Project Documentation:

1 INTRODUCTION

1.1 INTRODUCTION TO DATA VISUALIZATION:

Data visualization is the graphical portrayal of data and information.By utilizing visual components like outlines, charts, graphs, and maps, data visualization tools give an open method to see and get patterns, anomalies, and examples in information.Let me explain to you each of these terms one by one.

1.2 PURPOSE THE USE OF DATA VISUALIZATION:

Data visualization helps us a clear idea of what the information means by giving it visual context through maps or graphs. This makes the data more characteristic for the human mind to comprehend and therefore makes it easier to identify trends, patterns, and outliers within large data sets.In this project I will explain my storytelling dashboard using bar chart and tree maps. Here I analysis the Country wise India Exports by Principal Commodity in 2011-12 and 2012-13. Using some queries for exploratory data analysis i.e., To which country most of the primary goods are sent fromIndia,Which goods are highly exported from India?Highly paid primary goods export from India (2011-12)Highly paid primary goods export from India(2012-13).

2 LITERATURE SURVEY:

[1]Hadoop based Analysis and Visualization of Diabetes Data through Tableau[2019]

Due to rapid development of diverse healthcare practices, various procedures used in healthcare, produce data. This healthcare data has been scaled to a bigger size, thus, there is a dire need to analyse this scaled data in an efficient manner. Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of commodity computers. Analytical processing of big data with Hadoop is very helpful in performing significant actual-point in time analysis on massive amount of data and is capable to forecast an emergency situation. In this paper Hadoop driven analysis has been performed on a diabetes case study through comparison among Pig, Hive, and Tableau The superiority of Tableau is established as it allows users with minimal statistical background to visualize the results of analysis in an easy ‘onbutton-click’ manner.

[2]Leveraging Column Family to Improve Multidimensional Query Performance in HBase[2017]

Apache HBase is a widely used non-relational database in the Hadoop ecosystem. However, it will be inefficient if users perform multidimensional queries. Some of existing approaches incur extra costs in write performance or consistency maintenance, others are limited to specific applications. In this paper, we propose a novel data model called CFIDM,

short for Column Family Indexed Data Model. In CFIDM, we convert the queried column into multiple column families. Values in the specific column are partitioned. Each partition is manifested by a column family, turning column family into an index with no additional cost. Then we provide guides to build this data model. Finally, we evaluate the effectiveness and versatility of CFIDM on the Bixi data set and the TPC-DS benchmark. Results show that CFIDM can save 6.6% disk space for Bixi and 35% for TPC-DS, maximally speeding up the queries by 5X and 5.5X respectively.

[3]A performance evaluation of Hive for scientific data management[2013]

It is very important to evaluate the MapReduce-based frameworks for scientific data processing applications. Scientists need a low-cost, scalable, easy-to-use and fault-tolerance platform for large volume data processing eagerly. This paper presents an implementation of a scientific data management benchmark, SSDB, on Hive, a MapReduce-based data warehouse. A complete strategy of migrating SSDB to Hive is described in detail including query HQL implementation, data partition schema and adjustments of underlying storage facilities. We have tuned the performance using several system parameters provided by Hive, Hadoop and HDFS. This paper provides preliminary results and analysis. Evaluation results indicate that Hive achieves acceptable performance for some data analysis tasks even compared with some high efficient distributed parallel databases, but it needs subtle adjustments of underlying storage facilities and indexing mechanism.

[4]The Challenges of Big Data Governance in Healthcare[2018]

Big data starts to be employed in some industries but not yet widely or properly adopted in healthcare industry. This research paper aims at studying the usages and challenges of big data in healthcare sector. Governance of big data will include the domains of strategy, process, people, policy and technology and automation. Among the challenges identified in the healthcare sector, reliability and integrity are especially important because it is related to life and death. Big data governance for policy maker, authentication for data integrity, and future development of healthcare big data governance are discussed here. Moreover, some future development questions are raised in this paper for further study, which will improve the quality of life and lead to a better and healthier world under the proper and adequate big data governance environment.

[5]Different analytical techniques for big data analysis: A review[2017]

Big data refers to any collection of data so large and complex that it exceeds the processing capability of conventional data management system and techniques. Big data is the

hot topic of research now a day. The paper gives some ideas about the research area to look into. It has been done as a part of literature survey for the P.hd work. The paper can be taken as the literature review of the existing paper in the area of Security, Machine learning, Health care, Neural Network, Rough Set Theory and Fuzzy Logic. The Security and privacy are the hot topic of research in Big data. Major contribution to this paper comes from Healthcare sector. The Machine learning too has its own role to in the big data analytics. This paper also covers the rough set theory and the neural network contributions to the big data. It also includes the challenges and the scope for research for the each paper. It also gives some insights to all the research paper in the area of big data, helpful for those who are doing research in big data.

2.1 Existing System:

Due to rapid development of diverse healthcare practices, various procedures used in healthcare, produce data. This healthcare data has been scaled to a bigger size, thus, there is a dire need to analyze this scaled data in an efficient manner. Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of commodity computers. Analytical processing of big data with Hadoop is very helpful in performing significant actual-point in time analysis on massive amount of data and is capable to forecast an emergency situation. In this paper Hadoop driven analysis has been performed on a diabetes case study through comparison among Pig, Hive, and Tableau The superiority of Tableau is established as it allows users with minimal statistical background to visualize the results of analysis in an easy ‘onbutton-click’ manner.

2.2 Proposed solution:

This project covers the main requirements of our project, system architecture and query-processing pipeline, Data Visualization are extract the text file in tableau,queries for exploratory data analysis are sort out, To which country most of the primary goods are sent from India,Which goods are highly export from India?,Highly paid primary goods export from india (2011-12),Highly paid primary goods export from india (2012-13), Finally compare the Comparison of Goods export from india (2011-12) and Goods export from india (2012-13).

3 THEORITICAL ANALYSIS :

3.1 Hardware requirements:

- System :Intel[R] Core[TM] i3-1005 G1 CPU @ 1.20GHZ 1.19 GHZ
- Hard Disk :500 GB
- Monitor :15 VGA Colour
- Mouse :Logitech - Optical
- RAM :4.00 GB

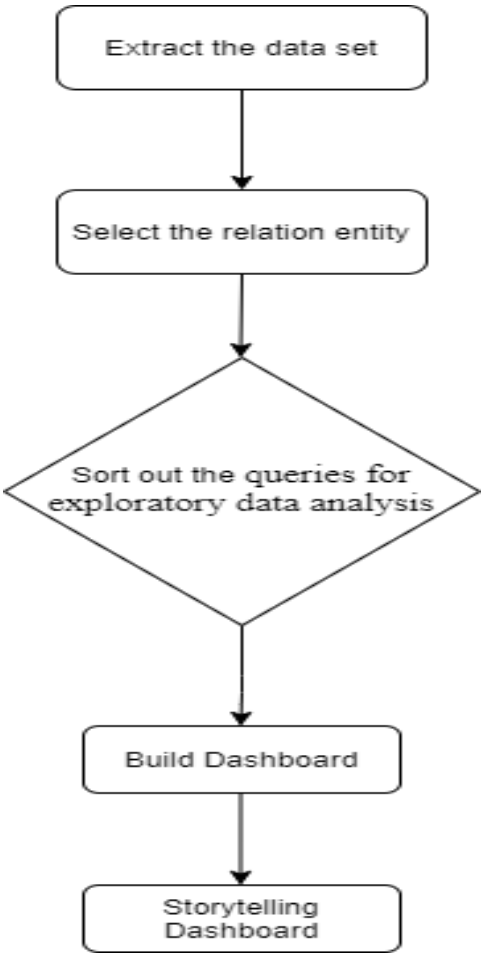
3.2 Software requirements:

- Tableau Desktop Professional 2021.1.
- Browser example: Chrome or Firefox.

4.EXPERIMENTAL INVESTIGATIONS:

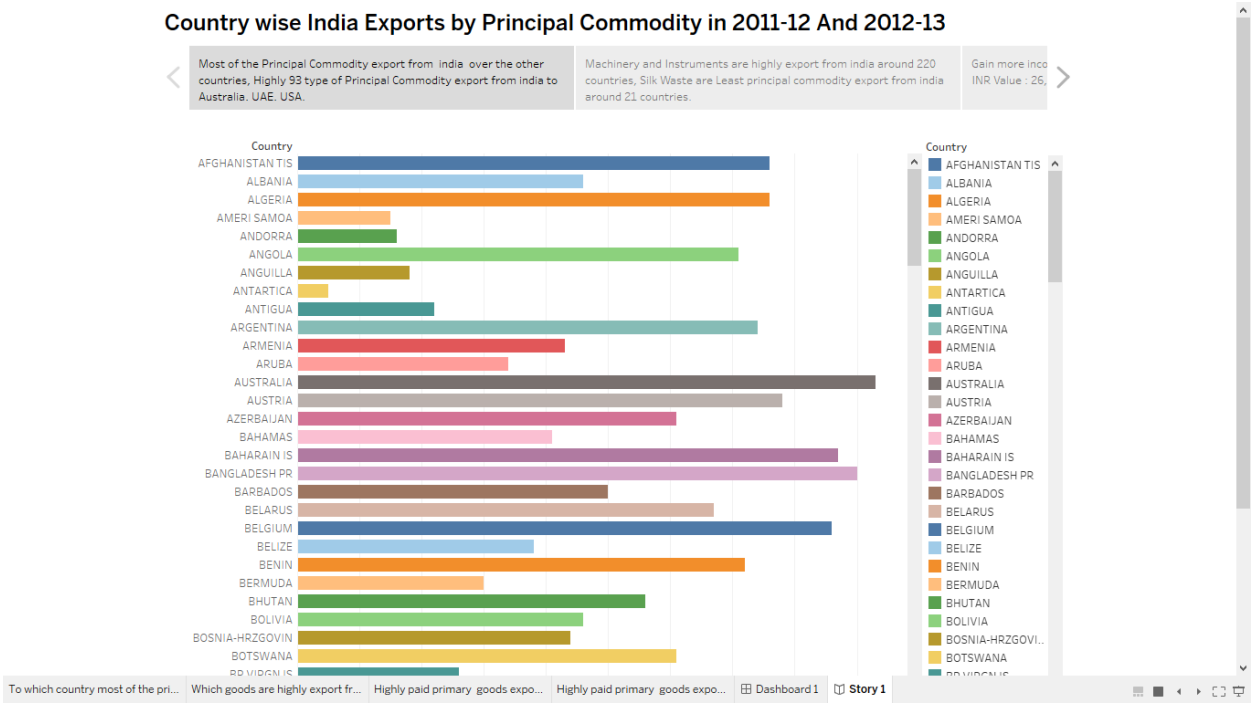
Queries for exploratory data analysis are sort out, To which country most of the primary goods are sent from India,Which goods are highly export from India?,Highly paid primary goods export from india (2011-12),Highly paid primary goods export from india (2012-13), Finally compare the Comparison of Goods export from india (2011-12) and Goods export from india (2012-13).

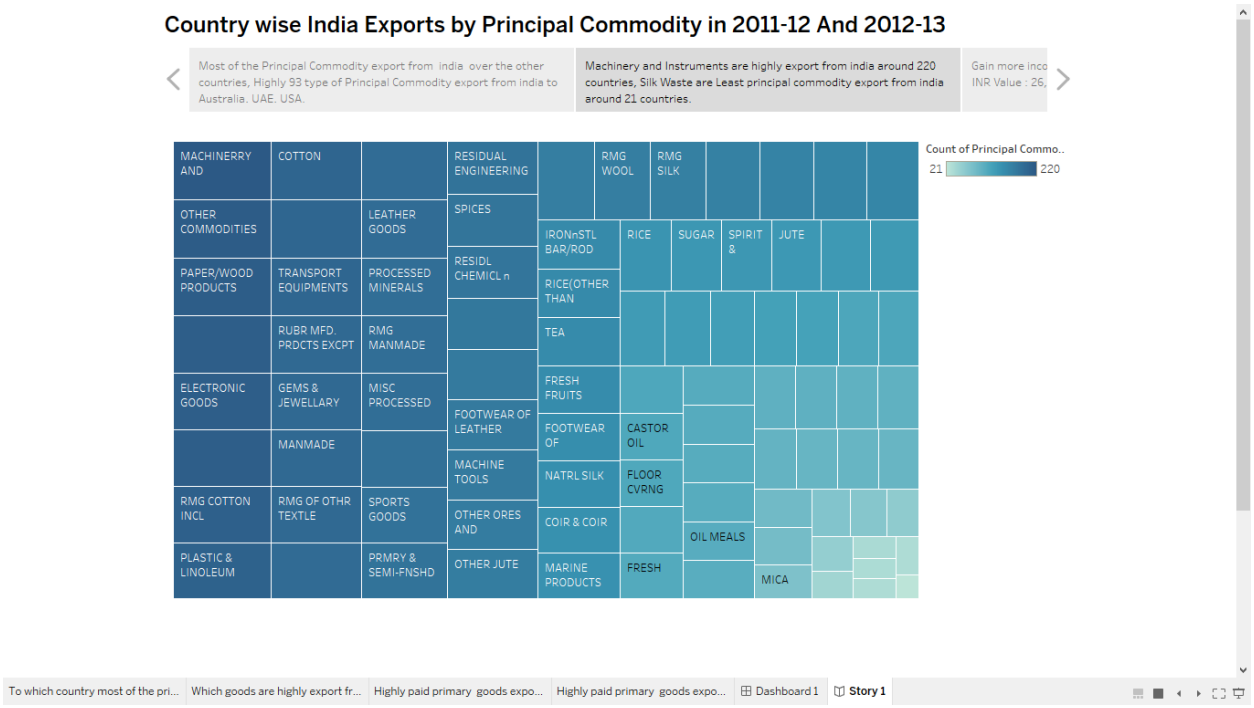
5.FLOWCHART Diagram:



6.RESULT Output of the project along with screen shots:

Here I was attached to the storytelling dashboard with screen shots.

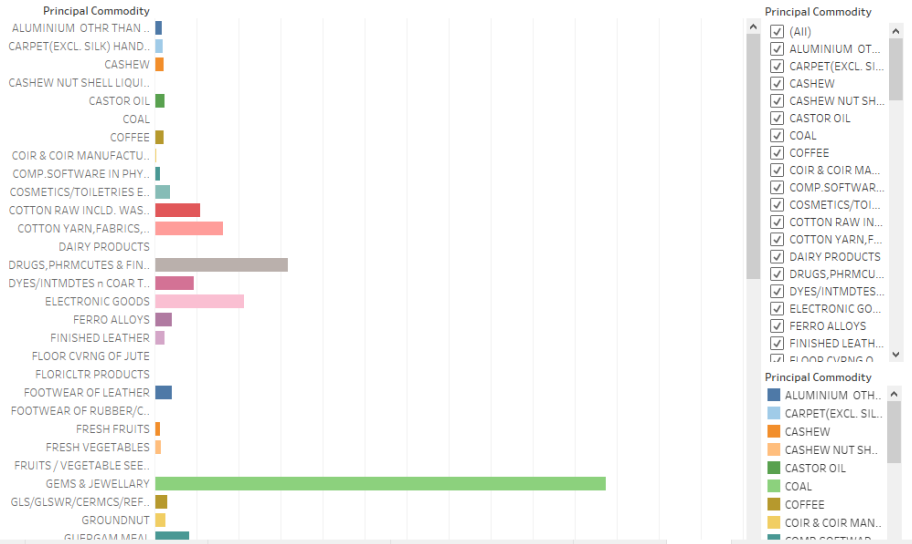




Country wise India Exports by Principal Commodity in 2011-12 And 2012-13

Machinery and Instruments are highly export from india around 220 countries, Silk Waste are Least principal commodity export from india around 21 countries.

Gain more income from Petroleum and Crude products at 2011-12
INR Value : 26,76,14,82,99,440.

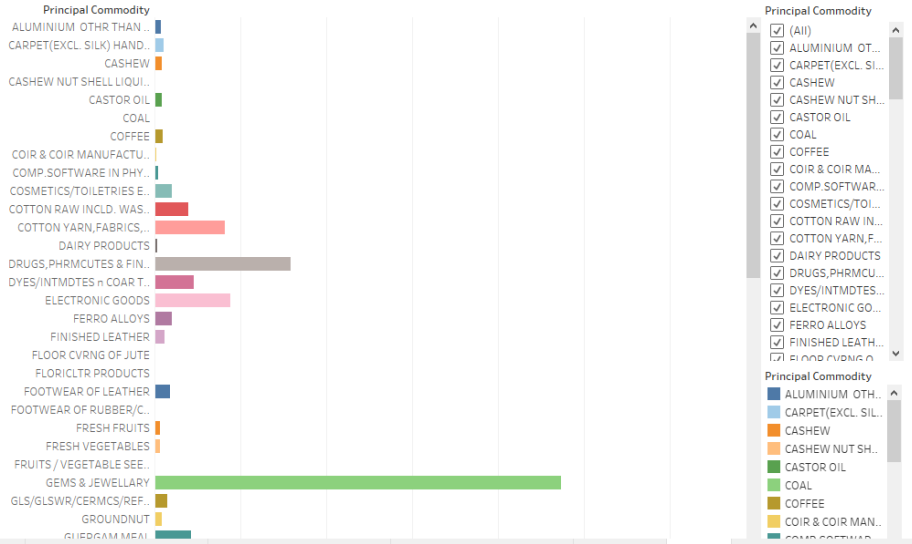


To which country most of the pri... | Which goods are highly export fr... | Highly paid primary goods expo... | Highly paid primary goods expo... | Dashboard 1 | Story 1

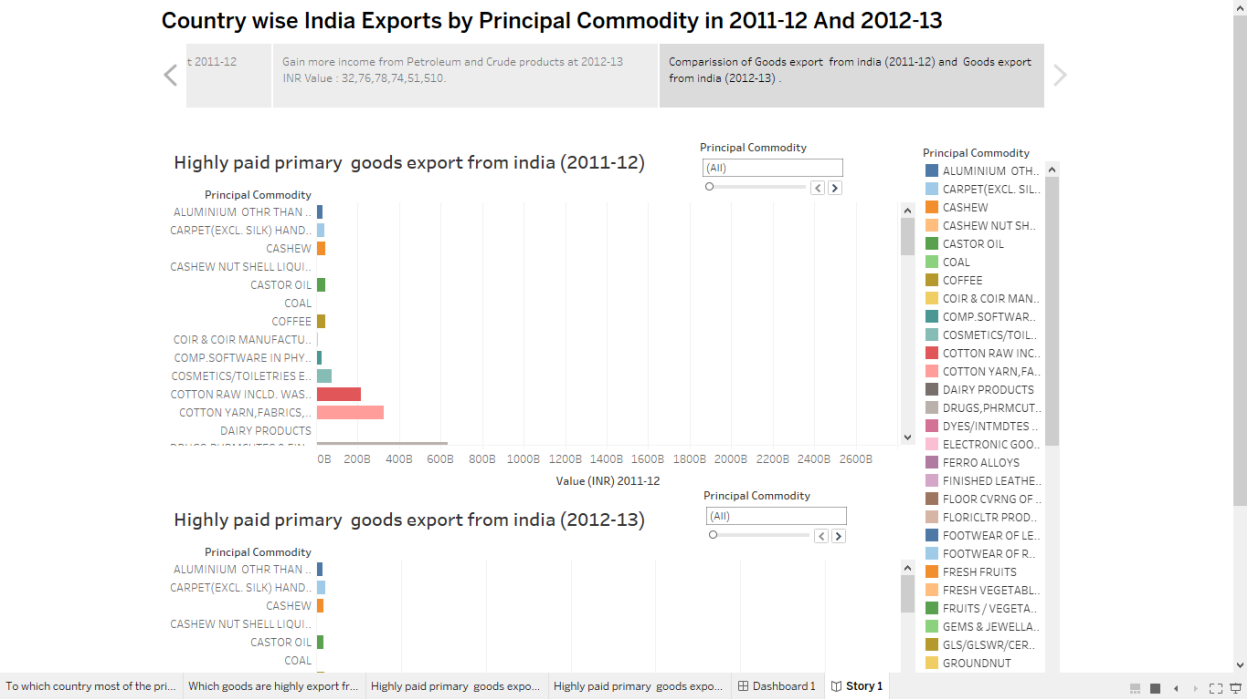
Country wise India Exports by Principal Commodity in 2011-12 And 2012-13

Gain more income from Petroleum and Crude products at 2011-12
INR Value : 26,76,14,82,99,440.

Gain more income from Petroleum and Crude products at 2012-13
INR Value : 32,76,78,74,51,510.



To which country most of the pri... | Which goods are highly export fr... | Highly paid primary goods expo... | Highly paid primary goods expo... | Dashboard 1 | Story 1



7.1.ADVANTAGES:

- Remarkable Visualization.
- Multiple Data Source Connections.
- High Performance.

7.2.DISADVANTAGES:

- High Cost.
- Poor After-Sales Support.
- Security Issues.

8.APPLICATIONS:

- Desktop.
- Public.
- Online.
- Server.
- Reader.

9.CONCLUSION:

Finally Data Visualization queries for exploratory data analysis are sort out, To which country most of the primary goods are sent from India,Which goods are highly export from India?,Highly paid primary goods export from india (2011-12),Highly paid primary goods export from india (2012-13), Finally compare the Comparison of Goods export from india (2011-12) and Goods export from india (2012-13).

Most of the Principal Commodity export from india over the other countries, Highly 93 type of Principal Commodity export from india to Australia, UAE, USA,Machinery and Instruments are highly export from india around 220 countries, Silk Waste are Least principal commodity export from india around 21 countries,Gain more income from Petroleum and Crude products at 2011-12 INR Value : 26,76,14,82,99,440,Gain more income from Petroleum and Crude products at 2012-13 INR Value : 32,76,78,74,51,510,Comparison of Goods export from india (2011-12) and Goods export from india (2012-13) .

10.FUTURE SCOPE:

Tableau is the most used software for data visualization. This software provides rapid visualizations, and as a result, it helps businesses to make decisions quickly. Tableau Desktop is a famous new generation BI (Business Intelligence) tool, which is also known as self-service data discovery and visualization tool.

11.BIBILOGRAPHY REFERENCES:

[1] P. Bhardwaj and N. Baliyan, "Big Data Analysis in Healthcare", in Smart Healthcare Systems, 1st ed., A. Sinha and M. Rathi, Ed. CRC Press(in press), ISBN 9780367030568 - CAT# K405452,2019 .

[2] Manogaran, G., Thota, C., Lopez, D., Vijayakumar, V., Abbas, K. M., &Sundarsekar, R. Big data knowledge system in healthcare. In Internet of things and big data technologies for next generation healthcare (pp. 133-157). Springer, Cham.(2017)

[3] Cao, C., Wang, W., Zhang, Y., & Ma, X. (2017, June). Leveraging Column Family to Improve Multidimensional Query Performance in HBase. In Cloud Computing (CLOUD), IEEE 10th International Conference on (pp. 106-113). IEEE.2017

[4] A. Thusoo, J.S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, R.Murthy, Hive: a warehousing solution over a map-reduceframework. Proc. VLDB Endow 2(2),1626–1629 (2009)

[5] Z. Shao, A. Thusoo, J.S. Sarma, N. Jain, Hive-a petabyte scale data warehouse using hadoop,inData Engineering (ICDE) (2010)

[6] C. Olston, B. Reed, U. Srivastava, R. Kumar, A. Tomkins, Pig latin: a not-so-foreign languagefor data processing, in Proceedings of the ACM SIGMOD International Conference onManagement of Data, ACM ,pp. 1099–1110.(2008)

[7] T. Liu, J. Liu, H. Liu,W. Li, A performance evaluation of Hive for scientific data management,inIEEE International Conference on Big Data, pp. 39–46.(2013)

[8] Wang, Y., Kung, L., & Byrd, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3-13.(2018)

[9] Tse, D., Chow, C. K., Ly, T. P., Tong, C. Y., & Tam, K. W. (2018, August). The Challenges of Big Data Governance in Healthcare. In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE) (pp. 1632-1636). IEEE.(2018,August)

[10] Popovic, J. R. Distributed data networks: a blueprint for Big Data sharing and healthcare analytics. Annals of the New York Academy of Sciences, 1387(1), 105-111.

[11] Wang, Y., & Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. Journal of Business Research, 70, 287- 299.(2017)

[12] Al Mayahi, S., Al-Badi, A., & Tarhini, A. Exploring the Potential Benefits of Big Data Analytics in Providing Smart Healthcare. In International Conference for Emerging Technologies in Computing (pp. 247-258). Springer, Cham.(2018, August)

[13] Narayanan, U., Paul, V., & Joseph, S. Different analytical techniques for big data analysis: A review. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 372-382). IEEE. (2017, August).

[14] Johri, P., Singh, T., Yadav, A., & Rajput, A. K. Advanced patient matching: Recognizable patient view for decision support in healthcare using big data analytics. In Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS), 2017 International Conference on (pp. 652-656). IEEE.(2017, December)

[15] Suo, Q., Ma, F., Yuan, Y., Huai, M., Zhong, W., Gao, J., & Zhang, A. Deep Patient Similarity Learning for Personalized Healthcare. IEEE Transactions on NanoBioscience.(2018)