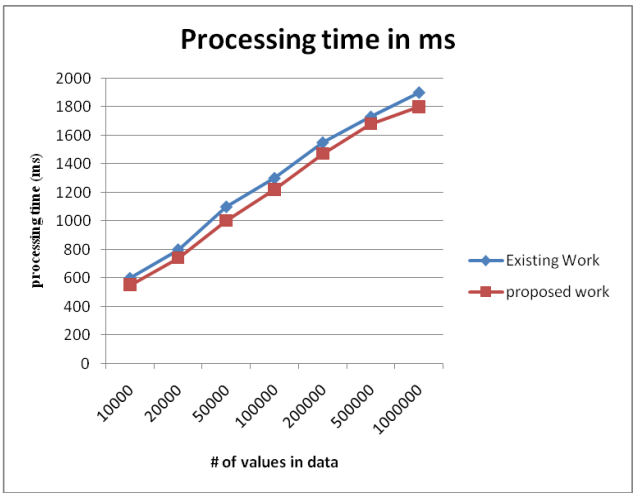

Figure 2. Comparison of Processing Time

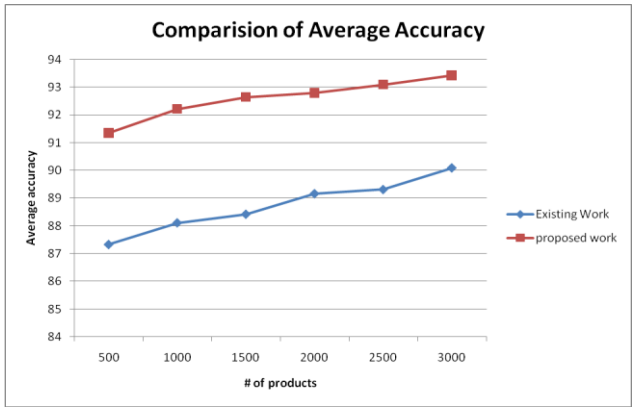

Figure 3. Comparison of Average Accuracy

5. CONCLUSION

The Proposed Methodology implemented here for the analysis of Big Data Applications such as E-Commerce Application using Load balancing and Clustering provides efficient Accuracy of Analysis of Data and Low Processing Time. The Experimental results performed on E-Commerce Data shows the performance of the proposed methodology.

6. REFERENCES

[1] G. Ilieva*, T. Yankova, S. Klisarova, 2015 ," big data based system model of electroniccommerce", Trakia Journal of Sciences, Vol. 13, Suppl. 1, pp 407-413

[2] Cuzzocrea, D. Saccà, and J. D. Ullman, "Big Data: A research agenda," in Proc. Int. Database Eng. Appl. Symp. (IDEAS'13), Barcelona, Spain,Oct. 09–11, 2013.

[3] D. A. Landgrebe, Signal Theory Methods in Multispectral Remote Sensing. Hoboken, NJ, USA: Wiley, 2003.

[4] C.I. Chang, Hyperspectral Imaging: Techniques for Spectral Detection and Classification. Norwell, MA, USA: Kluwer, 2003.

[5] J. A. Richards and X. Jia, Remote Sensing Digital Image Analysis: An Introduction. New York, NY, USA: Springer, 2006.

[6] D. Agrawal, S. Das, and A. E. Abbadi, "Big Data and cloud computing: Current state and future opportunities," in Proc. Int. Conf. Extending Database Technol. (EDBT), 2011, pp.530–533.

[7] R. A. Dugane and A. B. Raut, "A survey on Big Data in real-time," Int. J. Recent Innov. Trends Comput. Commun., vol. 2, no. 4, pp. 794– 797, Apr. 2014.

[8] P. M. Mather and M. Koch, Computer Processing of Remotely-Sensed Images: An Introduction, 4th ed. Wiley, January 2011.

[9] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," National Science Review, vol. 1, no. 2, pp. 293–314, June 2014.

[10] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie, "Remote sensing big data computing: Challenges and opportunities," Future Generation Computer Systems, vol. 51, pp. 47–60, 2015.

[11] Jiao Shi, Jiaji Wu, Anand Paul, Licheng Jiao, and Maoguo Gong, "Change Detection in Synthetic

Aperture Radar Images Based on Fuzzy Active Contour Models and Genetic Algorithms" Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2014, Article ID 870936.

[12] Muhammad Mazhar Ullah Rathore, Anand Paul, Awais Ahmad, Bo-Wei Chen, Bormin Huang, and Wen Ji, "Real-Time Big Data Analytical Architecture for Remote Sensing Application" IEEE Journal of Selected Topics In Applied Earth Observations And Remote Sensing, IEEE, 2015.

[13] Antonio Plaza a, Jon Atli Benediktsson , Joseph W. Boardman , Jason Brazile, "Recent advances in techniques for hyper spectral image processing" Remote Sensing of Environment 113 (2009) S110–S122.

[14] Martnez-Prieto, M. A., Cuesta, C. E., Arias, M., Fernndez, J. D., "The Solid architecture for real-time management of big semantic data", Future Generation Computer Systems, 2014.

[15] McGuire, T. (2013). Making data analytics work: Three key challenges. Retrieved from http://www.mckinsey.com/insights/business_technology/making_data_analytics_work on the 17th April 2015.