An Investigation of Data Management in Real Time Traffic: Agility and Functionality Analysis

B.Lavanya[1], Dr.P.Ganesh kumar[2]

[1]Research Scholar, Department of computer science and Engineering, K.L.N College of Engineering, Sivagangai.
[2]Professor, Department of computer science and Engineering, K.L.N College of Engineering, Sivagangai.
[1]Corresponding Author email: lavanyabalakrishnan89@gmail.com

Abstract: Big data is a large volume data set of dynamically growing data accessed only by supercomputing parallel accessing software. These data are mostly collected anonymously and accessed throughout the database. In this investigation, we cares about the properties of the Big data set with their heterogeneous behavior, anonymity and the complex association characteristics. Also the demands in the prediction of features without compromising the privacy if the data storage. An experimental investigation has also been carried out in this research to find out the estimation time of the nearest K- neighbor miniature with the regular arrangement in the traffic speed of Chennai metro traffic in the peak hours of the weekdays. The uncompromised results show that the nearest K-neighbor takes the highest estimation time especially in the evening peak time.

Index terms: Big Data, Demands, Lineaments, Agility, Traffic Miniature, Nearest Neighbor

I. Introduction

The amount of data transferred and stored per day has been tremendously increased by billions of internet users. These data sets can be process only with high end data processing application software [21]. The basic V^3 approaches associated with Big Data is Velocity, Volume and Variety [1]. In recent times the big data platform is used to mine certain sets of attributes from the whole data set, called predictive analysis depends on the user interactions. Researches show that [2] there is an uncertainty in the large amount of data available.

The usefulness of mining these data will be used in the fields such as clients targeting, resource mapping, crime prevention, disease identification and much more [3]. But, more often, the researchers and big data manipulators are experiencing complications while accessing the big data from the sources of cyberspace, trade information, bureaucracy, health care repositories etc. Due to the evolvement of Internet of Things (IoT) and several cost efficient information sensing mobile devices (smart watches, smart TV’s, Google IO’s etc.) in our day today life, the data sets have grown exponentially. So as that of this growth, worlds per year investment hits the hike for big data repositories [4]. Generally, big data represent a huge amount of complex data sets. Mostly these big data-sets are freely available; these are the digital footprints of the users [5, 6].

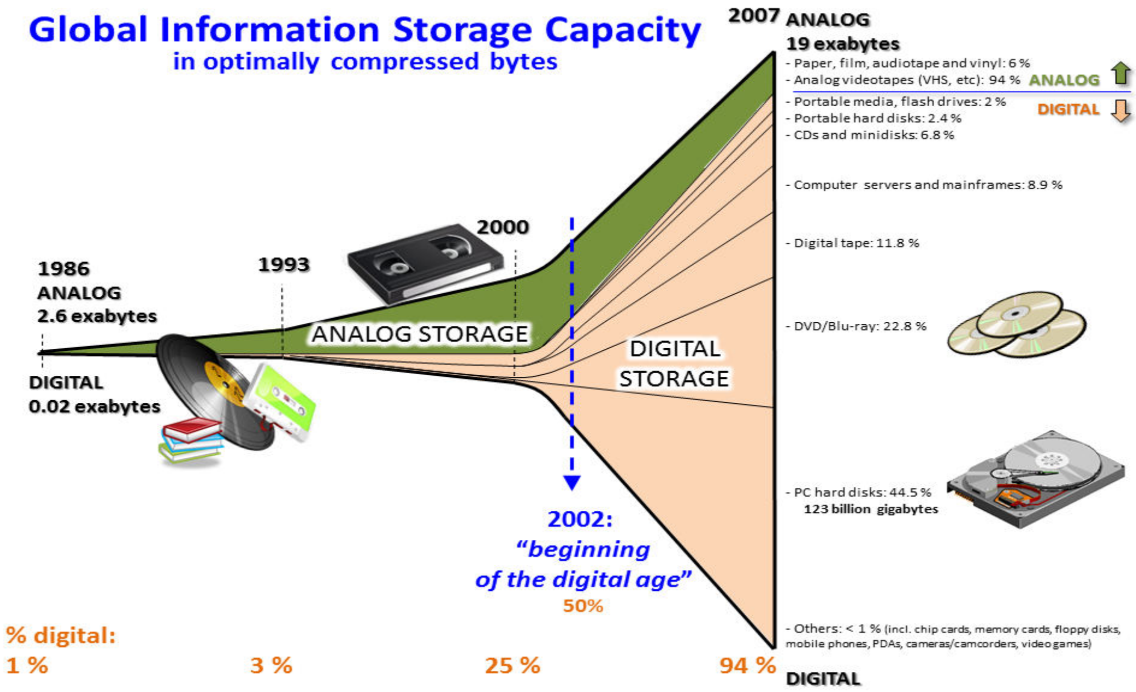

Fig.1. Global Information Storage Growth [18]

The term big data has been coined by Mr. John Mashey [7, 8, 9] in the 90's. The data-sets available with the big data cannot be handled by the usual computers and software applications. It is very difficult to unfold and practice with these data with our general computer machine; it takes a longer time to process. Normally, the big data consist of ordered, disordered and semi-ordered data-sets; anyhow the main convergence goes to the semi-ordered data-sets [10]. The size of the data-sets in the big data increased [11] from a few couple of terabytes to as many petabytes as of 2018. In Order to process these massive, complicated and discrete data-sets, we require advanced integration algorithms and protocols [12]. According to an article by the business intelligence resources of Villanova University [13], a new V, called veracity is included additionally to the V^3.

II. Big Data Representation

The concept of Big Data includes a high quantity of Heterogeneous and Autonomous provenance [14] with shared and disseminated discipline; which inquires the Complex and Evolving correlation between the data-sets.

Because of this complex distinctive, the process of mining the data in these large data-set is very tedious. Xindong Wu, 2014 assuming this big data mining process as follows; a number of blind persons are trying to analyze a big elephant, here the author says the elephant as Big Data. And the motivation of the blind persons is to draw the figure of the elephant by what they observed during the analyzing process. Fig.2. shows the diagrammatic representation of this assumption.

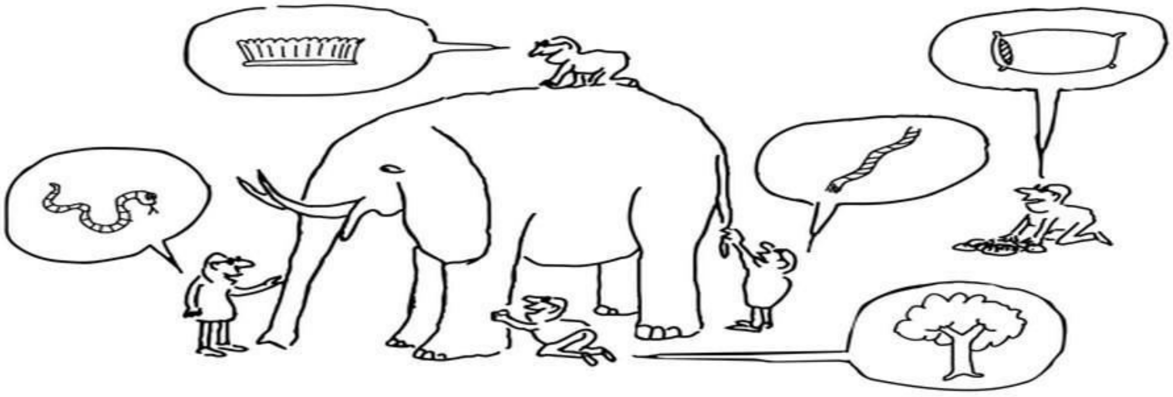

Fig.2. Illustration of blind persons analyzing the the elephant [19] (Big Data).

A) Heterogeneous Aspect

To provide a clear cut explanation of the heterogeneous aspect, we are continuing with the previous assumption that, the estimated conclusion of each blind person is peculiar, as the location they are estimating about the elephant is different [14]. One concludes the elephant’s trunk as snake, other one estimates the leg as tree trunk, the other one tells, tail as a rope and so on. In order to explain the complexity of the Big Data even better, the author hypothesize: a) the elephant is growing rapidly in size. b) Each blind person is having their own perspective of concluding the distinct region. c) The blind person shares their estimation and agrees to other person. Special Case: Each blind person may have different mother tongue and there will be the data security issues while discussing about the data (elephant). Now, we can assume a little bit about the complexity of the big data.

B) Anonymous Aspect

One of the critical aspects of the Big Data is the anonymity of the data with distributed regulation [HACE, IEEE]. Due to this characteristic, the every individual data source is capable of aggregating the data without a common gateway or control center. This is one good advantage for the data-sets; If all the data source are connected to one common control the data security may be highly prone to application dysfunction and malicious attacks. More precisely we can see that, the promotional behavior of Amazon is different for different countries. This is possible with the ordered and semi-ordered data collected from the individual user interaction with Google, Facebook and other social media platforms.

C) Complicated dependencies

We have already briefed a little bit through an illustration of elephant and blind person estimations with special cases. So, we can say that the complexity of the data is directly proportional to the volume of the data-sets in the Big Data [14]. During the mining process, the major objective of the data dependent system is to obtain the best feature values. This can be illustrated easily by taking the top rated social media and online marketing platforms like Facebook, Twitter, Amazon and EBay etc.. these internet platforms collects the user data such as age, single or in relationship, friends circle, hobbies, sex, previous purchased items, previous search history etc.., these data will be stored as Big Data. During the data mining process, the relation between adjacent data will create the complexity. Here in this process, there are two different categories are available, they are representation and relationship of the feature selection. In the feature representation, the two entities are considered as the same, if they correspond to same features. But, in feature relationship, two entities can be connected together even though they are not having anything in common. This feature implies the individuals and the connections represent the friends circle in our real-time world. These data will transform according to the change in spatial and temporal aspects. In order to explore a useful figure from -

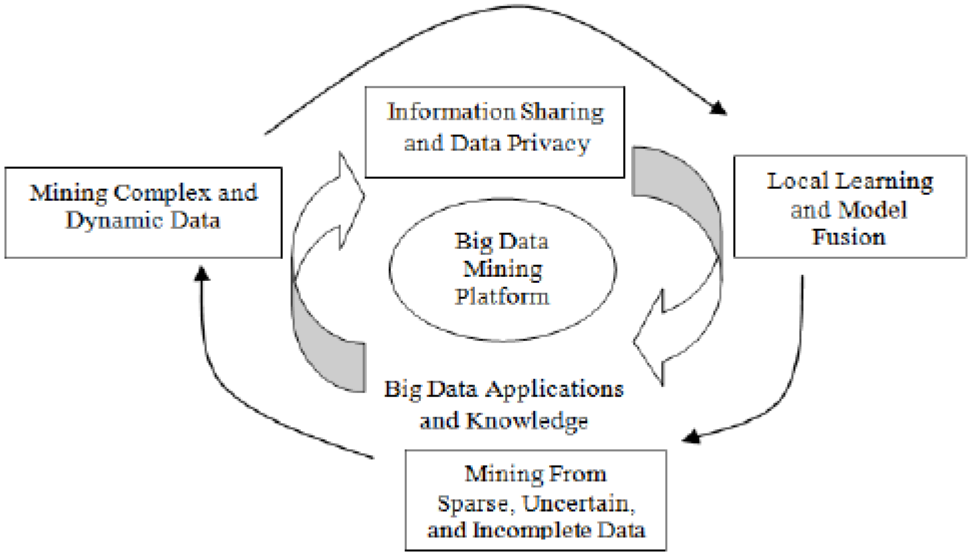
Information Sharing and Data Privacy
Mining Complex and Dynamic Data
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining From Sparse, Uncertain, and Incomplete Data

Fig.3. Processing Structure [20] of the Big Data

This complicated Big Data Ocean, the important key is to find out the non-linear data correlation of the dynamic data-sets.

III. Demands In Mining The Big Data

One of the important goals of any smart learning system is to estimate the features from the large volume of the Big Data sets [15]. However an illustration of the Big Data processing framework has been showed already in Fig.3, this section will briefs the category wise demands to be focused on approaching the data. Therefore the processing structure of the Big Data can be categorized as follows: data approaching and calculating (Category - I), data security and field learning (Category - II) and protocols for the Big Data sets.

A) Platform

The requirement of cluster computing is not a big deal for the estimation of Big data sets, as a complex data mining operation needs critical data inquiry and observations. Hence the data mining operation is executed by certain high level parallel programming software platforms such as ECL called the Enterprise Control Language or the Cluster or the MapReduce programming tool [14]. The actual operation of these parallel programming tools is to estimate a typical query in a Big Data set by means of multiple divided small operations running in single or multiple cluster nodes.

B) Data Interpretation

The information is randomly distributed to various locations physically for a larger volume of data-sets like Big Data; therefore no copies of the data are available in any host region. Hence, there are two constraints are existing for large volume of data-sets they are as follows: a) copies of the data are not stored regionally b) beyond defying the privacy settings, the estimation have to be done with the available data storage. Mostly the enterprises prevents the third parties to access the original data from the data-sets. One of the common key to prevent this data privacy issue is encryption. Several researches have been made to provide the data accessing approach without compromising the data privacy and still it is a blooming research platform for the researchers.

C) Mining protocols

There are several data mining protocol developed by the researchers considering the enhancement of the single peer-knowledge observation method [16], variable origin aspect [17] and data estimation approach. Since the improvement of the real time data availability and the dynamic data accessibility, more efficient data mining approaches have to be framed to achieve the knowledge discovery and variable data mining.

IV. Experimental Investigation

A speed expectation demonstrates is a connection between at least one quality that affect activity speed at the forecast time and prescient speed. Give s a chance to be the present time; Ts is the vector of the qualities that are utilized for anticipating speed, i.e., the vector of pertinent highlights; and the connection $g*$ be the speed expectation display. At that point, $g*(Ts)$ is the prescient speed at the time$(s+o)$, indicated by $h*(s+o)$.

A) Agility Indication

Various kinds of $g*$ can be obtained by utilizing diverse learning techniques or important highlights. The accompanying parts present the highlights and the acquiring protocols. Leverage of receiving k-NN as prescient miniature is that it doesn't require any unequivocal classical constructions. Rather, $k-NN$ sets aside a long opportunity to make an expectation since it needs to scan for the closest neighbors at forecast time. In our examinations, we construct a few diverse $k-NN$ miniature utilizing distinctive arrangements of highlights which are the present and past activity velocities of the objective connection and the neighbor interfaces in its up/down-stream. The exhibitions are analyzed by estimating the normal expectation exactness and the normal time taken to make a forecast utilizing different measures of information.

B) Lineaments

To foresee with high precision, it is essential to choose highlights that are identified with the objective prescient esteem. Fundamentally, the future activity state of the objective connection relies upon the present and ongoing movement conditions. The movement states of neighbor joins associated with the objective connection is likewise identified with the future activity condition of the objective connection. For instance, if neighbor joins toward which the vehicle is streaming—i.e., upstream connections—are congested, downstream connections are likewise bit by bit congested. Thus, if the movement blockage of downstream connections is casual, that of upstream connections will likewise be progressively settled.

C) Representation of $k - NN$ model

Forecast by the k-NN technique does not require any model building. Rather, a model scholarly with the k-closest neighbor (k-NN) calculation has an arrangement of preparing occurrences since it is a sort of nonparametric occasion based learning calculation. It predicts the objective mark an incentive by estimating the separations between the preparation cases and the objective occasion and taking the normal or weighted normal of the name estimations of the preparation cases, which are chosen as the k most comparable occurrences in light of the separation measure.

V. Results and Discussion

This evaluation has been observed in the state of Tamilnadu, especially in the Chennai Metro traffic, in the Adayar Signal from August to November 2017, this information has been gathered from the Chennai traffic police monitoring information repository. The congestion data from October 2017 is taken for validation and the features are tested with November 2017 congestion data.

Analysis Duration in Weeks:

Table - I

Total Weeks	Arrangement		K- NN (Adjacent)		Present
	Day	Week	Day	Week	
2	1.17	18.45	1.57	21.05	1
4	3.18	34.42	3.95	41.74	
6	4.87	40.87	7.45	64.09	

Analysis throughout the Day:

Table - II

Total Weeks	Arrangement		Adjacent		Present
	Day	Week	Day	Week	
2	3.32	2.82	3.29	2.75	3.15
4	2.45	2.63	2.65	2.54	
6	2.65	2.42	2.41	2.33	

Analysis in Weekdays (Monday to Saturday):

Table - III

Total Weeks	Arrangement		Adjacent		Present
	Day	Week	Day	Week	

2	3.29	2.76	3.21	2.68	3.12
4	2.36	2.54	2.58	2.45	
6	2.58	2.37	2.33	2.37	

Analysis in Weekend (Sunday):

Table - IV

Total Weeks	Arrangement		Adjacent		Present
	Day	Week	Day	Week	
2	3.45	3.92	3.61	3.84	2.98
4	3.12	3.24	3.42	3.64	
6	2.98	3.45	2.71	3.63	

Analysis in Peak Hours (8.30 A.M):

Table - V

Total Weeks	Arrangement		Adjacent		Present
	Day	Week	Day	Week	
2	4.32	4.11	4.63	4.08	4.05
4	3.84	3.62	3.23	2.98	
6	3.76	3.42	3.12	3.31	

Analysis in Peak Hours (5.00 P.M):

Table - VI

Total Weeks	Arrangement		Adjacent		Present
	Day	Week	Day	Week	
2	3.87	3.65	3.65	3.45	3.63
4	3.34	3.45	3.23	3.18	
6	3.54	3.79	3.43	3.11	

Overall results for PeakHours

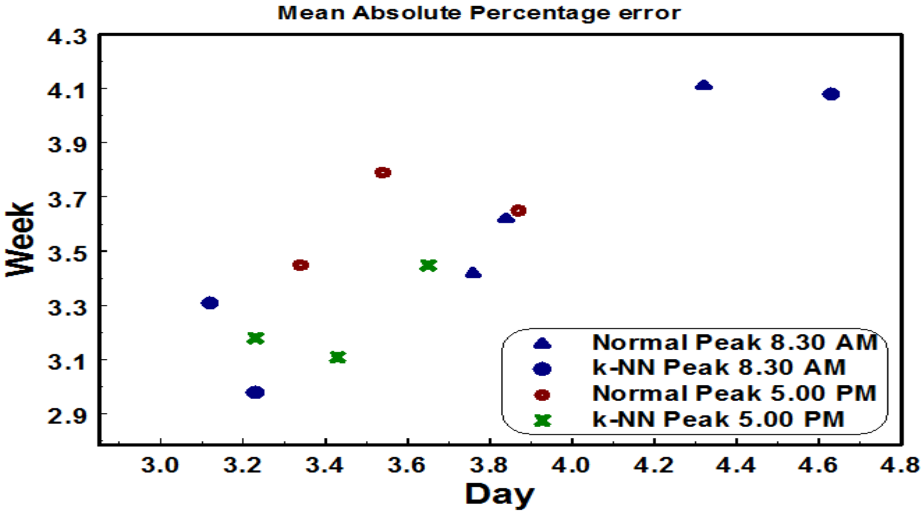

Fig.4. Cumulative MAPE – representation.

So, from the above analysis of the Mean Absolute Percentage Error (MAPE), the nearest neighbor model takes a very long time when compared with to the normal pattern, especially in the evening peak hours.

VI. Conclusion

Because of its usefulness in various applications in the real time data world, big data is becoming a blooming initiative. The term big data itself shows its characteristic large volume and data-set, this paper investigates the big data representation in heterogeneous association and complex evolving aspects. We have provided the demands in the big data regarding the aspects of running platform, security and data interpretation respectively. As a part of this investigation, we have taken the real-time city traffic data base of the Chennai metro traffic signal of Adayar to examine the *k*-NN estimation. Statistical observations were obtained and tabulated with the MAPE - Mean Absolute Percentage Error. From the obtained table it is observed that, *k* closest neighbor elapsed a longer time than the usual feature data-set.

REFERENCES

[1] Laney, Doug. "3D data management: Controlling data volume, velocity and variety." META group research note 6.70 (2001): 1.

[2] Boyd, dana; Crawford, Kate (21 September 2011). "Six Provocations for Big Data". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. doi:10.2139/ssrn.1926431.

[3] "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.

[4] Hilbert, Martin; López, Priscila (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information". Science. 332 (6025): 60–65.

[5] Hilbert, Martin, Part of the University of California course: "Digital Technology & Social Change", Aug 12, 2015.

[6] "Digital Technology & Social Change". Canvas.instructure.com. Retrieved 8 October 2017.

[7] John R. Mashey (25 April 1998). "Big Data ... and the Next Wave of InfraStress" (PDF). Slides from invited talk. Usenix. Retrieved 28 September 2016.

[8] Steve Lohr (1 February 2013). "The Origins of 'Big Data': An Etymological Detective Story". The New York Times. Retrieved 28 September 2016.

[9] Gil Press. "A Very Short History of Big Data". Forbes.com. Retrieved 2018-04-24.

[10] Dedić, N.; Stanier, C. (2017). "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery". 285. Berlin ; Heidelberg: Springer International Publishing. ISSN 1865-1356. OCLC 909580101.

[11] Everts, Sarah (2016). "Information Overload". Distillations. 2 (2): 26–33. Retrieved 22 March 2018.

[12] Ibrahim; Targio Hashem, Abaker; Yaqoob, Ibrar; Badrul Anuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "big data" on cloud computing: Review and open research issues". Information Systems. 47: 98–115. doi:10.1016/j.is.2014.07.006.

[13] "What is Big Data?". Business Intelligent Resources repository, Villanova University. Nov. 2017.

[14] X. Wu, X. Zhu, G. Q. Wu and W. Ding, "Data mining with big data," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014. doi: 10.1109/TKDE.2013.109

[15] J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining," Proc. 22nd VLDB Conf., 1996.

[16] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

[17] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 2, pp. 353-367, Mar./Apr. 2003.

[18] Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025), 60 –65. doi:10.1126/science.1200970.

[19] The Interaction Design Foundation ApS [DK], 6 Blind Men Walk Into a Bar… The UX Punchline, Nov. 2017.

[20] Sr Aghabozorgi, June, 2014, An Approachable Analytical Study on Big Educational Data Mining, An Educational Data Mining Project, 10.1007/978-3-319-09156-3_50.

[21] Wikipedia contributors. "Big data." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 4 Jul. 2018. Web. 10 Jul. 2018.