Research Journal of Computer and Information Technology Sciences ______________________ISSN 2320 – 6527
Vol. **8(1),** 1-8, June **(2020)** *Res. J. Computer and IT Sci.*

A solution approach to big data regarding parameter estimation problems in predictive analytics model

Abdul Alim[1*] and Diwakar Shukla[2]
[1]Department of Computer Science and Applications, Dr. Harisingh Gour Vishwavidyalaya, India
[2]Department of Mathematics and Statistics, Dr. Harisingh Gour Vishwavidyalaya, India
abdulaleem1990@gmail.com

Available online at: www.isca.in
Received 26th October 2019, revised 4th February 2020, accepted 3rd April 2020

Abstract

The existence of big data is everywhere because of social media and business organizations move forwards into online services. Big data is not just a considering volume of data, it is a concept which explains about the gathering, organizing, analyzing the data and extract information from those data sets. Big data analytics concept used in our daily life for various purposes such as weather forecasting, market trends and deals with heterogeneous data. The problem of parameter estimation in big data may be looked upon into three aspects volume, variety and velocity which are known as 3Vs. In big data environment, the users are receiving and sending variety of data (text, images, videos) over the Internet due to it is a challenging task to process and getting valuable solution with minimum data processing speed. In this paper we have picked big data parameters estimation problem and proposed a prediction model to estimate big data parameter based on sampling estimation technique. The model is applicable on dynamic nature dataset. In our proposed method we have applied stratified random sampling techniques for estimate those unknown parameters and compare the result with another sampling techniques.

Keywords: Big data, predictive analytics, big data parameters, data mining algorithms, stratified sampling.

Introduction

A Government organization as well as private sectors companies is getting collected big data sets from different sources. The collected data is not in a single format, it has various formats like text, images, videos, audios, and logs etc. big data analytics has opportunities possible future useful information outcomes. The big data is very complex nature and become challenging task to store, process, and analyze those data. The big data has different parameters such as volume, variety, velocity and more, we need to estimate these parameters and predict relevant information for future use. It can be apply in different areas like health informatics, business, future investment etc. The big data is mostly in high unstructured with divisive nature and to deal with complexity that's why we need to new mining techniques for retrieval of hidden information. The big data analytics offers knowledge which can benefits several applications domain. For example find out marketing trends, political election, geographical information data, sensor-generated data and social media very popular to generate lots of data in every second. The following figure shows the framework of big data analytics.

The Figure-1 has represented the process of big data analytics and knowledge discovery. A large datasets which is continuously increase in the volume and captured by the social media, Internet of Things (IoT), multimedia etc. The data sets format can be either structured or unstructured format, unstructured datasets arises difficulty to process and store it (cleaning, transform, normalization).

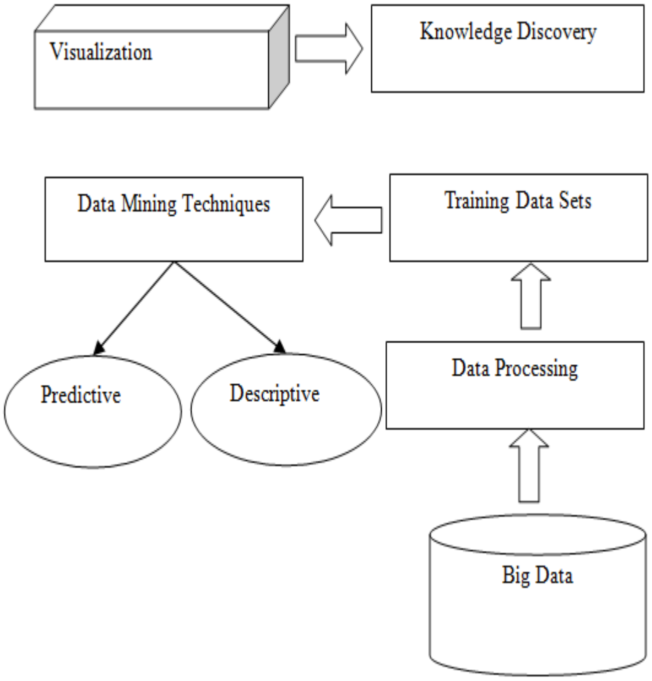

Figure-1: The Graphical Representation of Big Data Analytics Framework[1].

For the data processing many organizations has been adopted the MapReduce based systems for long-running batch job and one of the most powerful real-time processed based on big data tools that is a distributed computing platform, that is also known as S4 (Simple Scalable Streaming System). It is allow to developing applications for processing continuous unbounded stream of data[2]. The training datasets have processed by the appropriate data mining techniques and visualized the output in user understandable form or in proper visualization format.

Big Data Various Parameters

Volume is the first parameter comes to mind considering the term big data. Big data has initial three parameters which are known as 3Vs such as volume, variety and velocity, in other words big data has high volume, high velocity and high variety. According to one report that the Facebook process up to one million photographs per second and stored 260 billion photos using storage space of over 20 petabytes[3], and according Cisco Internet Business Solution Group, there will be 50 billion connected devices to the Internet reached by 2020[4]. After the three Vs some authors and research organization has extended other Vs which is also considered big data parameters. The following Figure-2 is present the different parameters of big data.

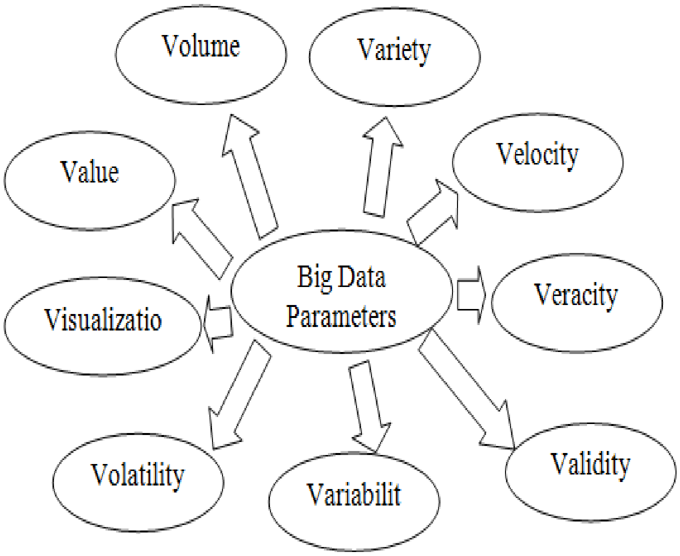

Figure-2: The Different Parameters in Big Data.

In the Figure-2 there are nine parameters of big data which also known as 9Vs of big data[5]. Volume: the volume is referred about the scale of data and it will reach to 40 zettabytes by 2020. Velocity: velocity is concerned about the speed at which large datasets are growing thus demanding analysis of streaming data over the Internet. Variety: the social media and other organization are generating the data in various format like text, videos, images, log files, XML etc., those are unstructured data. Some other data sets can be in structured format also. Veracity: it is refer to with inaccuracy of the data. When we analyzing the data sets, it should be in actual needed are really a cumbersome activity. There are several statistical and analytical process have to go for data cleansing for choosing intrinsic data for decision making[6]. Value: data value concern to derive useful information in decision making because it has direct impact on business profits and also play very important role in big data. There are huge amount of data available around us in various sectors, it is complex task to extract valuable data. Validity: validity refers to the result in proper understandable format, the given data set may not have any veracity problems and if it is not proper understandable so datasets is not valid. The estimated result is going to be used for the processed of decision making thus big data-producing sources and consequent analysis must be correct. Variability: the term variability to inconsistency data flow. This parameter has produced difficulty by increasingly access of digital media, which is the main way of peak in data loads. Volatility: volatility depends on volume, velocity and variety of big data because basically it refers to the life duration and liability. Due to the rapid changing in dynamic data which is generated from different areas, so volatility going to be decide how long time data is valid and should be stored. We need to understand and determine at the time of estimation about requirements, availability and lifetime of data[7]. Visualization: in big data era, the visualization is a actual representation of data in proper format, it can be in pictorial and graphical. When the data sets in vast volume so visualization and exploration become challenging task that is how to represent the data correct and understandable format. The data mining techniques is very useful for data visualization. Several approaches has been developed in the context of data visualization[8].

Parameter Estimation Based on Sampling Techniques

The sample selection procedure can be selected by a judgment or random procedure. The judgment based methodology leads to a sample of biased units. Due to biased the conclusion and prediction both get effected about the parameters. The random sampling are biased free but bear a level of uncertainty regarding better representation of population. When the population is vast then it is very difficult to process all data at a time therefore it necessary to develop different methods of sample selection for estimation that are provide precise enough for a specific purpose. The sampling technique is very useful and helpful when lack of time, prestige bias, lack of understanding, self-interest, etc.[9]. In most of the estimated population parameters, random sampling is used, there are following some parameter estimation techniques-

Simple Random Sampling (SRS)

The simple random sampling technique concerned with the drawing sample in such a way that each and every unit of population has an equal and independent chance of being included in sample. The probability of simple random sampling without replacement is $P=\frac{1}{N_{C_n}}$ being selected where N_{C_n} total possible sample from the population size N. in the SRS the sample drawn unit by unit and the unit in the population are numeric 1 to N and within the SRS the sample selection process must given an equal chance to every number in the population. Suppose we have a population N units then the probability of first unit sample selection is n/N and the second draw the probability that someone the remaining $(n-1)$ specified unit selected is $n-1/N-1$, and so on. The probability of n all specified unit are selected in n drawn is –

$$\frac{n}{N}\cdot\frac{n-1}{N-1}\cdot\frac{n-2}{N-2}\cdots\frac{1}{N-n+1}=\frac{n!(N-n)!}{N!}=\frac{1}{N_{C_n}} \quad (1)$$

In a sampling procedure we consider on certain properties that we use to measure and record for every unit which comes into the sample. The population are denoted by - $X_1, X_2, \ldots X_N$, and the corresponding values for the units in the sample are denoted by $x_1, x_2, \ldots x_n$, or we have the following definitions.

The total potation is-

$$X=\sum_{i=1}^{N} X_i = X_1+X_2+\cdots+X_N \quad (2)$$

The total sample is-

$$x=\sum_{i=1}^{n} x_i = x_1+x_2+\cdots+x_n \quad (3)$$

The population mean is-

$$\bar{X}=\frac{X_1+X_2+\cdots+X_N}{N}=\frac{1}{N}\sum_{i=1}^{N} X_i \quad (4)$$

The sample mean is

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n} x_i \quad (5)$$

The variance of the sample is-

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2 \quad (6)$$

The standard deviation of the sample is-

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}} \quad (7)$$

Where x the sample observation $\bar{x}$ is the mean of the sample and n is the total sample.

Confidence limit: The entire population is very difficult to calculate population mean $(\bar{X})$ or standard deviation (σ) but the confidence interval process can be reasonable straight forward for a sample. The confidence interval can be manipulated and then used for very useful of inference parameters of population. In the confidence interval used Z value for desire confidence probability (%) and find out lower bound and upper bound, where most of the sample values would be lie. By the definition the confidence interval is-

$$CI=\bar{y}_{st}\pm Z\frac{S}{\sqrt{n}} \quad (8)$$

Where Z is the desired confidence probability, $\bar{y}_{st}$ is an estimated mean of the sample, S is a variance and n is the sample size. The most common values are following[10,11].

Confidence Probability (%)	50	80	90	95	99
Value of Z	0.67	1.28	1.64	1.96	2.58

Stratified Random Sampling: The stratified sampling is widely used sampling for approximate query processing on the heterogeneous dataset after dividing in to homogeneous group. When the data is so large with different types of data is available then it is very difficult to apply simple random sampling on those data. The alternative sampling technique is stratified sampling, where the population is divided into groups called strata.

The stratified sampling provides flexibility to categorize the heterogeneous population into homogeneous strata and then process it. The concept of stratified sampling is that the population of N units is divided into K strata, each of size $N_1+N_2+N_3+\cdots+N_k$, $\sum_{i=1}^{k} N_i = N$ and sample size $n_1+n_2+n_3+\cdots+n_k$, $\sum_{i=1}^{k} n_i = n$. The sample allocation may be in three ways- Proportional, Neyman, Optimal allocation and arbitrary[12]. For example if there are N=10000 observation including image, videos, text, and log file and we want to calculate the mean of video so in this situation the simple random sampling is not sufficient first we should know the particular video files but here we have an opportunity that apply stratified sampling to divide population into stratum like strata 1 image, strata 2 video, strata 3 text and strata 4 log file then we can easily apply sampling techniques. The stratified sampling is precise techniques in comparison of simple random sampling.

Predictive Analytics Model: In big data era the textual content like structure, semi-structured and unstructured with multimedia content such as videos, images, audios, are producing bulk of data via different devices. The devices can be machine-to-machine communication, social media, sensor devices, Internet of Things (IoT), etc. So it is not possible to process all types of data at a particular time for present as well as future use also. In this scenario the forecasting method will help to predict useful information for present and future used by using sampling techniques.

The sampling-based parameter estimation techniques are very capable and useful to estimate different big data parameters and discover decision-based information. The predictive analytics is a forecasting method to determine the future possibilities through the sampling techniques. The following Figure-3 represents the different types of big data analytics which is used to predict decision-based values. The Figure-3 has explained steps by steps procedure for prediction analysis[13].

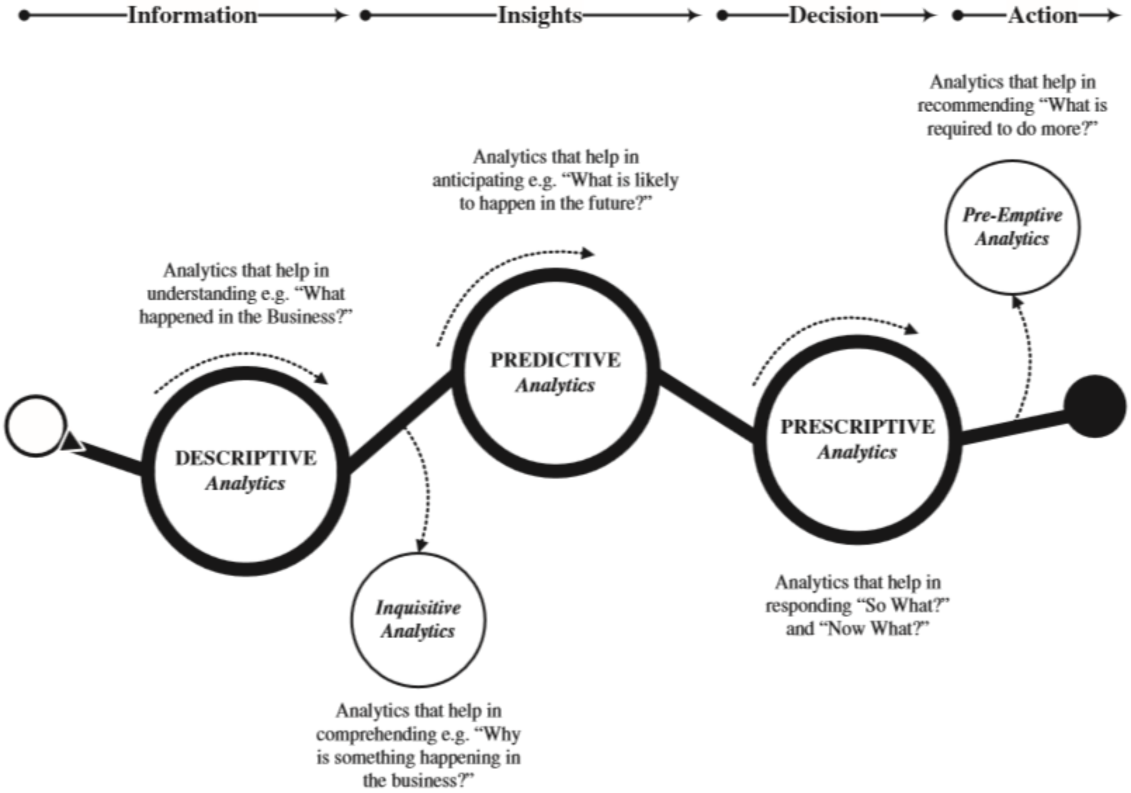
Information
Insights
Decision
Action
Analytics that help in recommending "What is required to do more?"
Analytics that help in anticipating e.g. "What is likely to happen in the future?"
Pre-Emptive Analytics
Analytics that help in understanding e.g. "What happened in the Business?"
PREDICTIVE Analytics
PRESCRIPTIVE Analytics
DESCRIPTIVE Analytics
Inquisitive Analytics
Analytics that help in responding "So What?" and "Now What?"
Analytics that help in comprehending e.g. "Why is something happening in the business?"

Figure-3: Classification of Big Data analytics[14].

In big data scenario the data sets are coming in dynamic nature from different devices or sources. The big data needs a dynamic prediction model for prediction and forecasting. The predictive analytics model not capable in single domain but it is also useful for different domain. For example the dynamic prediction of real-time road travel time on the traffic information platform. Stock price prediction model based satisfactory on short-term basis[15], help in the detection of illegal activities like money laundering and financial frauds[16]. The area of sampling is not limited, it can be used as per need on different purposed of different sampling methods such as simple random sampling, cluster sampling etc. the sampling techniques is applicable to find out accuracy measures, value of resulting data, economical decision, data collection, how long data available, average size of data generation in a particular time[17].

Parameter Estimation Problems under Big Data Model: Data visualization aims to make data more meaningful for user interpretation in heterogeneous data environment. The data represent one of the problem under visualization parameter estimation. Big data parameter (volume) has variety of data which is known as heterogeneous data sets. The heterogeneous data may create high-level redundancy of data, the most of the data generated by the sensor devices so the data should be not have repetition in other hands datasets must be free from unbiased[18]. Furthermore challenges facing by the humanity-population and resources, rich-poor gap, health issues, education, peace and conflict, energy management, status of women, educational decision going to be implemented[19]. Big data classification problems, crime prediction based on multi-level classification task with higher accuracy, crime trend forecasting[20]. Health parameter estimation problems[21]. The sampling-based estimation has variation of solution in different areas of big data mining when data is so large.

Proposed Prediction Model Based on Sampling Technique: In big data era, the data size and transmission speed of data is a challenging task for big data processing and discover valuable information. The 95% of big data is an unstructured format[22]. The social media are producing numerous in every data and WhatsApp is very famous communication media over social media. In a proposed system we have assumed that huge database of N numbers of user with concerned to WhatsApp users, even WhatsApp has more than one servers in different places, and one server are storing huge amount of data which are generating by the users such as text, images, videos. Here we have consider that large datasets which have heterogeneous data. In this paper we have developed a mathematical estimation model which able to calculate average size of volume which has been generated by the WhatsApp user in different time (t). The proposed estimation method used stratified random sampling technique. The following Figure-4 represents the overview of proposed framework.

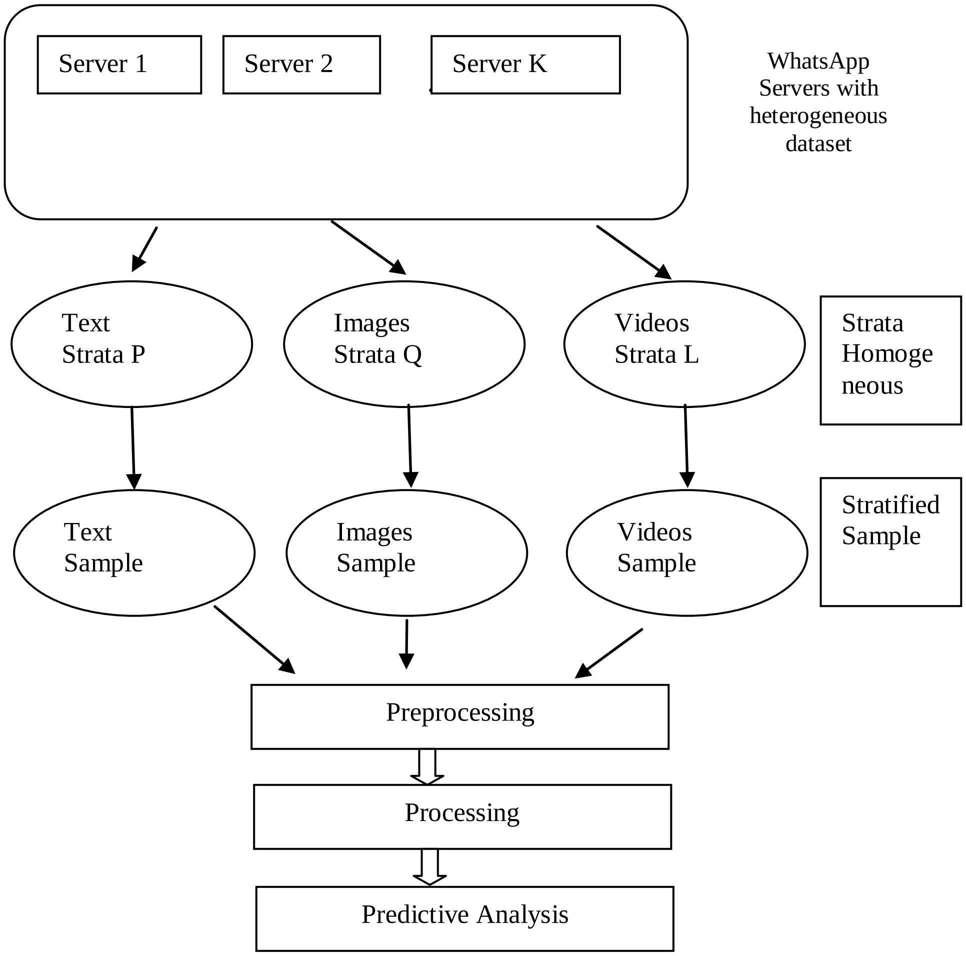

Figure-4: Graphical Representation of Proposed Model.

In the above-proposed model we have used stratified random sampling (SRS) because SRS allows dividing entire population into different strata (group) then we create a sample from strata and after that will perform predictive analysis by using SRS technique. In the proposed methodology we have assumed there are K different WhatsApp servers available in various places and store heterogeneous data like text, images, and videos. Furthermore we have divided in strata to all population (volume of data) as a homogeneous form and we have taken sample of data from the each strata.

Let's assume that there are K-1 server available on different places. The population (data) of N units is divided into three strata such as P, Q and L which is text, images, and videos respectively. Each of server size $N_1 + N_2 + N_3 + \cdots + N_k$, and sample size $n_1 + n_2 + n_3 + \cdots + n_k$ which are randomly selected from population N_i . Here the our assumption is to calculate the average size of WhatsApp data which has stored in different servers in a particular time. We have followed the following procedure to estimate the unknown parameter which is average size of volume at the given time (t).

Results and Discussion

The proposed based on the stratified sampling and in this we have estimated the one big data parameter as a volume (size). The given data set or population has stored three types of data image, video and text at different time interval like t_1, t_2 and t_3, the following table are showing the nature of dataset. In the given dataset we considered the give file size in Megabyte.

Here the assumption is that what is the average size of volume in a day which is generated by the user in different time and the user are increasing like 71, 80 and 88. We have estimate the volume separately like t_1, t_2 and t_3, after that pooled all result and finally have estimated (T). In this result confidence interval has calculate at 95% and 99% both. Furthermore we have compared our result with another sampling technique as a simple random sampling. We have followed the arbitrary sample selection method and taken 70% sample randomly from the dataset. The following Table-3 are showing the compared result.

Table-1: Proposed Algorithm.

Step 1	Collect the data from different servers, there are K-1 server in the population (volume). $N_1 + N_2 + N_3 + \cdots + N_k$
Step 2	Applying stratified sampling techniques and divide the entire population (N) into different strata (P, Q, L). P(text), Q(images), and L(Video)
Step 3	Draw sample (n) from the strata, the sample is $n_1 + n_2 + n_3 + \cdots + n_k$
Step 4	When the data is dynamic and time-dependent then we need to decide some general weight according to data generation time $W_i = \frac{N_i}{N}$
Step 5	Estimation of mean for the proposed method $\bar{y}_{st} = \sum_{i=1}^{m} W_i \bar{y}_i$ to estimate population mean $Y = \sum_{j=0}^{n} W_j \bar{Y}_j$ of the study parameter (Y)
Step 6	Estimation of variance of $\bar{y}_{st}$ which is unbiased estimator of the population $V(\bar{y}_{st}) = \sum_{i=1}^{m} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i{}^2 S_i{}^2$ where $S_i{}^2(t) = \frac{1}{N_i - 1} \sum_{i=1}^{N} \left(Y_{ij} - \bar{y}_i\right)^2$
Step 7	The pooling value of $\bar{y}_{st}$ on time T is $(\bar{y}_{st})T = \left[\theta_1(\bar{y}_{st})t_1 + \theta_2(\bar{y}_{st})t_2 + \cdots + \theta_q(\bar{y}_{st})t_q\right]$ where θ is weight and the confidence interval from the population strata on a different time is- $V\,(\bar{y}_{st})T \pm Z\sqrt{\mathrm{V}\,(\bar{y}_{st})T}$ where Z desire confidence probability (%).

There are seven steps in our proposed estimation method to find out the average size of servers data on different time t.

Table-2: Nature of the Dataset in Different Time Interval.

t_1				t_2				t_3			
User ID	Images	Videos	Text	User ID	Images	Videos	Text	User ID	Images	Videos	Text
1	300	800	100	1	190	550	113	1	450	1270	218
2	400	750	120	2	290	500	133	2	500	1220	238
3	200	300	10	3	140	50	23	3	400	770	128
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
71	523	856	98	80	600	500	23	88	150	720	105

Table-3: The Proposed Algorithm Result.

Time	n_i	w_i	$\bar{y}$	σ	$S_i{}^2$	$(\bar{y}_{st})_{t_i}$	$V(\bar{y}_{st})_{t_i}$	$(\bar{y}_{st})_T$	$V(\bar{y}_{st})_T$	σ (Avg)
t_1	n1=50	w1=0.33	524.36	288.13	83016.28	623.58	4755.45	634.18	1458.19	816.38
	n2=50	w2=0.33	1253.44	2051.62	4209135.52					
	N3=50	w3=0.33	106.54	178.37	31814.78					
t_2	n1=56	w1=0.33	397.55	284.95	81196.54	486.63	4249.91			
	n2=56	w2=0.33	959.59	1937.43	3753631.48					
	N3=56	w3=0.33	113.46	169.48	28724.54					
t_3	n1=62	w1=0.33	560.02	351.66	123665.92	754.34	4118.38			
	n2=62	w2=0.33	1517.47	1892.90	3583087.43					
	N3=62	w3=0.33	208.39	192.55	37229.42					
Confidence Interval (95% and 99%)						[559-709], [520-749]				

Table-4: The Simple Random Sampling Result

Time	n_i	$\bar{y}$	σ	${S_i}^2$	$(\bar{y}_{st})_T$	$V(\bar{y}_{st})_T$	σ (Avg)
t_1	n1=50	628.11	1283.78	1648095.36	626.76	1529012.71	1235.82
	n2=50						
	N3=50						
t_2	n1=56	490.20	1181.85	1396780.85			
	n2=56						
	N3=56						
t_3	n1=62	761.96	1241.84	1542161.92			
	n2=62						
	N3=62						
Confidence Interval (95% and 99%)			[519-735], [486-768]				

In our result illustration the proposed solution is given in Table-3 and the simple random sampling result is given in Table-4 so here we see the difference in variability and the standard error. The proposed algorithm's standard error is 816.38 and the SRS standard error is 1235.82 that is more than to proposed algorithm. The variance is also high in simple random sampling. The population estimated means is $\bar{Y}_{st} = 561.63$ in stratified sampling and $\bar{Y}_{st} = 555.77$, both are lies in confidence interval. So we can say that the proposed solution is better than the simple random sampling.

Conclusion

In this research paper we have described the big data and its various techniques which are very useful for processing of large amounts of data such as big data mining techniques. This paper has focused on parametric estimation of big data and we have discussed different parameters like 9Vs of big data. Furthermore we have also discussed about predictive analysis and proposed sampling-based estimation method for calculating average size of unstructured data at different time with dynamic nature. This research paper has presented two solution based on sampling techniques. This paper focused on stratified sampling and we have proposed algorithm based on this sampling method. Also we have compared our solution with the simple random sampling and we have found that the proposed solution is precise in comparison of simple random sampling.

References

1. Chauhan R. and Kaur H. (2015). A Spectrum of Big Data Applications for Data Analytics. *Computational Intelligence for Big Data Analysis*, Vol 19, Springer, Cham, pp 165-179. ISBN: 978-3-319-16598-1

2. Targio Hashem, Ibrahim Abaker, Yaqoob, Ibrar, Anuar, Nor Badrul, Mokhtar, Salimah, Gani, Abdullah, Khan, Samee Ullah (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47, 98-115.

3. Gandomi, Amir and Haider, Murtaza (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144.

4. Barrachina, Duque Arantxa and O'Driscoll, Aisling (2014). A big data methodology for categorising technical support requests using Hadoop and Mahout. *Journal of Big Data*, 1(1), 3-11.

5. Alim, Abdul and Shukla, Diwakar (2018). Big Data: Myth, Reality and Parametric Relationship. *International Journal of Advanced in Management, Technology and Engineering Sciences*, 8(3), 1235-1244.

6. Kune, Raghavendra, Konugurthi, Pramod, Agarwal, Arun, Rao, C., and Buyya, Rajkumar (2015). The Anatomy of Big Data Computing. *Software: Practice and Experience*, 46(1), 79-105.

7. Khan, Nawsher, Alsaqer, Mohammed, Shah, Habib, Badsha, Gran, Ahmad Abbasi, Aftab & Salehian, Solmaz (2018). The 10 Vs, Issues and Challenges of Big Data. Proceeding in International Conference on Big Data and Education. Honolulu, HI, USA, 09th-11th March, pp 52-56. ISBN: 978-1-4503-6358-7

8. Bikakis, N. (2018). Big Data Visualization Tools. Encyclopedia of Big Data Technologies, Springer, Cham. ISBN: 978-3-319-77525-8

9. Shukla, Diwakar and Singh Thakur, Narendra (2014). Imputation Methods in Sampling. Aman Prakashan Sagar, India. pp 129-148, ISBN:978-93-80296-31-9.

10. Shukla, Diwakar and Rajput, Y.S. (2010). Graph Sampling. Aman Prakashan Sagar, India. pp 1-176, ISBN: 978-93-80296-03-6.

11. Cochran, William G. (1977). Sampling Techniques. John & Sons, USA, pp 1-442. ISBN: 0-471-16240-X

12. Nguyen, Trong Duc, Shih, Ming-Hung and Srivatava, Divesh (2019). Stratified Random Sampling from Streaming and Stored Data. Proceedings of the 22nd International Conference on Extending Database Technology (EDBT), 26th-29th March, pp 25-36. ISBN: 978-3-89318-081-3

13. Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263-286.

14. Yang, Zhao-Xia and Zhu, Ming-Hua (2019). A Dynamic Prediction Model of Real-Time Link Travel Time Based on Traffic Big Data. Proceeding in International Conferece on Intelligent Transporation, Big Data & Smart City (ICITBS). 330-333, DOI: 10.1109/ICITBS.2019.00087

15. Adebiyi, Ayodele, Adewumi, Aderemi and Ayo, Charles (2014). Stock price prediction using the ARIMA model. Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim, Changsha, China, China, 12th – 13th Jan, pp 105-111.

16. Peng, Zhihao (2019). Stocks Analysis and Prediction Using Big Data Analytics. Proceeding in International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Changsha, China, China, 12th – 13th Jan, pp 309-312.

17. Hashemian, M. S., Abkar, A. A., & Fatemi, S. B. (2004). Study of sampling methods for accuracy assessment of classified remotely sensed data. In International congress for photogrammetry and remote sensing, pp. 1682-1750. Available at: https://www.researchgate.net/publication/252668114_STUDY_OF_SAMPLING_METHODS_FOR_ACCURACY_ASSESSMENT_OF_CLASSIFIED_REMOTELY_SENSED_DATA. Accessed on 10.11.19.

18. Chen, Min, Mao, Shiwen and Liu, Yunhao (2014). Big Data: A Survey. *Mobile New App*, 19(2), 171-209.

19. Lee, Jae-Gil and Minseo, Kang (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, 2(2), 74-81.

20. Feng, Mingchen, Zheng, Jiangbin, Han, Yukang, Ren, Jinchang and Liu, Qiaoyuan (2019). Big Data Analytics and Mining for Crime Data Analysis Visualization and Prediction. In *IEEE Access*, vol. 7, pp 106111-106123. DOI: 10.1109/ACCESS.2019.2930410

21. Venkatesh, R., Balasubramanian, C. and Kaliappan, M. (2019). Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique. *Journal of Medical Systems*, 43(8), 1-8.

22. Amir Gandomi and Murtaza Haider (2015). Beyond the hype: Big data concepts methods and analytics. *International Journal of Information Management*, 35(2), 137-144.