Advances and Applications in Mathematical Sciences
Volume 17, Issue 1, November 2017, Pages 213-229
© 2017 Mili Publications

A SURVEY ON TOOLS USED IN BIG DATA PLATFORM

R. S. RAGHAV[1], J. AMUDHAVEL[2] and P. DHAVACHELVAN[3]

[1,3]Department of Computer Science
Pondicherry University
Puducherry, India
E-mail: vpmrags@gmail.com
dhavachelvan@gmail.com

[2]Department of CSE
KL University
Andhra Pradesh, India
E-mail: info.amudhavel@gmail.com

Abstract

The advancement of data in the current world is drastically increased, where millions of data is generated from a variety fields. This massive growth of data displays the value of data in every aspect. To tackle these millions of data, the current technology need to work effectively for understanding the three different V's of big data like variety, volume and velocity of Data. It is a difficult job for an organization to research and visualize massive volume of datasets. Every organization should allocate some methodology to have perfect insight from analysis of large volume of data. This strategy helps a company to improvise their work flow and to find out a different path from their competitors. In this survey, we discuss the major tools in big data environment by understanding the quality of tool with different factors. It also describes about the different techniques and tools currently used for the handling due sets of data.

1. Introduction

The role of big data is to gather variety of unstructured data, collected from raw data across the globe. This is a difficult process to gather information and knowledge from the raw data. To solve these issues the company deployed a technique for storing and processing the data [1]. The analysis of big data is a sequence; set of activities is needed to tackle with

2010 Mathematics Subject Classification: 68N13, 97N50.
Keywords: big data, business intelligence tools, data transfer tools, big data file system.
Received March 10, 2017; Accepted July 20, 2017

perfect solution. Thus company requires a different methodology to have a united set of solutions for analyzing big data. It also collects the data, for getting perfect decisions. Many organizations carry out these jobs by organizing the use of both commercial and open source tools. The integrated architecture for analyzing big data designed to derive a better solution for a complex problem. To analyze big data it requires each piece of information from large volumes of different data [2]. The company creates a model to study the quality of data, to know every quality of data and its features. The collection of data also improves the quality of the web service [23, 24].

2. Requirements For Understanding Big Data

Finding nature of data. In many situations, the company don't really understand the quality of data, so it requires a separate process for exploration and discovery [3].

Iteration. The data processing is based on the relationships, which is not known by the organization. The company should have a perfect methodology to discover the path for getting better decisions. It requires some iterative process to understanding the data and its complexity [4]. This can be achieved by using the big data process, where the industry creates a design which handles plenty of data with short span of time.

Flexible Storage. The usage of big data database for handling large data, it needs some special methods to organize and to spend more time to solve complex task [5].

Mining. The mining process is carried effectively in big data platform without any difficulties [3, 4]. It contains some data mining tools which easily grab the data from the large data sets. The processing of the data is held quickly without any delay.

Predicting Future. Big data analysis consist of predicting methodology used for predicting the future to take accurate decisions [6]. It also gives an idea about the requirements of the customer and to satisfy their needs without any fail.

Decision Management. The organization needs to plan a predictive model for managing decision without any trouble. It collects every detail like data velocity, volume and value of data. According to these things, the company will design a proper model, which can take better and quick decision.

3. Miscellanies Tools in Big Data Platform

The big data platform needs some special build tools for collecting large data and it also requires some effective tools for easy data transferring. It needs file system for reducing the data into small chunks for easy processing.

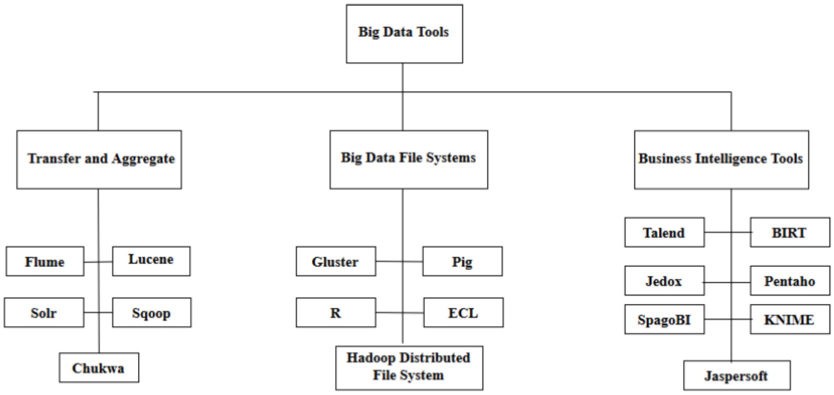

Figure 1. Classification of Big data Tools.

3.1. Data Transfer and Data Aggregation

The job of transferring and aggregating massive volume of data sets for processing [11]. The processing of unstructured data needs some effective tools for indexing a huge data set. The transferring data needs advanced tool, which can quickly process large data sets. The tool needs to have a tendency for processing complex data.

Lucene

It depends upon the collaborative processes which has high quality software. It produces very quick indexing and effective searching for massive dataset modern architecture and it can be altered according to the variety of data sets and it also defines as OS Independent [12]. The queries processing contains quick many powerful query types: phrase related queries, wildcard based queries, proximity queries range based queries etc.

Solr

It is defined as a platform for carrying effective searching technique and it is completely based on the Lucene tools. It provides a powerful searching process in a large volume of data. It collects all types of data from variety of websites to understand the user needs [13, 22]. It is OS Independent, where it doesn't need any platform for processing the tools.

Sqoop

The term Sqoop the tool consumed for migrating large sets data, it carries operation of importing or exporting from the variety of data sources. The data sources can be traditional database or NoSQL database and data warehouses [15]. It completely used for exporting data from external system into HDFS and the transferring of data to H Base or hive carried by Sqoop.

Flume

It gathers data from variety of sources from different environment. The aggregation of data gives reliability and durability of data for powerful processing [16]. The data gathered from distributed sources is centralized by using flume and the flow carries by migrating massive volume of data in a single space. It tells flume to collects quick data processing and data movement to the Hadoop's file system.

Chukwa

The working process of tool is to collects data from widely spreaded system. It is placed on top of HDFS and Map Reduce [21]. The tool can use for data visualization to gain deep insights about the data. As a result it gives a clear cut idea for the organization to take better results.

3.2. Open Source Big Data File Systems and Programming Languages

Mostly the tools used in big data platform are meant to be an open source. The file system used in big data analytics is to reduce the complexity, by breaking down the large data into small chunks [25, 26]. This process is carried by the Master node and the job is allocated to the slave node for effective processing. The different set of programming languages used in big data platform, for carrying query processing in effective way.

Gluster

The Gluster is a file, which contains object for storing major volume of data. The capacity of data processing in the gluster is 72 bronto bytes. The extension of gluster can be used to achieve the demerits of HDFS.

Hadoop Distributed File System

The HDFS is one of the key components used in the hadoop ecosystem; it handles more nodes in same rack. It can handle multiple racks in the cluster and it is also known as a primary storage system for Hadoop. The data replication process carried onto several nodes in a cluster to have reliable, fast performance.

Quantcast File System (QFS)

It is said to be another source to the Hadoop Distributed File System (HDFS). This QFS can handle massive amount of data in terms of batch processing. He Processing power used by the OFS is less and the error correction rate of the QFS is high when compared to HDFS.

LUSTRE

It is other file system tool used for cluster computing of high volume of data. The type of processing used by the LUSTRE is parallel distributed file system. It provides high performance file systems for handling the clusters [19]. The range of the cluster varies in size from small workgroup clusters to large work group. The lustre has the ability how to tackle variety of cluster to produce proper outcome.

Pig

The Pig is known for a high-level language for carrying executing of data analysis programs. The program evaluation carried by combining together with infrastructure. The design of the pig program can easily handle the parallelization for huge data sets [20]. It makes to write the coding easily, by understanding and maintain programs, to complete the major task assigned by the master node to the slave nodes in a quick way.

R

R is a programming language for a big data platform to carry statistical computing and complex graphics task similar to *S*. The platform includes a set of tools, which has the ability to manage difficult task to solve in an efficient way. The data manipulation process carried in *R* is very easy and the evaluation of data can be easily done by the *R* platform. It also gives deep insights about the data by analyzing it a proper way, and presenting in terms of charts and graphs.

ECL

ECL ("Enterprise Control Language") is the language for working with HPCC. IT contains a Varity set of tools for handling large file system. It contains an IDE and a debugger for error correction.

3.3. Big Data Business Intelligence Tools

The Business analytics of the company used to calculate the performance. It used for evaluating their status in the market and to know the area to improve their work flow [13]. It uses statistical methods for specific product or process by the company. The main theme of company is to run the business analytics for monitoring their business flow and to find the demerits of the existing processes and highlight meaningful data.

Talend

Talend consist of variety of business intelligence and data warehouse products. Talend Open Studio for Big Data contains a data integration tools which can be utilized by the hadoop ecosystem components. It teaches to handle the metadata connection with the database; it gives clear idea about the retrieve schemas from *DB* to metadata using a proper connection. It also indicates the integration process carried by the *DB* and Meta data and some other components of Hadoop.

Jaspersoft

Jaspersoft creates a cost effective and flexible environment for handling massive unstructured data without error. They are mostly deployed in the companies with more number of user interactions and the quick communication of queries. It creates a bridge for data analyst to have deep insights about the data [5, 8]. The variety of data related application and to the business intelligence people can be highly interacted by using this business intelligence tool.

Jedox

Jedox is known as a streamlines consolidation of data with proper reporting of the work flow carried by the data analyst. It enhances the business processes by working in an effective way for large data volume. The work flow is carried transparently, where the error can be easily identifies, it can be solved in a short span of time. This process makes the tool as one of the main business intelligence tool in large big data platform.

Table 1. Comparison of Tools for Data Transfer and Data.

Features	Lucene	Solr	Sqoop	Flume	Chukwa
Scalable	High	High	High	High	Avg
Performance	Very High	High	High	High	Low
Powerful	High	High	Avg	High	Avg
Accurate	High	High	High	Avg	Low
Queries	High	High	Avg	Avg	Avg
Cross-Platform Solution	Java	XML JSON and HTTP	Java	Java	Java
Fault Tolerant	Avg	High	High	High	Avg
Security and Monitoring	Avg	High	Avg	Avg	High

Table 2. Comparison of Tools for File Systems and Programming Languages.

Features	Gluster	Hadoop Distributed File System	Quantcast File System	LUSTRE	Pig	R	ECL
Scalable	High	High	High	High	High	High	Avg
Performance	Very High	High	Very High	Avg	High	High	High
Programming	Hard	Easy	Easy	Easy	Easy	Easy	Easy
Fault Tolerant	Avg	High	High	High	Avg	High	Avg
Data Handling	Avg	High	High	Avg	High	Avg	High

Table 3. Comparison of Tools for Business Intelligence Tools.

Features	Talend	Jaspersoft	Jedox	Pentaho	SpagoBI	KNIME	BIRT
Scalable	High	High	High	High	Avg	High	Avg
Performance	High	Avg	High	High	Avg	High	High
Security	High	Low	High	Low	Low	High	Low
Decision Making	Avg	High	Avg	High	Avg	High	Avg
Data Handling	High	High	High	Avg	Avg	High	High

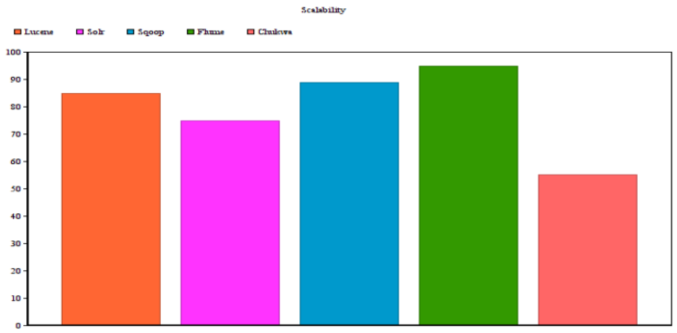

Figure 2. Scalability of Data Transfer and Data Aggregation.

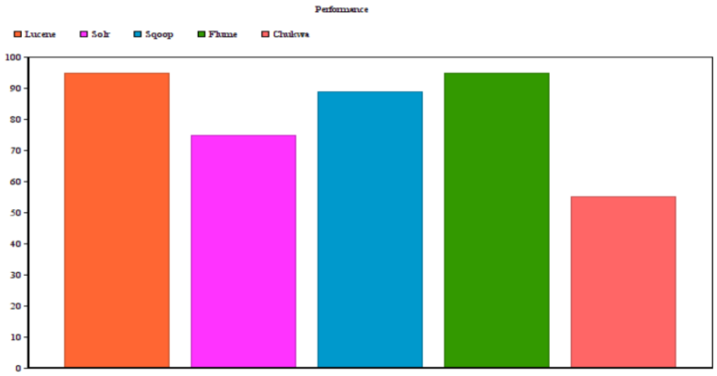

Figure 3. Performance of Data Transfer and Data Aggregation.

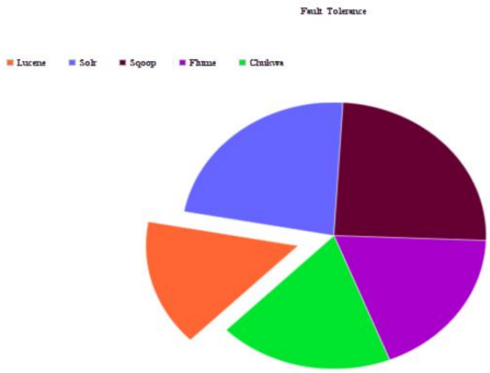

Figure 4. Fault Tolerance of Data Transfer and Data Aggregation.

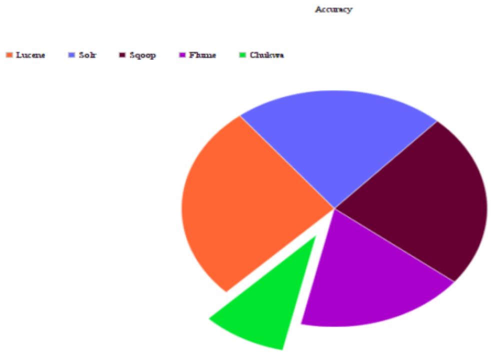

Figure 5. Accuracy of Data Transfer and Data Aggregation.

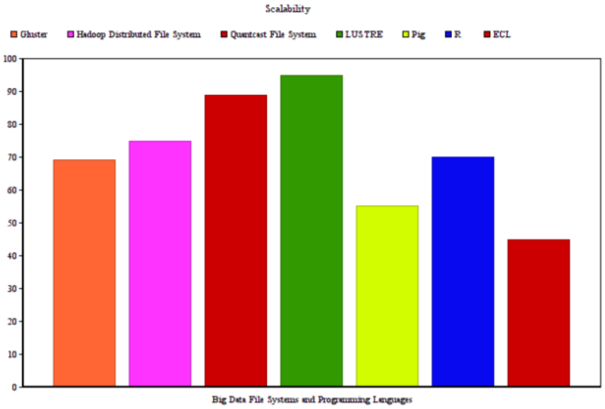

Figure 6. Security of Data Transfer and Data Aggregation.

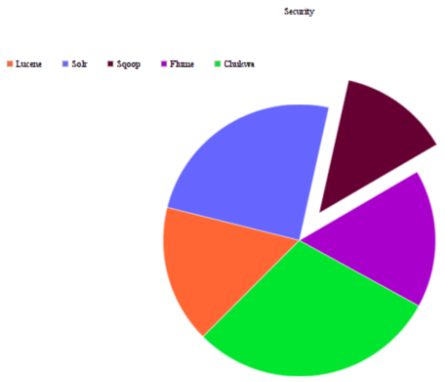

Figure 7. Scalability of File Systems Programming Languages.

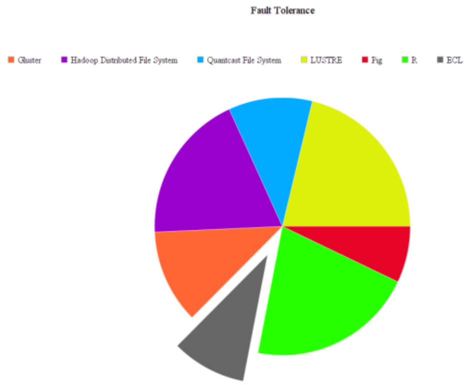

Figure 8. Fault Tolerant of File Systems Programming Languages.

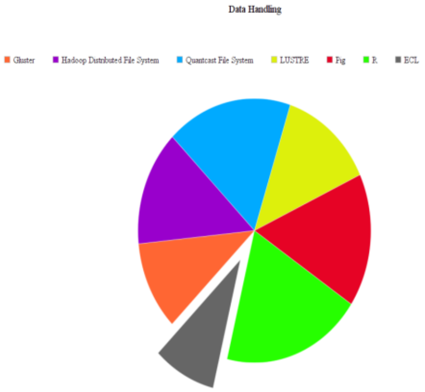

Figure 9. Data Handling of File Systems Programming Languages.

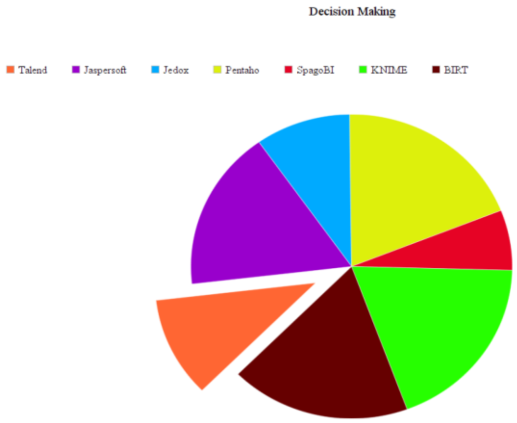

Figure 10. Scalability of Business Intelligence tool.

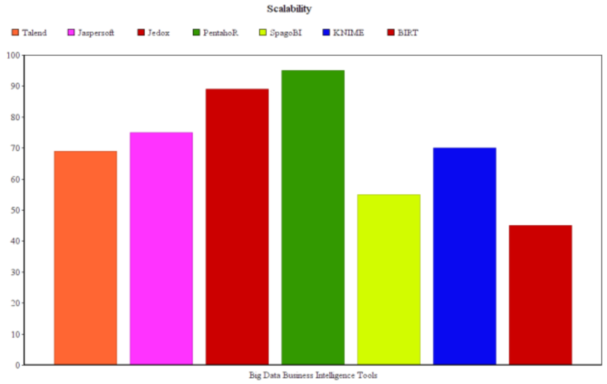

Figure 11. Decision making of Business Intelligence tool.

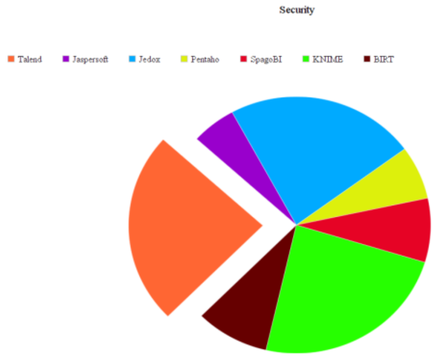

Figure 12. Security and Monitoring of Business Intelligence tool.

Pentaho

It is known as a Comprehensive Data Integration and Business Analytics Platform which is used by more than 10,000 companies. It provides information about the business and act as a big data analytics tools [19]. It also contains some data mining process for collecting more number of data and processing it in an easy way. The analysts can easily understand the business flow with high data accuracy, by visualizing the reports or by using the dashboard functionalities.

SpagoBI

It is an open source business intelligence tool particular used for tackling the huge variety of data generated by the dynamic changing environments. It gives knowledge and expertise sharing of the data for the companies. The integration of technologies into a clumsy moving of data with large environment of complex queries can be easily solved by tool.

KNIME

The Konstanz Information Miner offers user-friendly interaction, easily integration of data, and effective but simple way of processing the queries. The analysis of data carried powerfully by giving deep insights to the client and user to have high knowledge about the data. The exploration of business work flow and end to end analytics can easily carried by using the tool.

BIRT

It is also known as "Business Intelligence and Reporting Tools" which is mainly used for data visualization and business intelligence. It is completely based on Eclipse which adds reporting features to Java applications.

4. Discussions

In this section we discuss which type of tools used by the organization for carrying better decision. The tools should give some prior knowledge about the data and helps the business people should gain deep insights about multiple of data generated from widely spread devices and system. The Table 1 express about the Data Transfer and Data Aggregation tools used in big data environment. It gives an idea to the user by considering some key factors of the tools such as Scalability of users or loads balancing, Performance of the tool, Powerful while handling large volume of data to make quick results, Queries processing, Cross Platform solution Security and monitoring of the error, and fault tolerant. These are some of the key factors noted by the business people while processing large amount of data. In Figure 2 the scalability factor is considered, where the tools used for data transfer is take place. It indicates how the tool can be tackling more number of users in an effective without facing any problem. According to factor the tools discussed above have perfect solution to handle multiple data in short span of time. It perfectly balances the load without showing any dump on the operational time. The Figure 3 explains about the performance of the tool when there is a presence of multiple unstructured data. From the previous factors we can understand the above mentioned tools can satisfy more number of users and unstructured data. So the performance of the tools varies slightly and they don't show any vast difference in processing the data. In Figure 4 fault tolerance factors is shown, even though they can more volume of data in an effective way. The tools face some problem while handling error related problem. The tools Luence don't have the ability to find error for some types of data, where it will take time to solve the issue. The Figure 5 shows accuracy factor, here the term accuracy refers to read the data and help the user to produce deep insight, where the flume tool faces some problem where it take some time to show accuracy. The comparison of above mentioned tools flume gives low accuracy, and there are some tools which doesn't give proper accuracy to the business people. The Figure 6 discuss about the security and monitoring of numerous data, the tool should show monitor the movement of data in frequent way, to help the business people to identifying the issues and correct the error within short duration.

The Figure 7 shows the scalability of data for Big Data File Systems and Programming Languages. The file system is sued for handling large data sets in an effective way; the procedure for handling these data is reducing into small files. The process of manipulating data is allocated to their slave nodes, to get quick output. The load balancing concept is done effectively by tackling the more number of users. The programming language has the same vision to allocate the jobs quickly and it fetch the response quickly for the queries. In Figure 8 fault tolerance of the tool and programming languages are shown, the file system mentioned in the Table 2 shows the low amount of fault tolerant tool. The integration of the file system and programming language is done by assuming the language is used for allocating the data to the slave nodes by the file system. The Figure 9 describes about the handling of data, this factor explains about the quality and processing of data by the tool referred in the Table 2. The tool which has low level of handling data is shown out of the pie chart.

The Figure 10 explains about the scalability of the business intelligence tool, by refereeing the Table 3 the user can clearly understand which tool will give better scalability for handling large number of data coming from different fields. The Figure 11 describes about the decision making of the tool, if an analyst needs an idea about the product they should have a deep knowledge about the manipulation of data. To achieve this task the business intelligence tool used some clear dashboards or reports giving tools, to fetch the proper information from the derived output. This strategy helps the business people to know the value of each data and error occurred by the improper data. It also helps the analyst to remove unwanted data from the reports to makes the result as a finite one, to make better decisions. The Figure 12 shows the security of the tool; where it explains about the nature of tool by exporting unnecessary data which suppress the performance of the tool. This helps the decision makers to avoid the unwanted data or data which contains some unrelated information. The tool which doesn't have the ability to show the unrelated data will not give fruitful result to the business intelligent people and they will not gain any knowledge from the results.

5. Conclusions

The value of data in world is massive and they need some skilful processing strategies to handle large amount of data for fetching quick and effective results. The handling and analyzing the massive amount of data, the company need to carry pre process like proper extraction of data from variety of data sources. The organization should understand the basic 3v's of data such as volume, variety, velocity and value etc. According to it, the company should select the proper tool, database, scripting language and visualization tool for carrying powerful analyses. In this paper we discussed about the variety of tools and its features. This survey helps the business people to understand every piece of information and its work flow. This paper shows how the company can process and analyze data it by knowing the factors of the tool, to improve their business value. By selecting a perfect tool and programming language the company can gain deep insights about the nature of data and they can take quick decisions.

Acknowledgements

This work is a part of the Research Projects sponsored by Visvesvaraya Ph.D Scheme for Electronics & IT, Ministry of Electronics & Information Technology, India, Reference Nos.: PHD-MLA-4(44)/2015-16, dated August 2015. The authors would like to express their thanks for the financial supports offered by the Sponsored Agency.

References

[1] C. L. Philip Chen and Chun-Yang Zhang, Dataintensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences 275 (2014), 314-347

[2] T. Giri Babu and G. Anjan Babu, A Survey on Data Science Technologies and Big Data Analytics, International Journal of Advanced Research in Computer Science and Software Engineering 6(2) (2016), 322-327.

[3] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl and D. Keim, Visual analytics for the big data era-A comparative review of state-of-the-art commercial systems, In Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on (pp. 173-182), IEEE. (2012).

[4] Guo-Dao, Rong-Hua Liang and Shi-Xia Liu, A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges, Journal of Computer Science and Technology 28(5) (2013), 852-867 DOI 10.1007/s11390-013-1383-8

[5] Ekaterina Olshannikova, Aleksandr Ometov, Yevgeni Koucheryavy and Thomas Olsson, Visualizing Big Data with augmented and virtual reality: challenges and research agenda, Journal of Big Data (2015), 2-22.

[6] Ravi, Likhitha, et al., A Survey of Visualization Techniques and Tools for Environmental Data, Proceedings of the 2013 Intl. Conference on Computers and Their Applications (CATA 2013).

[7] Shixia Liu, Weiwei Cui, Yingcai Wu and Mengchen Liu, A survey on information visualization: recent advances and challenges, The Visual Computer 30(12), 1373-1393.

[8] R. Mahalakshmi and S. Suseela, Big-SoSA: Social Sentiment Analysis and Data Visualization on Big Data, International Journal of Advanced Research in Computer and Communication Engineering 4(4) (2015).

[9] S. Syed Fiaz, N. Asha, D. Sumathi and A. S. Syed Navaz, Data Visualization: Enhancing Big Data More Adaptable and Valuable, International Journal of Applied Engineering Research 11(4) (2016), 2801-2804.

[10] M. Khan and S. S. Khan, Data and Information Visualization Methods and Interactive Mechanisms: A Survey, International Journal of Computer Applications 34(1) (2011), 1-14.

[11] P. Simon, The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions, Harvard Business Review 13 (2014), 1-8.

[12] C. L. P. Chen and C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences 275(10) (2014), 314-347.

[13] B. Porte, Visualizing Big Data in Drupal: Using Data Visualizations to Drive Knowledge Discovery, Report, University of Washington, (2012), 1-38.

[14] T. A. Keahey, Using visualization to understand big data, Technical Report, IBM Corporation (2013), 1-16.

[15] P. Fox and J. Hendler, Changing the Equation on Scientific Data Visualization, Science 331(11) (2011), 705-708.

[16] B. Otjacques, UniGR Workshop, Big Data- The challenge of visualizing big data, Report, Gabriel Lippmann (2013), 1-24.

[17] Hazen, T. Benjamin et al., Big Data and predictive analytics for supply chain sustainability: A theory-driven research agenda, Computers & Industrial Engineering 101 (2016), 592-598.

[18] Wang, Gang, et al., Big data analytics in logistics and supply chain management: Certain investigations for research and applications, International Journal of Production Economics 176 (2016), 98-110.

[19] Sofiya Mujawar and Aishwarya Joshi, Data Analytics Types, Tools and their Comparison, International Journal of Advanced Research in Computer and Communication Engineering 4(2) (2015),

[20] Tsai, Chun-Wei, et al., Big data analytics: a survey, Journal of Big Data 2.1 (2015).

[21] R. Raju et al., A heuristic fault tolerant Map Reduce framework for minimizing make span in Hybrid Cloud Environment, Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on. IEEE, 2014.

[22] J. Amudhavel, et al., Survey and Analysis of web service Composition Strategies: A state of art performance study, Indian Journal of Science and Technology 9(11) (2016).

[23] J. Amudhavel, et al., A comprehensive analysis and performance assessment on QoS aware web service selection algorithms, Communication Technologies (GCCT), 2015 Global Conference on. IEEE, 2015

[24] M. Rajeswari, et al., Appraisal and analysis on various web service composition approaches based on QoS factors, Journal of King Saud University-Computer and Information Sciences 26(1) (2014), 143-152.

[25] J. Amudhavel, et al., Big Data Scalability, Methods and its Implications: A Survey of Current Practice, Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015). ACM, 2015.

[26] P. Karthikeyan, et al., A comprehensive survey on variants and its extensions of big data in cloud environment, Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015). ACM, 2015.