

Consolidation Approach for Critical Data Identification in Optimal Clusters towards Big Data Security

Angeline Benitta¹ and S.P.Victor²

²Associate Professor, Department of Computer Science, St.Xaviers College, Tirunelveli,

¹Research Scholar, Manonmaniam Sundaranar University, Tirunelveli.

Abstract

Big data generally focuses on analysis based consolidated approach with the expectation of desired results in a standard way of responses. The role of critical data is very effective when used in relation to a big data dealing with specific clusters. Big data in business organizations are entitled with specific field surroundings which are able to make twist and turns based on the value it occurs or on the frequency it changes often. The process of identifying the critical data is the lead to identify the characteristics that affect the business organizations, in order to improve the process the critical data identification and reasons for the change cause along with the current impact predictions are essential to improve the monitoring process in an organization. This proposed method of consolidation approach allowing only the most relevant information being provided to yield a higher quality of critical data results. This paper deals with the consolidation approach for providing a critical data identification with its effects in terms of its roles towards Big data indirectly for its security. In near future the focus will be to implement the neuro fuzzy approaches in real time using neural network conceptual schema for Big data security.

Keywords: Big data, critical data, cluster, consolidation, optimality

I.INTRODUCTION

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software [2]. Big data is a term that describes the large volume of data both structured and unstructured, that affects a business on a day-to-day basis [1].



Fig-1: Big data applications

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups [3]. In other words Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups [4]. The quality of cluster depends on the method used. Clustering is also called as data segmentation, because it partitions large data sets into groups according to their similarity [5].

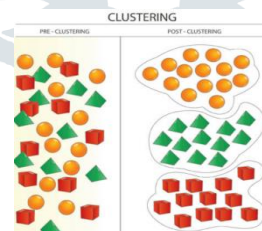


Fig-2: Clustering in Big data

Critical data plays a vital role in information management system that are used as criteria for processing, searching and matching the business or organization critical events, data categorization adjustment, actions to be taken on critical events found, and decisions around data survivorship. [7]. Data critical to one business area may not be critical for another [8].

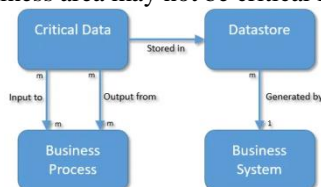


Fig-3: Critical data in big data

II. PROPOSED METHODOLOGY

The proposed methodology describes the process of identification and analysis of critical data from big data in a sequential process. Initially the big data in CSV/Excel/SQL tables are preprocessed by cleaning noisy data, removing blanks, and proper conversions and given as an input to the big data handling tool for the formation of clusters. The resultant output file with csv/xlsx/tables is taken into consideration for further processing with respect to the clusters formed by the tools. The output clusters contain the data value with proper possible primary key and point of sale to organization based on the business criteria such as sales/income/orders/process outputs as an individual impact towards the effective governance. The interpretation of the particular critical data field value are analyzed due to its value makes the twists and turns in business logic. Finally the feedback analysis system is given for the improvement of effective business or organization governance Big data analysis of critical data. Hence the security is essential for this critical data value in Big data which will be focused in near future.

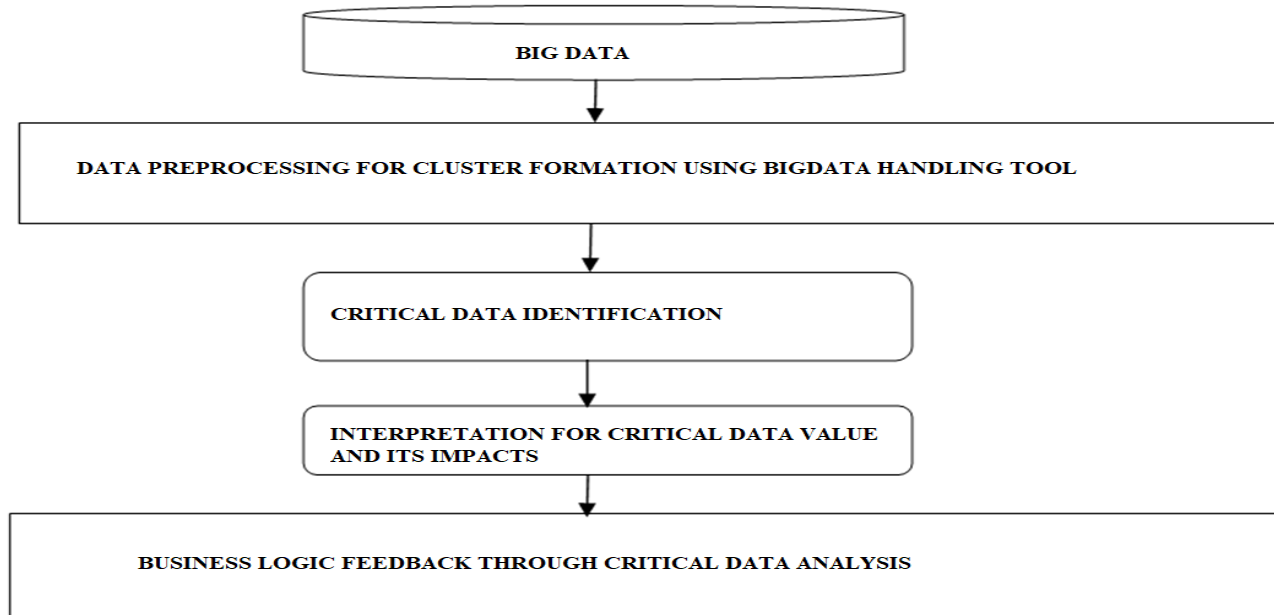


Fig 4: Proposed Critical data Identification and Analysis in Big data

The following algorithm steps explores the schema for the proposed methodology for any Business Big data governance,

Algorithmic Approach:

Step-1:

Input-Big data File; Tool-RStudio; Process-Cluster Formation; Output-Cluster data file

Step-2:

Identify the critical data field with its relation to Primary key

Step-3:

Critical data Value Analysis :{ Min, Average, Max}

Step-4:

Compare Existing/Actual output vs. Expected Normal curve based output

Step-5:

Output-Feedback for Business Governance.

III. IMPLEMENTATION

The implementation of Big data in real-time data collection is a sequential process to handle it with proper care, this proposed methodology is implemented through the software R-Studio [10]. A real-time product development and cargo company big data with conventional and conversional datasets are taken into consideration for our proposed methodology implementation, this product development and cargo company may be named as XYZ company in the world. Data storage: initially the Big data contains nearly 65000 rows of Excel-CSV file storage. The data mining tools for vast data is not feasible in excel. Data formatting and alignment are properly done through the proper removal of unwanted rows and columns along with the primary fields, numeric fields and descriptive fields and additional fields are ordered and added properly. The empty fields are removed by GOTO option in Excel for special field selection as Blank cells and identify the corresponding row selection for removal or deletion. The NaN are removed by setting NaN<-NULL in RStudio.

Fig 5: Big Data CSV file

Loading the file in RStudio

Fig 6: RStudio file loading implementation methodology

The output file generated by RSTUDIO which produces 1441 clusters based on the vendor name is as follows

Table with 48 columns and 48 rows of company names and their respective categories or locations.

Table with 48 columns and 48 rows of company names and their respective categories or locations.

Table with 48 columns and 48 rows of company names and their respective categories or locations.

Table with 4 columns: Rank, Company Name, Rank, Company Name, Rank, Company Name, Rank, Company Name. Lists various industrial and engineering companies such as Marcel and Marcel Nig Ltd, Mexen Integrated Services, Nadudech Nigeria, etc.

Table with 4 columns: Rank, Company Name, Rank, Company Name, Rank, Company Name, Rank, Company Name. Lists companies like Webster Inspection Services, Wegg Equipamentos Elctricos S/A, Weir Bnk Valves, etc.

Fig 7: Cluster Field Values for Critical data Identification in Big data

Procedure to Identify Critical data:

The critical data identification based on the number of orders handled by the cargo company yields the analytical view of the organization performance towards feasible handling of package deliveries. The tool used for processing is XLMiner and Excel -Addins [11] for data analytics package.

Table 1: Critical data Analysis

Order range	Type Value	Companies Count
Minimum orders	1	228
Maximum orders	3855	1
More Than 100	>100	130
More Than 200	>200	61
More Than 300	>300	34
More Than 400	>400	21
More Than 500	>500	17
More Than 600	>600	12
More Than 700	>700	9
More Than 800	>800	8
More Than 900	>900	6
More Than 1000	>1000	6
More Than 1100	>1100	5
More Than 1200	>1200	2
More Than 1300	>1300	2
More Than 1400	>1400	2
More Than 1500	>1500	2
More Than 1600	>1600	2
More Than 1700	>1700	2
More Than 1800	>1800	2
More Than 1900	>1900	2
More Than 2000	>2000	1
More Than 2100	>2100	1
More Than 2200	>2200	1
More Than 2300	>2300	1
More Than 2400	>2400	1
More Than 2500	>2500	1
More Than 2600	>2600	1
More Than 2700	>2700	1
More Than 2800	>2800	1
More Than 2900	>2900	1
More Than 3000	>3000	1
More Than 3100	>3100	1
More Than 3200	>3200	1
More Than 3300	>3300	1
More Than 3400	>3400	1
More Than 3500	>3500	1
More Than 3600	>3600	1
More Than 3700	>3700	1
More Than 3800	>3800	1
More Than 3900	>3900	0
But 2 to 99	>1 & <99	439

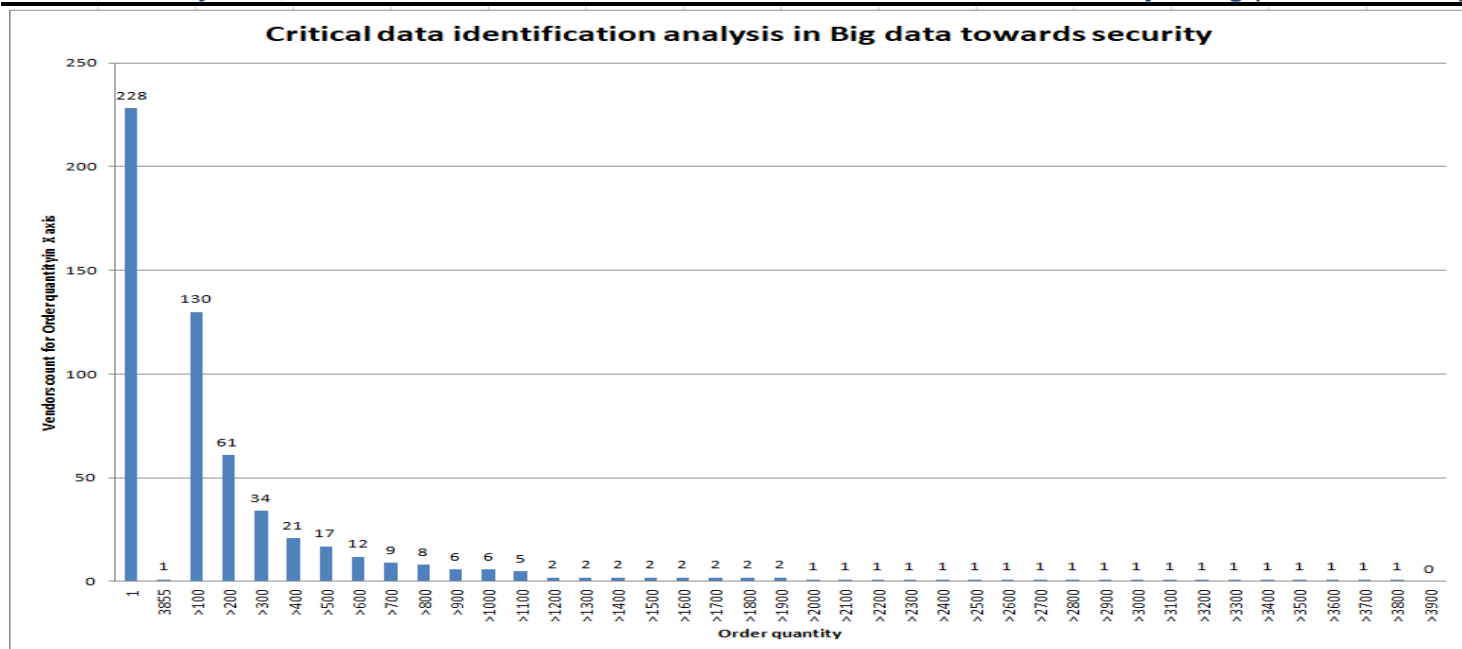


Fig 8: Critical data value Identification

Critical Data Analysis:

The following data analysis by add-ins in excels produces the various categoral lead values in critical data field for further processing towards organization governance.

3a Technologies 5	5 Minimum orders	1	228
A-Plus Enterprises Pvt. Ltd. 88	88 Maximum orders	3855	1
A. Dee Global Resources Limite 5	5 Between 2 and 10	2 to 10	644
A.B.L.K. El-Lo & Sons Resource 11	11 Between 11 and 20	11 to 20	161
A.N Nwakile Gen. Ent. Ltd 92	92 Between 21 and 30	21 to 30	82
A.S.Kura Brothers Ltd 1	1 Between 31 and 40	31 to 40	52
A.U & Sons Facilities Enterpri 125	125 Between 41 and 50	41 to 50	43
A.Z. Hollink South Africa (Pty 8	8 Between 51 and 60	51 to 60	31
Aabchuss Ventures Ltd 18	18 Between 61 and 70	61 to 70	22
Abbnig Limited 4	4 Between 71 and 80	71 to 80	17
Abgelisco Resources Nig 34	34 Between 81 and 90	81 to 90	14
Abiodun Osoba 17	17 Between 91 and 100	91 to 100	17
Able Instruments & Controls 3	3		
Abuchi Global Services 47	47		
Acm Nig Ltd 3	3 Between 101 and 200	101 to 200	69
Active It Distribution Fzco 5	5 Between 201 and 300	201 to 300	27
Adams Armaturen GmbH 17	17 Between 301 and 400	301 to 400	13
Addak Power Engineering Ltd 2	2 Between 401 and 500	401 to 500	4
Adef Ventures 2	2 Between 501 and 600	501 to 600	5
Adeka Palmarole Sas 1	1 Between 601 and 700	601 to 700	3
Adept & John Nig Ltd 3	3 Between 701 and 800	701 to 800	1
Adetunji Ilori Stores 8	8 Between 801 and 900	801 to 900	2
Advance Compseals Pvt Ltd 20	20 Between 901 and 1000	901 to 1000	0
Advanced Filtration Solutions 2	2 Between 1001 and 2000	1001 to 2000	5
Afesemi Investment Limited 12	12 Between 2001 and 3000	2001 to 3000	0
Afisemi Nig. Ltd. 6	6 Between 3001 and 4000	3001 to 4000	1
Aforma (Nig) Coy 35	35 More than 4000 orders	>4000	0
African Fertilizer & Chemicals 119	119		1442
African Petroleum Plc 51	51 Subtract the repeated Max order		1
Afza Material Handling And Sto 2	2 Total Clusters		1441
Ageco International Fze 2	2		
Agogos Resources Nigeria 32	32		
Ahmid Idris Limited 6	6		
Aibes Communications 2	2		
Air Liquid Nigeria Plc 4	4		
Air Liquide Uk Limited 84	84		
Air Wave Limited 5	5		
Ajomesco Concept Enterprises 45	45		
Aju-Doh And Sons Ltd. 3	3		
Akadan Global Services 4	4		

Fig 9: Critical data Analysis

Step-1: Focus on the Max ordered values

The critical data of max 3855 orders issued and processed by the single company is "yautec international limited". The following table illustrates the high range of orders processed during the five years.

Table 2: Trusted customer companies

Highly Satisfied orders	Range	Number of Companies
Between 1001 and 2000	1001 to 2000	5
Between 2001 and 3000	2001 to 3000	0

Between 3001 and 4000	3001 to 4000	1
More than 4000 orders	>4000	0

The following figure shows the max ordering companies for the organization,



Fig 10: Maximum order raising companies

Step-2: Focus on the Min ordered values

The critical data of min only 1 order issued and processed by the companies=228.

The critical data of 2 to 10 order issued and processed by the companies=644.

In a span of 5 years (2011 to 2015) the number of companies accessed our products and cargo company for processing is 872 out of 1441. The following table illustrates the minimal orders processed by the organization during the period of 5 years.

Table 3: Minimum order supplied by customer companies

Processed orders during 5 years	Companies count(Clusters)
Only one order	228
2 to 10 orders	644
Total No of Min orders	872
Total Companies/Clusters	1441
Min order Companies percentage %	60.51

The following figure shows the actual scenario for the organization.

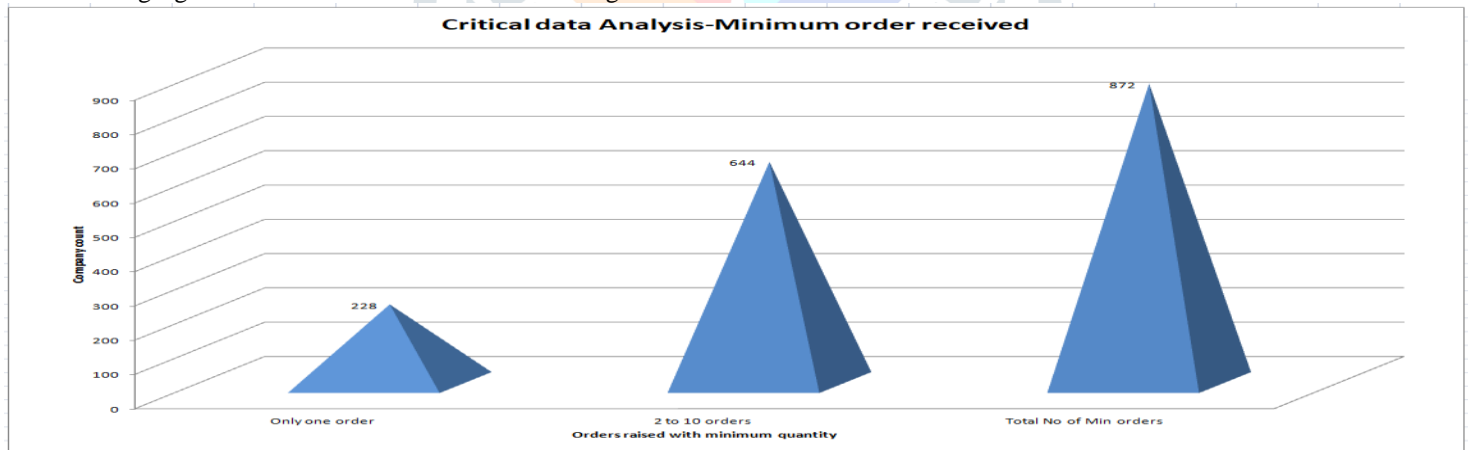


Fig 11: Minimum ordering companies

Step-3: Focus on the middle ordered values

The critical data of middle or tolerable mean orders issued and processed by the companies=1441-872-(Max) =1441-878=563

Table 4: Average order supplied customer companies

Processed orders during 5 years	Companies count(Clusters)
Max ordered Companies	6
Min ordered Companies	872
Total Companies/Clusters	1441
Average order companies	1441-878=563
Average order Companies percentage %	39.07

The following figure shows the actual scenario for the organization.

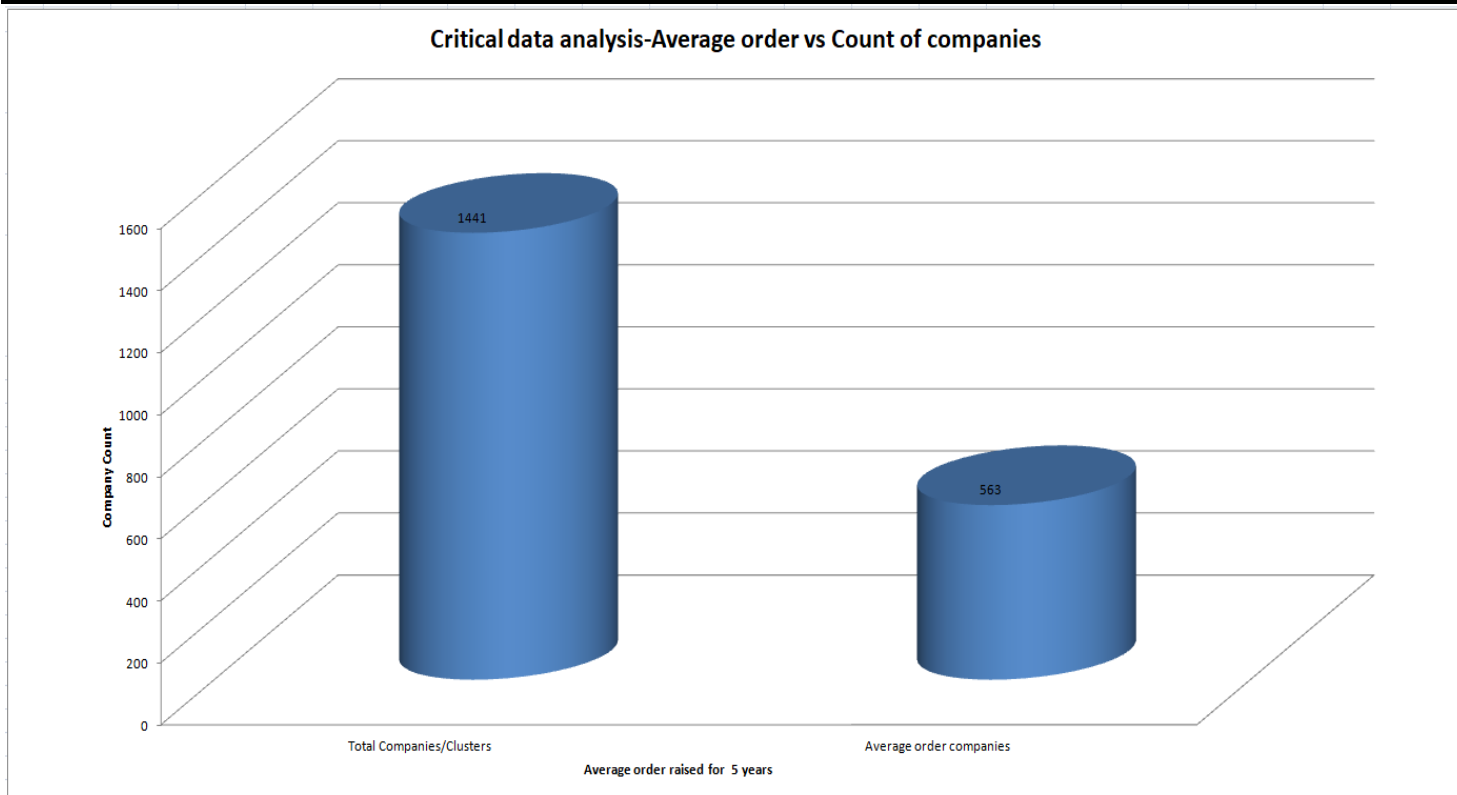


Fig 12: Analysis of Average ordering of companies

IV.RESULTS AND DISCUSSION

The critical data analysis for the Big data yields the sensitive information about the organization performance and moreover it acts as the datum to be confidentially maintained by any of the organization. The XLMiner tool works well in the Data analysis of 1441 clusters with critical data field validations and verifications. The different directional perspective view of critical data field value produces the results for the organization as in the following table. A well performing organization processing results must be in a Normal curve of 20% Min orders, 60% average orders and 20% Max orders.

Table 5: Organization performance based on critical data analysis

Orders for the past 5 years	No of orders	Performance % for the duration of 5 years	Expected Performance %	Deviation
Minimum orders	872	60.51	20	40.51
Average orders	563	39.07	60	-20.93
Maximum orders	6	00.42	20	-19.58
Total orders	1441			

The following figure shows the organization performance for dealing with min, max and average orders for the past 5 years.

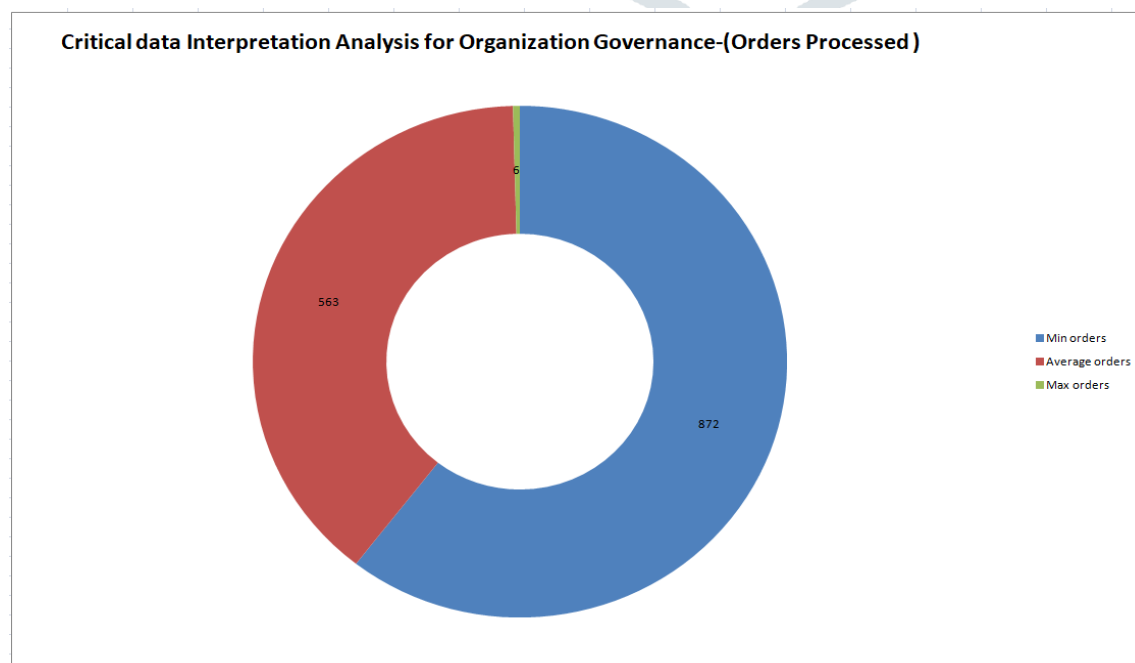


Fig 13: Critical data analysis results for the Organization order processed schema

The following figure shows the need for improvement in terms of comparison with the nominal expected output of a normal curve.

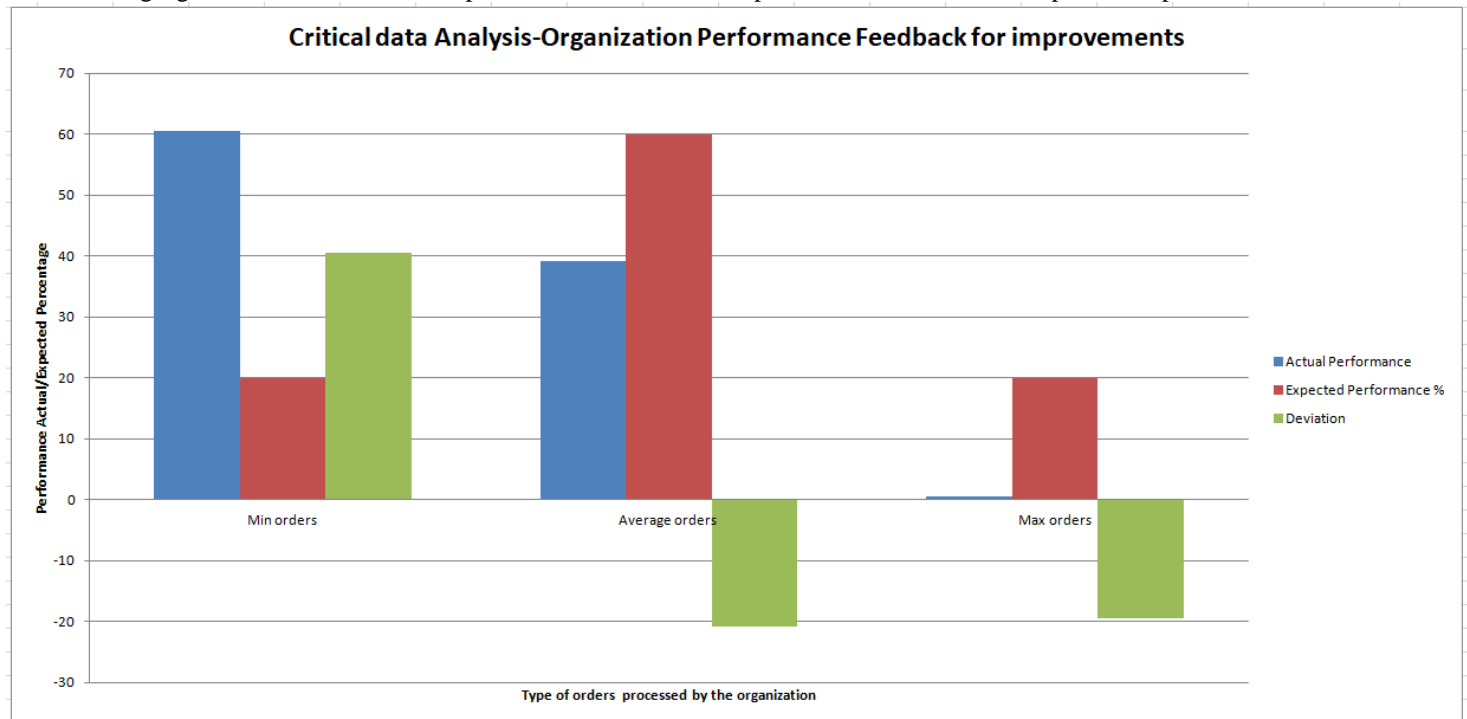


Fig 14: Comparison with the actual and nominal expected output performance based on critical data analysis

The proposed methodology shows that the deviation in performance of the organization resembles the need for the insight view of the business model for further improvements.

V.CONCLUSION

Big data is a data with vastness, variety, velocity of large magnitude of information's out of which analyzing the data using normal access or excel is not in a feasible condition so the inclusion of XLMiner or any add-ins for Excel or CSV is important to handle the analysis of data. The process of handling and extracting critical data from big data with individual component implementation of data cleansing for data preprocessing reduces the Big data analysis complexity, Critical Data field selection with its validity explores the business strategies with optimal magnitude in other words it will act as the glimpse or turning point for the entire organization business process. Our proposed critical data analysis schema reduces the time complexity and data complexity along with XLMiner tool or Excel Addins pack and Solver pack reduces the complexity in towards critical data interpretations, the final result yields 95% (Level of significance=5%) optimal output for the organization betterment than the normal implementation. The overall method proves to be highly efficient compared to normal Excel and access based approach because of its denial of service towards the data consistent issues also, dramatically reducing running time and number of features required for the efficient critical data issues. Moreover, the experimental results revealed that the critical data in a Big data for an organization business process is highly sensitive and it must be kept confidential with higher level of security algorithms applied to it. In near future the focus will be to propose a fuzzy based neural network oriented critical data security for Big data domain.

References

- [1] Shirudkar, Kalyani, and Dilip Motwani. "Big-Data Security." Department of Computer (2015).
- [2] Maturdi, Bardi, et al. "Big Data security and privacy: A review." China Communications 11.14 (2014): 135-145.
- [3] Jensen, Meiko. "Challenges of privacy protection in big data analytics." Big Data (Big Data Congress), 2013 IEEE International Congress on. IEEE, 2013.
- [4] Wang, Cong, et al. "Privacy-preserving public auditing for data storage security in cloud computing." Infocom, 2010 proceedings IEEE, 2010.
- [5] Clifton, Chris, et al. "Tools for privacy preserving distributed data mining." ACM Explorations Newsletter 4.2 (2002): 28-34.
- [6] Conn S.S., 2015, *OLTP and OLAP Data Integration: a Review of Feasible Implementation Methods and Architectures for Real Time Data Analysis*, <http://academic.regis.edu/cias/Library/pid54211.pdf>, Access on: 15.02. 2015.
- [7] Lillian Clark, I-Hsien Ting, Chris Kimble, Peter Wright, Daniel Kudenko (2006) "Combining ethnographic and clickstream data to identify user Web browsing strategies" Journal of Information Research, Vol. 11 No. 2, January 2006
- [8] Eirinaki, M., Vazirgiannis, M. (2003) "Data Mining for Web Personalization", ACM Transactions on Internet Technology, Vol.3, No.1, February 2003
- [9] Mobasher, B., Cooley, R. and Srivastava, J. (2000) "Automatic Personalization based on data Mining" Communications of the ACM, Vol. 43, No.8, pp. 142-151.
- [10] RStudio -Open Source software for Big data.
- [11] XLMiner and solver Addins.