ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

Practical Evaluation and Comparative Study of Big Data Analytical Tools

Dr. Sailesh .S. Iyer, Dr. Kamaljit Lakhtaria
Kadi Sarva Vishwavidyalaya, Gandhinagar, India
Gujarat University, Ahmedabad, India

ABSTRACT: Big Data Analytics has emerged as an industry which has great applications for the Industry at large. Big Data Architecture is studied to analyze the existing environment and suggest changes to optimize the same.

However big data challenges include capture, storage, search, sharing, transfer, analysis, and visualization. Many Big Data Analytic Tools and Visualization tools are available. This paper is used to analyze and compare various tools and provide a practical evaluation of the above tools.

KEYWORDS: Big Data Analytics, Search, Storage, Sharing, Transfer, Architecture, Visualization, Practical Evaluation**.**

I. INTRODUCTION (DIGITAL ANALYTICS)

Digital Analytics has had a major impact on corporate decision making. Traditional data has been replaced by Big Data. The characteristics of Big data like velocity, veracity, value and volume make it challenging to process Big data. Digital Analytics [1] overview is given below:

Fig.1 Digital Analytics [1] an Overview.

The various applications of Big Data Analytics be it Retail, Healthcare, Education, etc. make it imperative for any organization to deploy Big Data Analytics in one form or another. 87% of enterprises believe Big Data analytics will redefine the competitive landscape of their industries within the next three years. 89% believe that companies that do not adopt a Big Data analytics strategy in the next year risk losing market share and momentum. Selection of an appropriate Big Data tool is a key driver to growth in an organization.

Rexer Analytics Survey[2] reveals startling facts about Analytic Tools being used. Most of the widely and commonly used Analytic tool is Excel at 75.6% followed by Statistical and Mining Platform R at 35% and SAS @34.1% forming major players.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

Some of the smaller player include MS Access, SPSS, Tableau, SQL, Python, Statistica, SAP, Matlab etc..

Python has been consistently growing and in this survey its application has been limited to Data Science. R is in fourth place in growth, and given its second place in overall market share, it is in an enviable position. SPSS and SAS usage has declined considerably. Hadoop is being replaced by Spark and H2O.

.

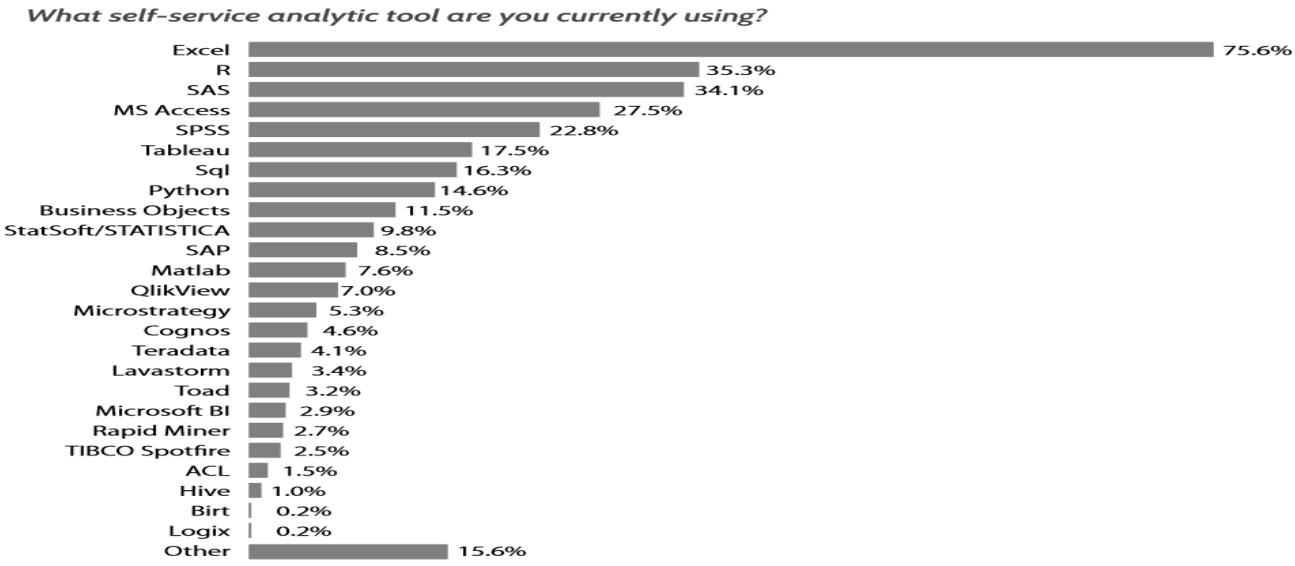

Fig 2. Analytics tools used by respondents to the 2015 Rexer Analytics Survey[2].

Tableau being a Visualization tool also has been widely accepted in the market and stands at a whopping 17.5% which is higher given the fact that it is used for Visualization

Figure 3 [2] displays Tool wise Salary survey statistics have been provided which indicate market acceptability of each tool.

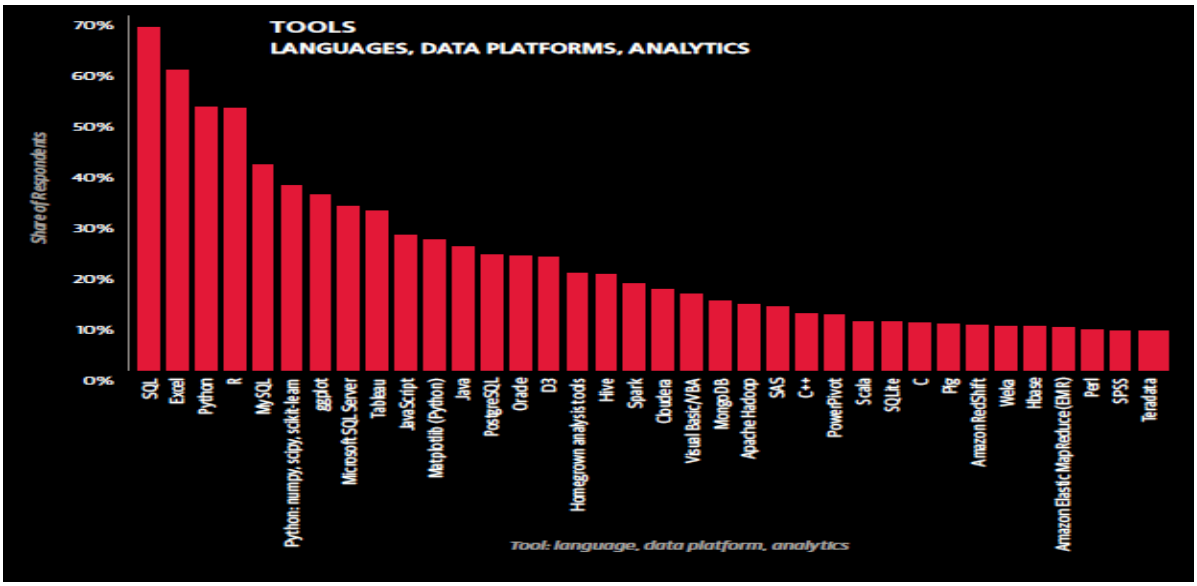

Figure 3. Tools used by 2015 respondents to O'Reilly 2015 salary survey[2].

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

II. RELATED PAPER AND TOOL SURVEY

The number of Scholarly articles authored in 2015 [2] with reference to the Analytic Software is shown in Figure 4. The number of articles is lead by SPSS which has major share followed by R, SAS, Stata, MATLAB, Java, Statistica, Hadoop, Python etc..

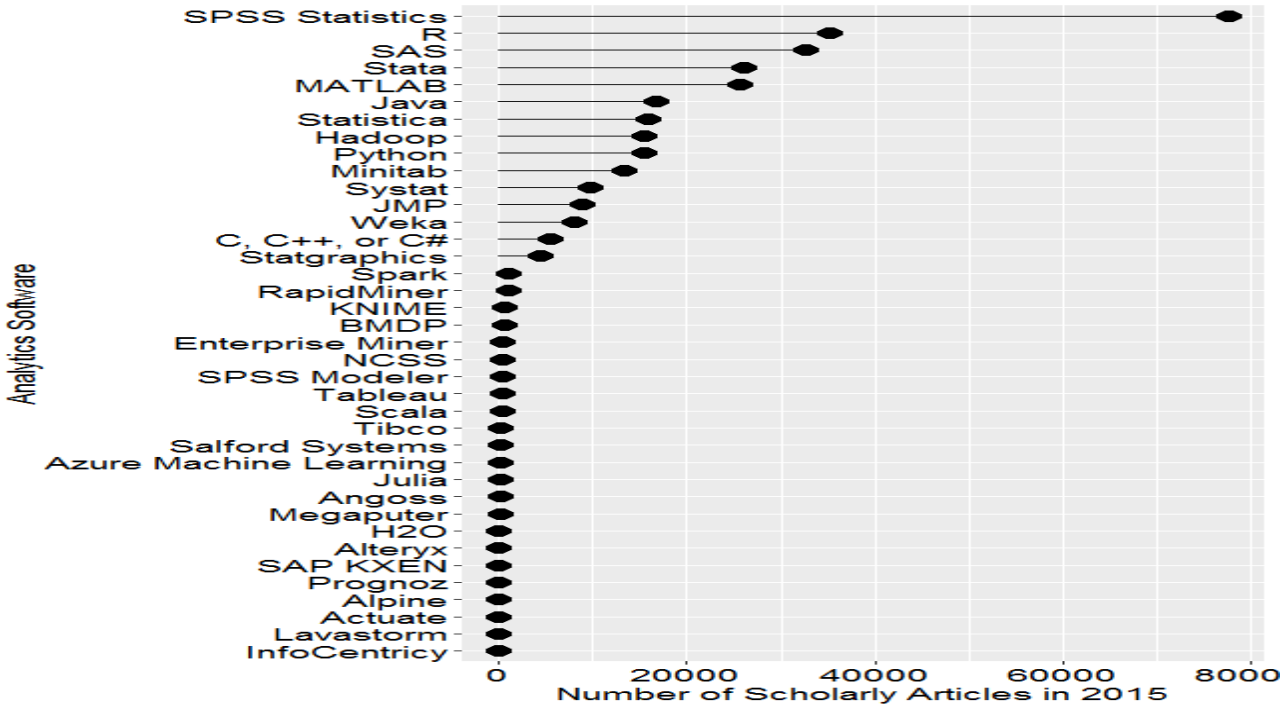

Figure 4. [1]Scholarly articles for each software package 2015.

Data Mining also form an integral part of Data Analytics. Table-1 [3] represents the number of papers published for DM Techniques year wise from year 2010 to 2015.

Year	Classification	Clustering	Regression
2010	21	12	15
2011	25	13	16
2012	28	17	21
2013	30	13	25
2014	33	13	26
2015	37	15	22

Table-1[3] Data Mining techniques papers published year wise.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

The graphical representation of the Data Mining Techniques is shown in Figure 5[3].

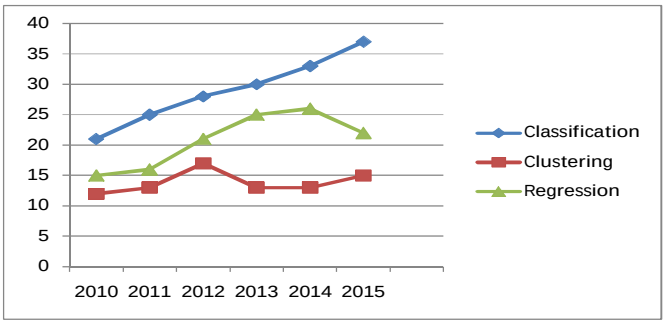

Figure-5 [3]. Graphical representation of DM Techniques.

III. BIG DATA ARCHITECTURE AND TOOLS COMPARISON [4]

Figure-6 shows the entire Big Data Architecture. The architecture is broadly divided into four main parts:

1. Data Storage and Preprocessing.
2. Data Warehouse.
3. Visualization and Analytics.
4. Online business Apps.

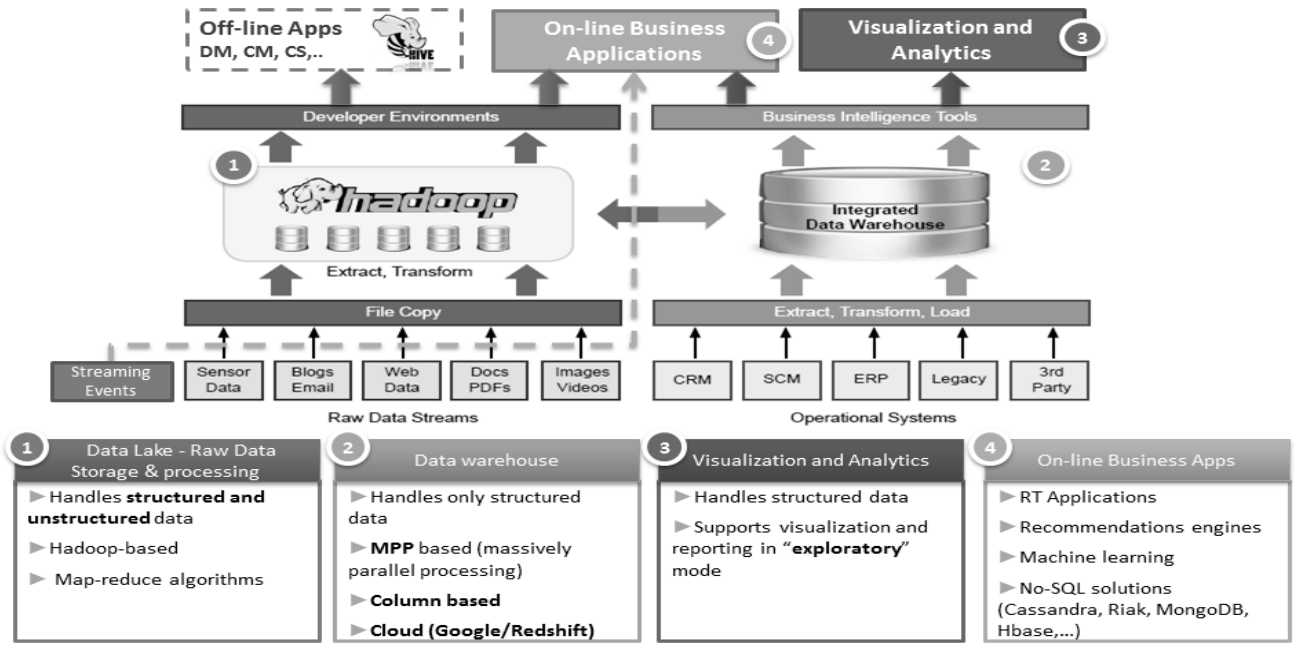

Figure 6[2] Big Data Architecture.

There are various tools available for all parts of Big Data namely Data Storage, Data Warehouse, Visualization and Analytics.
Some of the tools are discussed as follows:

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

1. Pentaho Business Analytics.
2. Talend Open Studio.
3. JasperSoft BI
4. Tableau.
5. Qlik.
6. Actuate.

Big Data Tools	Pros	Cons
Jaspersoft	Complete BI Solutions. Costing very low.	Less used in Companies. Below average performance and data volumes.
Pentaho	Ranking high among available tools. Cost very low.	Customer feedback and support below average. Not easy to use.
Tableau	Customer ranking high. Reusability, Embedding high.	High maintenance/support fees. High Governance Issues.
Qlik	Visualization Analytics high Easy to use Strong Dashboard & Big Data support	Not Enterprise ready. Risk to current customers.
Actuate	User friendly. Extended big data connectivity.	Non-Interactive. Not suitable for dashboards and visualization.

Table-2. Big Data Tools Comparison [12]

1. Jaspersoft BI Suite[1]:

The Jaspersoft package is open source leaders for producing reports from database columns. The JasperReports Server now offers software to suck up data from many of the major storage platforms, including MongoDB, Cassandra, Redis, Riak, CouchDB, and Neo4j. Hadoop is also well-represented, with JasperReports providing a Hive connector to reach inside of HBase.

2. Pentaho Business Analytics:

Pentaho software platform was a report generating engine, branching into big data by making it easier to absorb information from the new sources. Pentaho Tools can combine with NoSQL databases such as MongoDB and Cassandra. Once the databases are connected, you can drag and drop the columns into views and reports as if the information came from SQL databases.
Pentaho also provides software for drawing HDFS file data and HBase data from Hadoop clusters. One of the more intriguing tools is the graphical programming interface known as either Pentaho Data Integration.

3. Karmasphere Studio and Analyst:

Karmasphere Studio is a set of plug-ins built on top of Eclipse. It's a specialized IDE that makes it easier to create and run Hadoop jobs.

Karmasphere also distributes a tool called Karmasphere Analyst, which is designed to simplify the process of plowing through all of the data in a Hadoop cluster. It comes with many useful building blocks for programming a good Hadoop job, like subroutines for uncompressing Zipped log files.

4. Talend Open Studio:

Talend offers an Eclipse-based IDE for stringing together data processing jobs with Hadoop. Its tools are designed to help with data integration, data quality, and data management, all with subroutines tuned to these jobs.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

Talend Studio allows you to build up your jobs by dragging and dropping little icons onto a canvas.

Talend also maintains TalendForge, a collection of open source extensions that make it easier to work with the company's products. Most of the tools seem to be filters or libraries that link Talend's software to other major products such as Salesforce.com and SugarCRM.

5. Skytree Server:

Skytree offers a bundle that performs many of the more sophisticated machine-learning algorithms.

Skytree Server is optimized to run a number of classic machine-learning algorithms on your data using an implementation the company claims can be 10,000 times faster than other packages. It can search through your data looking for clusters of mathematically similar items, then invert this to identify outliers that may be problems, opportunities, or both.

6. Tableau Desktop and Server:

Tableau Desktop is a visualization tool that makes it easy to look at your data in new ways, then slice it up and look at it in a different way. The tool is optimized to give you all the columns for the data and let you mix them before stuffing it into one of the dozens of graphical templates provided.

Tableau uses Hive to structure the queries, then tries its best to cache as much information in memory to allow the tool to be interactive. Tableau offers an interactive mechanism so that you can slice and dice your data again and again.

7. Splunk:

Splunk creates an index of your data as if data were a book or a block of text.

Splunk comes already tuned to my particular application, making sense of log files, and it sucked them right up. It's also sold in a number of different solution packages, including one for monitoring a Microsoft Exchange server and another for detecting Web attacks. The index helps correlate the data in these and several other common server-side scenarios.

8. Lumify:

Lumify analyze relationships, automatically discover paths between entities, and establish new links in 2D or 3D.

Pluggable map providers [4] allow integration of corporate GIS solutions. Organize work into separate workspaces that can be shared with colleagues. Updates are pushed to everyone viewing the workspace in real-time.

9. SpagoBI

SpagoBI [5] claims to be "the only entirely open source business intelligence suite." Commercial support, training and services are available. Operating System: OS Independent.

10. Terracotta

Terracotta's "Big Memory"[6] technology allows enterprise applications to store and manage big data in server memory, dramatically speeding performance. The company offers both open source and commercial versions of its Terracotta platform, BigMemory, Ehcache and Quartz software. Operating System: OS Independent.

11. Avro

Apache Avro is a data serialization system based on JSON-defined schemas. APIs [7] are available for Java, C, C++ and C#. Operating System: OS Independent.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

12. Oozie

Apache project designed to coordinate the scheduling of Hadoop jobs. It can trigger jobs at a scheduled time or based on data availability. Operating System: Linux, OS X.

13. Zookeeper

Zookeeper [8] is "a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services." APIs are available for Java and C, with Python, Perl, and REST interfaces planned. Operating System: Linux, Windows (development only), OS X (development only).

These are few commonly used tools in arena of Business Intelligence, Visualization and Big Data Analytic Tools. These tools are compared on various parameters like OS, API's, Commercial support and Visualization capabilities.

Table 3 gives a comparative study of Data Science tools involved in Data Mining and Visualization [9] like R, Weka, Orange and Rapid Miner. The evaluation parameters consisted of Data Mining techniques [10] like Association Rule Mining, k-Means, Decision Tree, Naive Bayes Classifier, Time Series, Text Analytics, Big Data Processing and Visualization of Data

Evaluation	R	Weka	Orange	Rapid Miner
Association Rule Mining	Yes	Yes	Yes	Yes
K-Means	Yes	Yes	Yes	Yes
Decision Tree	Yes	Yes	Yes	Yes
Naïve Bayes Classifier	Yes	Yes	Yes	Yes
Time Series	Yes	Yes	No	Partial
Text Analytics	Yes	Yes	Yes	Yes
Big Data Processing	Yes	Yes	No	No
Visual Data Workflows	No	Yes	Yes	Yes

Table-3[12]. Comparative Table of Data Science Tools.

Table 4[13] displays the Framework Comparison of MapReduce, Big SQL, MPP, No SQL, In Memory with parameters Data Scale, Cluster Scale, Drivers, Data Type and Analytics Type.
Framework comparison also try to classify the data type as Structured, Semi-structured and Unstructured Data leading to Analytic type as Descriptive, Diagnostic, Predictive and Perspective Analytics

Frame work	Data Scale	Cluster Scale	Drivers	Data Type	Analytics Type
MapReduce: Hadoop	100's PB	10 to 1000's	Cost	Unstructured	Predictive
Big SQL	PB's	10 to 100's	Cost	Semi- Social Media	Diagnostic, Predictive
MPP: GPDB	PB's	10 to 100's	Performance	Structured	Descriptive, Diagnostic
No SQL	Trillions of rows	10 to 100's	Cost	Structured	Diagnostic, Predictive
In-Memory	Billions of rows	10 to 100's	Performance	Structured	Prespective

Table-4. Framework Comparison [13].

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

IV. CONCLUSION

The Qlik big data tools satisfy all the parameters like Visualization, Dashboard Analytics etc.. but proves to be risky for existing customers. Tableau tool ranks high on all other parameters except maintenance cost which is higher. However, Qlik Tool scores over Tableau and other big data tools.

The data science tools compared prove that Weka is undisputedly the best data mining and visualization tool followed by Rstudio.

Framework comparison proves that In-Memory provides higher performance but works for only Structured data. MapReduce proves that Data Scale is limited, data is unstructured and Predictive Analytics.

These studies and comparison prove that Qlik Tool for Big Data, Weka Tool & R Studio for Data Mining and Visualization, In-Memory and MapReduce to be used as framework provide much better advantages over traditional methods like effective Data Capture, processing and Analytics.

REFERENCES

[1] http://campus.afaqs.com/blog-details/11_Digital_analytics_ %E2%80%93_What_is_it?_(Part_2_of_Web_Analytics_S.

[2] http://r4stats.com/articles/popularity

[3] Sailesh Iyer and Lakhtaria Kamaljit .I. “Clustering Algorithm for Text Steganography”, International Journal of Advanced Research in Computer and communication Engineering” ISSN(O): 2278-1021, ISSN(P): 2319-5940, Vol. 5. Special Issue 3, November 2016.

[4] Lakhtaria Kamaljit .I., ed. Next Generation Wireless Network Security and Privacy, IGI Global 2015.

[5] https://www.slideshare.net/ScottMitchell14/tech-only-sf-bay-user-group-july-.

[6] https://mydataexperiments.com/2015/02/11/decision-matrix-for-big-data-tools-and-technologies/

[7] http://r4stats.com/2016/06/08/r-passes-sas-in-scholarly-use- finally.

[8] Jaseena K.U. and Julie M. David, “Issues, Challenges and Solutions:Big Data Mining Solution”, IJCSIT, Vol 5(2), 2014.

[9] Priya P. Sharma, Chandrakant P. Navdeti, “Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution”, IJCSIT, Vol 5(2), 2014.

[10] Albert Bifet, (2013), “Mining Big data in Real time”, Informatica 37, pp15-20.

[11] http://www.innoventsolutions.com/comparison-matrix.html.

[12] Puneet Singh Duggal, Sanchita Paul, (2013), “Big Data Analysis:Challenges and Solutions”, Int. Conf. on Cloud, Big Data and Trust, RGPV.

[13] http://www.predictiveanalyticstoday.com/bigdata-platforms-bigdata-analytics-software.

BIOGRAPHY

Dr. Sailesh Iyer is currently serving as a Senior Faculty in MCA Department, Kadi Sarva Vishwavidyalaya, Gandhinagar.He as Ph.D Degree in Computer Science and Research concentrated on developing and implementing an algorithm for Text Steganography. His research interests include Linguistic Steganography, Image Processing, Data Mining, Software Engineering, Project Optimization and Big Data Analytics. He is a Computer Society of India (CSI) Lifetime member and has to his credit various publications in International Journals of repute. He has also presented many Research Papers in International and National Conferences. He has served as a Judge for various events, delivered expert talks and organized several events including AICTE sponsored National Symposium.

Dr. Kamaljit I Lakhtaria is working as Associate Professor in Department of Computer Science, Gujarat University. He obtained Ph. D. in Computer Science in the area “Next Generation Networking Service Prototyping & Modeling”. He holds an edge in Next Generation Network, Web Services, Mobile Ad Hoc Networks, Network Security and Cryptography. He is author of 9 Reference Books in the area of Computer Science. He Published 3 chapters in International Editorial Volumes. He presented many Research Papers in National and International Conferences. His Papers are published in the proceedings of IEEE, Springer and Elsevier. His 5 Ph.D. students graduate under his guidance. He is Life time member ISTE, IAENG and many Research Groups. He hold the post of Editor, Associate Editor in many International Research Journal. He is Program Committee member of many International Conferences. He is reviewer in IEEE WSN, Inderscience and Elsevier Journals.