

Performance Metrics for Decision Support in Big Data vs. Traditional RDBMS Tools & Technologies

Alazar Baharu

Lecturer, Computer Science & IT Department,
AMiT, AMU
Ethiopia

Durga Prasad Sharma

Professor, CS&IT
AMiT, AMU
MOEFDRE

Abstract—In IT industry research communities and data scientists have observed that Big Data has challenged the legacy of solutions. ‘Big Data’ term used for any collection of data or data sets which is so large and complex and difficult to process and manage using traditional data processing applications and existing Relational Data Base Management Systems (RDBMSs). In Big Data; the most important challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization and privacy. As the data increases in various dimensions with various features like structured, semi structured and unstructured with high velocity, high volume and high variety; the RDBMSs face another fold of challenges to be studied and analyzed. Due to the aforesaid limitations of RDBMSs, data scientists and information managers forced to rethink about alternative solutions for handling such data with 3Vs. Initially research study focused on to develop an intelligent base for decision makers so that alternative solutions for long term suitable solutions and handle the data and information with 3Vs can be designed. In this research attempts has been made to analyze the feature based capabilities of RDBMSs and then performance experimentation, observation and analysis has been done with Big Data handling tools and technologies. The features considered for scientific observation and analysis were resource consumption, execution time, on demand scalability, maximum data size, structure of the data, data visualization, and ease of deployment, cost and security. Finally the research provides a decision support metrics for decision makers in selecting the appropriate tool or technology based on the nature of data to be handled in the target organizations.

Keywords—Big Data; RDBMSs; big data tools; Variety; velocity; volume; Metrics

I. INTRODUCTION

Currently big data analysis is an emerging domain of research and has become a new paradigm for business intelligence, predictions and forecasting in salient disciplines. In order to choose appropriate technology for data capture, curation and analysis; still there no clear decision support metrics to assist top level executives. A strong need is anticipated for development of a decision support system metrics for organizations who wants to handle different size data sets with different types and varied velocity and volume. There is a strong need for research oriented data handling mechanisms in order to support top level technocrats and executives for their decision making processes.

In the end of 1959 scientists have tried to trouble shoot these problems related to huge amount of data handling by

emerging hierarchical DBMS’s within organization having large amount of data with high computing power. This phenomenon was continued from 1960’s – 1970’s. After that a new era emerged i.e. was the era of EF Cod’s RDBMSs that overcome the limitations of DBMSs [1]. These solutions worked fine tuned till 2007 but after that there was again a limitation i.e. how to handle a huge amount of data which is un-structured or semi structured and has been increasing in Zettabyte per day with critical time complexity due to the introduction of different technologies such as e-commerce, smart city surveillance camera, GPS systems and social networking Medias. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s as of 2012, every day 2.5 Exabyte (2.5×10^{18}) of data were created; as of 2014, every day 2.3 Zettabyte (2.3×10^{21}) of data were created [2].

This dramatic increase in data with high variety, high velocity and high volume became difficult to handle by existing principles and mechanisms therefore a concept of Big Data has been evolved.

When the scope and importance of this research is narrowed down in Ethiopian context, it has been observed that business houses, millionaires, decision makers here believe on numbers rather than theoretical aspects and predictions. However the thorough analysis of literature review clearly indicate that there is no or very limited research studies conducted in the area especially for Ethiopian need and context, which may be a serious issue in future endeavors and key indicator for Ethiopian investors, business houses and industrialists.

This research study strive to conduct the performance analysis of Big Data vs. RDBMS tools and technologies to develop a crystal clear performance metrics that can support the decision makers to select the appropriate tool or technology from amongst the RDBMS and Big Data. Further, the parameters considered in this research are time complexity of search queries, memory management, data visualizations, scalability, deployment cost etc.

II. REVIEW OF LITERATURE

In the simplest way 58,300 results were found when searching for the term “difference between RDBMSs and Big Data” in Google. This can show as how much this topic is confusing and it needs to be clearly and scientifically explained [3]. In addition to this fact can also show as there is a gap of

concept and research works. This in turn confirms the research ability of the topic selected. Research related to Big Data emerged in the 1970s but has seen an explosion of publications since 2008. Big Data is an all-encompassing term for any collection of data sets so large or complex that it becomes difficult to process using traditional data processing applications like RDBMSs and Data warehousing tools and technologies. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. If this flood of Big Data challenges the legacy of the RDBMSs; a solution is needed for long term sustainability in order to gain the full potentials of hidden insights in Big Data. Data is exploding so fast and the promise of deeper insights is so compelling that IT managers are highly motivated to turn big data into an asset they can manage and exploit for their organizations. Emerging technologies such as the Hadoop framework and Map Reduce offer new and exciting ways to process and transform big data defied as complex, unstructured, or large amounts of data [4]. Why can't an analyst utilize databases with lots of disks to do large-scale batch analysis? Why is Map Reduce needed? The answer to these questions comes from another trend in disk drives: seek time is improving more slowly than transfer rate. Seeking is the process of moving the disk's head to a particular place on the disk to read or write data. It characterizes the latency of a disk operation, whereas the transfer rate corresponds to a disk's bandwidth. In many ways, Map Reduce can be seen as a complement to an RDBMS. Map Reduce is a good fit for problems that need to analyze the whole dataset, in a batch fashion, particularly for ad-hoc analysis. An RDBMS is good for point queries or updates, where the dataset has been indexed to deliver low-latency retrieval and update times of a relatively small amount of Data. Relational data is often normalized to retain its integrity, and remove redundancy. Normalization poses problems for Map Reduce, since it makes reading a record a nonlocal operation, and one of the central assumptions that Map Reduce makes is that it is possible to perform (high-speed) streaming reads and writes [5].

Applying Big Data analytics to the fuel of development faces several challenges. Some relate to the data including its acquisition and sharing, and the overarching concern over privacy. Others pertain to its analysis [6]. The most important affecting and challenging Factor in Big data is Privacy. It is the most sensitive issue, with conceptual, legal, and technological implications. However the three basic and important requirements for RDBMSs are confidentiality, integrity and availability. The stored data must be available when it is needed (availability), but only to authorized entities (confidentiality), and only modified by authorized entities (integrity). Traditional relational database management systems (RDBMS), like Oracle, SQL and MySQL, have been well-developed to meet the three requirements. In addition, enterprise RDBMS are further required to have ACID properties, Atomic, Consistency, Isolation, and Durability, that guarantee that database transactions are processed reliably. With such desirable properties, RDBMS have been widely used as the dominant data storage choice). RDBMS now are facing major performance problems in processing exponential growth of

unstructured data, such as documents, e-mail, multi-media or social media. Thus a new breed of non-relational, cloud-based distributed databases, called NoSQL, has emerged to satisfy the unprecedented needs for scalability, performance and storage)[7].

III. RESEARCH METHODOLOGY

This research study uses both qualitative and quantitative research methods. Quantitative technique used to collect and convert data into numerical form so that statistical calculations can be applied to conclusions.

Qualitative research methods applied to qualitatively analyze the research question and that of the quantitative research method helped to analyze based on numbers.

A. Data Collection Methods

The data collection methods used was formal interview with different professionals. Web articles, scholarly paper, white papers, flyers, company product specifications, and books related to the research topic were also analyzed critically as secondary sources of facts. In addition to this 60 IT professionals were selected randomly to gather facts about the commonly used Big Data and RDBMSs tools and technologies. The data set for experiment purpose is the *Olympic Games winners' dataset* which is acquired from the publically available data source *Talned*. Data analytical tools were selected based on the parameters like 1) Popularity statistics, 2) Market coverage statistics, 3) Commonly preferred Practices, 4) Professional recommendations statistics, 5) Professional recommendation collected through formal interview in the fact finding technique.

B. Organizing Data set

After collection of data sets; the data set has been organized in a manner to make it suitable for the experimentation purpose.

IV. PERFORMING EXPERIMENT ON THE DATASET

Experiments & analysis on data sets were done both quantitatively and qualitatively. The parameters selected for comparison metrics were CPU consumption, Memory (RAM) consumption, Execution Time, Scalability, Inter-Operability, Ease of deployment, System Security, data visualization, maximum data size and Cost.

A. Experimental procedure

Among the selected parameters for the performance analysis of Big Data and RDBMSs tools & technologies; CPU Time, Private working set, Execution time of each tools were measured and recorded by using the windows task manager and windows performance monitoring tools. In this performance analysis the two operations i.e. WRITE and READ operations were selected. Finally the performance of each tool has been measured by firing Reading and Writing Queries, then the result has been recoded to compare the performance of each tools.

1) Qualitative analyses

For qualitative comparison metrics performance analysis, System security, ease of deployment, Inter-operability and data visualization support parameters were analyzed. In doing this massive review of literature and fact finding techniques have been used.

2) Quantitative analysis

For quantities comparison metrics & performance analysis, the measurement of execution time, memory consumption, CPU consumption, cost, Max size of data, Scalability and data visualization (both quantitatively and qualitatively) were analyzed.

3) Comparative Analysis & Drawing Decision Support Metrics (System)

After extensive review of related research contributions, detailed experimentation, performance measurement and analysis; a comparative analysis was done to draw a decision support metrics for the future decision makers to select the most appropriate and most feasible RDBMS or Big Data tool or technology; based on their organizational needs and selection parameters.

V. SELECTION, EXPERIMENTATION AND DISCUSSION

A. Selection of Data analytical tools

The selection of the data analytics tools has been performed based on the two 1) Database engine ranking website and 2) From the formal interview conducted on selected IT professionals and data experts. According to the database engine ranking; website ranking of the database systems was done based on- 1) Number of mentions of the system on websites, 2) General interest in the system, 3) Frequency of technical discussions about the system, 4) Number of job offers, in which the system is mentioned and 5) Number of profiles in professional networks, in which the system is mentioned. Based on such criteria the top three RDBMSs found were following-

TABLE I. RDBMS RANKINGS ACCORDING TO THE DATABASE RANKING WEB SITE [TILL 1/8/2015]

Rank	Last Month	DBMS	Database Model	Score	Changes
1.	1.	Oracle	Relational DBMS	1439.16	-20.63
2.	2.	MySQL	Relational DBMS	1277.51	+8.93
3.	3.	Microsoft SQL Server	Relational DBMS	1198.61	-1.44

In addition to the above scientific observation and further verification; the IT professionals were interviewed. Based on the collected feedback from the interview it was confirmed that the most popular types of RDBMSs were the once ranked by the DB ranking engine and 69 percent of respondents responded that the introduction of Big Data Analytical technologies will bring additional features to data science.

On the collected opinion data and its analysis the Oracle, MYSQL and MS SQL Server were selected as top three RDBMSs from ten RDBMS. Expert & user opinion analysis

clearly indicates that domain specific people (experts and users) also like the RDBMSs in the same ranking manner i.e. first Oracle, second MySQL and third Microsoft SQL Server.

B. Selection of Big Data analytical tools

In computer and IT world there are several methods, techniques, tools and technologies which are used for database creation, storage, management and analysis of different types of data, information, text and documents to be analyzed. When the data is being generated with high volume, high variety and high velocity then alternative tools and technologies are available in the market for different kinds of analytics in real time manner. According to the Apache technology specifications; there are a number of different flavors and distributions of apache Hadoop that are available for Big Data analytics. Some of them include Amazon Web Services, Apache Bigtop, Cascading, Cloudera, Cloudspace, Datameer, Data Mine Lab, Datasalt, Hortonworks, HStreaming, IBM, MapR Technologies, Think Big Analytics, and WANdisco[8].

Among all these Hadoop flavors; Hortonworks data platform was selected as a most significant data analytics tool in this research study. The parameters used to select the Hortonworks were, ease of accessibility (i.e. open source), easy to deploy as it has user friendly and GUI interfaces, unstructured registration for accessibility, well established development community with sufficient deployment records in the real world and incorporation of free online and offline embedded tutorials. Rest of the big data analytical tools is very difficult to deploy them and to get them work on the machine specification prepared for the research study. So the selection of the data analytical tools from both parties looks like unbalanced but it was observed that each Big Data analytical tools uses the same technologies as a foundation like Apache Hadoop, Hive, PIG script, MAPR, Hcatlaog and so on[9].

C. Selection of supportive analytical tools

- Navicat premium

Among third party DBMS Connection and management tools; Navicat premium was selected to help in organizing the data. Navicat Premium is a database administration tool with 100,000 users across 7 continents in more than 138 countries and it allows to simultaneously connecting to MySQL, MariaDB, SQL Server, Oracle, PostgreSQL and SQLite databases from a single application [10].

- Oracle Virtual machine

There are a variety of virtualization tools available on the market among them Oracle virtual machine have been selected and used as a virtualization tool for hosting the Horton Works Sandbox [11].

- Windows task manger

Windows Task Manager has been used to display the programs, processes, and services that currently run on a computer. In addition to this Task Manager is used to monitor computer's performance. In monitoring resource usage of a given process, task manger used different metrics like CPU Usage, CPU Time, Memory - Working Set, Memory - Peak Working Set, Process time etc. This research used four task

manager columns like CPU Usage, CPU Time, Threads and Memory-Private Working Set.

- *Windows performance monitoring*

Windows Performance Monitor tool can be used to examine how programs running in a system are affecting the performance of the computer, both in real time and by collecting log data for later analysis. Windows Performance Monitor uses performance counters, event trace data, and configuration information, which can be combined into Data Collector Sets.

In this research study windows performance monitoring tools has been used to measure how much resources are used by a given process and to conform the results observed from the Task Manager [12].

D. Comparative and performance analysis

In this section, the RDBMS and Big data tools were compared to analyze the performance; based on selected. During analysis the measurement metrics; used are CPU USAGE, Memory (RAM) usage, execution time and number of threads of Big Data and RDBMSs tools and technologies.

1) A comparative analysis based on Structure of data

The structure of data which is generated by different sources was categorized as structured, semi structured and unstructured. In this study, it was assumed that if someone wants to analyze the data for getting further insights and knowledge discovery, it has to deal with these three structures of data. For that data analysis RDBMS and Big Data Analytical tools and technologies may play a greater role and contribution in the process of knowledge discovery and decision making. Till today, each tool has its own support related to the structure of the data and most of the RDBMS tools handle only structured data and they don't have any provision to analyze the semi-structured and unstructured data. However Big Data analytical tools and technologies supported all structures of data.

2) A comparative analysis on support of maximum data size

The data size limitation for a row and a column for each tool have been observed in the following table 2:

TABLE II. MAXIMUM DATA SIZE CAPABILITY

Tools Name	Max DB size	Max table size	Max row size	Max columns per row
MSSQL	524,272 TB	524,272 TB	8,060 bytes	30,000
MySQL	Unlimited	64 -256 TB	64 KB	4,096
Oracle	4 GB * block size/table size	4 GB * block size	8 KB	1,000
Horton works	Unlimited	Unlimited	Unlimited	Unlimited

Above table clearly indicates that BIG DATA technologies have created a pace for unlimited data representation and handling capabilities where the RDBMSs features have been comprises their limits in terms of data size.

3) A comparative analysis of CPU consumption, RAM consumption (PWS), Execution time and number of threads used

In any computation or communication, the quality of an algorithm or a given instruction are measured based on the amount of resources it consumes. Resources like Processor and Memory are the most important things to be measured when to evaluate the performance of a given instruction.

In this research study, a simple query prepared for data insertion and reading was used as instruction to be executed on both RDBMSs and Big Data tools and technologies to measure performance differences in the tools related to the consumption of CPU, RAM, Number of threads and execution time (elapsed) were measured by using data ranging from 100 up to 5,000,000rows in both read and write operations.

In measuring the performance of target tools& technologies; the query was fired using each tool. Afterwards, the windows task manager and windows performance monitoring tools were monitored and the effect of the fired query on resources like CPU and RAM (memory) were observed and recorded. In addition to this; the execution time and the no of threads used were also observed and recorded. The execution time for the query execution was recorded from the target RDBMS and BIG DATA tools and the number of threads used were recorded from the Performance monitoring tool data collector set log file.

The two queries executed in each tools are:

Query for inserting Writing data: INSERT [Olympic Athletes] ([Athlete], [Age], [Country], [Year], [Closing Ceremony Date], [Sport], [Gold Medals], [Silver Medals], [Bronze Medals], [Total Medals]) VALUES (N' Michael Phelps', 23, N' United States', N'2008', CAST(0x00009B0200000000 AS Date Time), N' Swimming', 8, 0, 0, 8)

*Query for reading (selecting) data: Select * from Olympic Athletes limit 100-5000000;*

Based on the above process the result of the performance evaluation and analysis was illustrated in the following Table 3. The Operations were divided in to two categories- READING and WRITING.

TABLE III. PERFORMANCE ANALYSIS OF EACH TOOL IN READ OPERATION

Name of the tool	No of rows	CPU time	PWS(K B)	Thread	Time(S)
MSSQL	100	2.9	144864	6	0.085
	1000	3.1	144884	7	0.106
	5000	6.5	144864	27	0.162
	10000	6.6	144890	17	0.228
	100000	6	144898	12	1.422
	1000000	12	144900	3	13.950
	5000000	48	144929	8	66.592
MySQL	100	3	230200	3	0.032
	1000	4.62	250300	3	0.004
	5000	6.15	260100	3	0.015
	10000	6.36	270000	3	0.029
	100000	12.72	360500	3	0.366
	1000000	23.08	390500	3	4.820
	5000000	30.74	430500	3	27.635
Oracle	100	5.37	400800	3	0.002
	1000	5.41	400900	3	0.009
	5000	7.73	410100	3	0.44
	10000	9.29	410700	3	0.082
	100000	23.3	650100	3	1.108
	1000000	37.51	304800	3	9.669
	5000000	42.57	136333	3	50.269
Horton works Hadoop	100	5.9	870900	3	66
	1000	11.3	876856	3	83
	5000	15.5	878698	3	89
	10000	19.9	879652	3	90.325
	100000	22.6	889453	3	110
	1000000	29.54	889657	3	128
	5000000	40.89	889759	3	170

Generally in read operation the performance of RDBMS increases even if the number of the data size increase up to max level, however Big Data tools and technologies performance somehow decreases as size of data increases.

The performance measurement of WRITE operation has been recorded and illustrated in table 4.as follows:

TABLE IV. PERFORMANCE MEASUREMENT IN WRITE

Name of the tool	No of rows to write	CPU Time (%)	PWS(KB)	Threads	Elapsed time(S)
MS SQL Server	100	0.6	262840	2	0.103
	1000	2	192492	3	0.893
	5000	2.2	237792	3	4.736
	10000	3.7	287916	3	9.843
MySQL Server	100	1.8	43668	2	2.115
	1000	12.7	43684	2	22.046
	5000	19.9	44698	2	137.599
	10000	22.2	44800	2	258.581
Oracle	100	1.7	162764	2	0.189
	1000	11.7	204556	2	1.184
	5000	15.9	311968	2	11.094
	10000	36.7	396964	2	16.995
Horton works data platform	100	0.5	125000	3	4
	1000	0.9	125162	3	9
	5000	3.6	125173	3	11
	10000	5.9	125189	3	14

CPU Time Consumption analysis on write operation

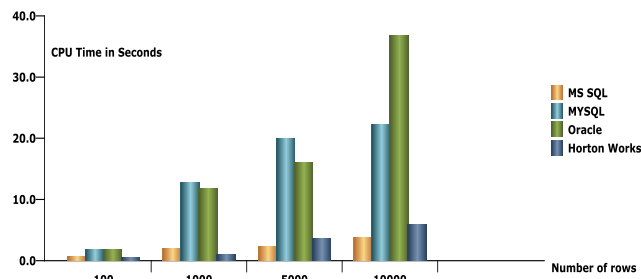


Fig. 1. CPU time consumption of each candidate tool in WRITE operation

RAM Consumption in Write operation

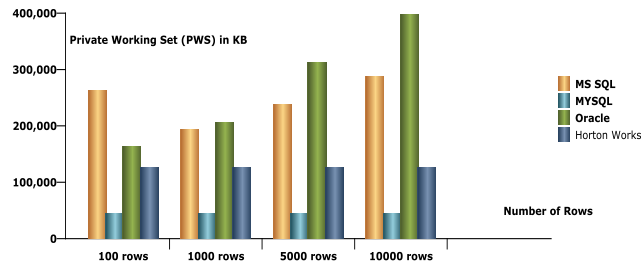


Fig. 2. RAM Consumption analysis of Each Candidate tool in WRITE operation

Execution time on Write operation

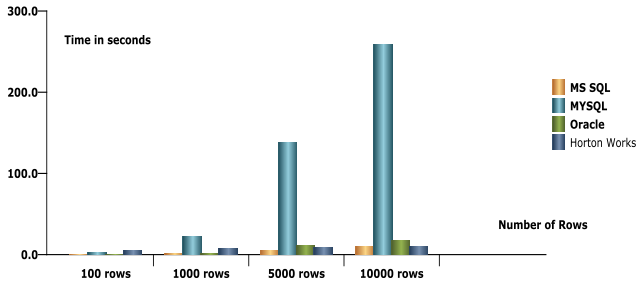


Fig. 3. Execution time analysis of each candidate tool in WRITE operation

In this phase it was observed that the performance of Big Data analytical tools and technologies increase in WRITE operation, however the performance of RDBMSs decrease when searching for the exact place or column table to put or to WRITE the data as the size of data increases.

E. A comparative analysis based on Software and operational cost

According to different scholar’s ideas software were classified under two broad. These categories are open source and proprietary software.

The purchase cost of the selected data analytics tools has been summarized on the following table 5:

TABLE V. COST ANALYSIS OF EACH TOOLS AND TECHNOLOGIES

Name of the tools	License type	Purchase coast
MYSQL	Open Source	Free
MSSQL	Proprietary	\$ 2,499.00
ORACLE	Proprietary	\$ 4,350.00
HORTON-Works data platform	Open Source	Free

F. A comparative analysis based on scalability

Scalability is one of the criteria to measure the capability of a given software or system. Scalability is all about the expansion support of a given system when there is a need for expansion in demand.

If the scalability of these tools is analyzed; most of the RDBMSs are not horizontally scalable however they are vertically scalable till there maximum limitations. As a matter of facts; scalability of Big Data tools is not limited horizontally as well as vertically.

G. Comparative analysis based on Data visualization support

Visualizations help people see things that were not obvious to them before. Even when data volumes are very large, patterns can be spotted quickly and easily [14]. Its fact that the data visualization will make data analysis results to be presented in a best possible way as crystal clear. Now every data analytics tools and technologies are including this feature in there system.

Using RDBMS one can do analytics on them but can’t visualize the result of the analysis, however using Big Data

tools and technologies every data analysis can be supported by the data visualizations.

H. Comparative analysis based on ease of deployment

Deployment is the first and sometimes the only experience system administrators have with an application. Ease of deployment is a key consideration for any systems. Most of the time installation of software on windows operating system is easy and simple it’s all about opening the executable file then following the prompt. However when it comes to installing software in non-windows operating system use needs a key knowledge working with the terminal and the command line scripting. Its fact that windows operating system have 55.42 % of users across the world and that of non-windows operating systems has 44.58 % of users [15].

Based on this fact and observation installing software like RDBMSs on windows machines can be an easy task even if it’s difficult for normal users to install them on non-windows operating system. However in Systems like Big Data tools, it has been observed that it’s very difficult to configure and deploy them on windows machines and also it’s a tough task to install them on non-windows operating systems without having a full knowledge terminal.

I. Comparative analysis based on System Security

In the digital age keeping information secured is very challenging task due to multiple threats imposed on the information systems keeping or storing the data for instance in 2013 Kaspersky lab have announced 5,188,740,554 cyber-attacks, identifies 104,427 newly modified malicious programs, 1,700,870,654 attacks on online resources from online attackers and 3 billion malware attacks were found [16].

In keeping the privacy and security of the information RDBMSs and Big Data analytical tools have their own solution or counter measures to every treats posed on them, however having and implementing all this tight security measures even didn’t kept the data from hackers. In this research study the security features of each tool from both RDBMS and Big Data analytical tools and technologies were analyzed based on the product owner specifications and it was founded that large volumes of RDBMSs have a high security measures deployed on them to keep the safety of the data. However if user see the security measures implemented in Big Data tools and technologies there is a high concern for security and privacy issue to be addressed in future.

Finally it was observed, analyzed and concluded that most of the RDBMSs have provided us a plenty of security features that can help to secure data and to protect it from authorized users, however most of the big data analytical tool challenges are confined to security and privacy issues only.

VI. DECISION SUPPORT METRICS FOR DATA DRIVEN ORGANIZATIONS

As stated in the objective, the research as a final contribution designed a suitable decision support metrics that can be used by Data analysts/ data scientists or decision makers in selecting data analytic technologies and tools. The decision support metrics is illustrated in table 6:

TABLE VI. SELECTION DECISION SUPPORT METRICS AND RECOMMENDATIONS

Selection parameters		RDBMS	Big Data Tools
Structure of data	Structured	Good	Good
	Semi-structured	Satisfactory	Good
	unstructured	Poor	Good
Resource consumption & execution time	Read operation	Good	Satisfactory
	Write operation	Satisfactory	Good
Scalability	Vertically	Good	Good
	Horizontally	Poor	Good
Cost		Satisfactory	Good
Ease of deployment		Good	Satisfactory
Data visualization		Satisfactory	Good
Large data set		Poor	Good
Security and privacy		Good	Satisfactory

VII. CONCLUSION

The main thrust of this research study started from the notion; if structured, unstructured and semi-structured data emerge with 3Vs then how to handle them efficiently. In case current technologies are not capable enough to handle; then what next? After the rigorous observation and analysis of different features of RDBMSs and Big Data, it is concluded that the major challenge in Big Data is storage capacity and it can be fulfilled by the Hadoop distributed file system (HDFS) and the analysis process can be handled by a Map and Reduce which can process data across different clusters in parallel manner. A metrics for purchase rent or deploy related decision support for data management or handling is also designed here. This 'Decision Support Metrics' can be used as an 'advisory base line' for selecting most fit tools from available in the market. This metrics presented seven parameters for the analysis. Based on these parameters; decision makers can select appropriate tools and technology for optimizing performance in desired domain of application. This base line metrics can be used by data scientists as a business intelligence support tool to select best fit tools from RDBMSs, DBMSs and Big Data technologies based on organizational needs.

REFERENCES

- [1] "Edgar F. Codd," 20 05 2014. [Online]. Available: http://en.wikipedia.org/wiki/Edgar_F._Codd.
- [2] "Big Data," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Big_data. [Accessed 23 05 2014].
- [3] "Differece between RDBMS and," Google, [Online]. Available: https://www.google.com.et/?gws_rd=ssl#q=difference+between+rdbms+and+big+data. [Accessed 15 07 2014].
- [4] Intel, "Peer Research Big Data Analytics," Intel, 2012.
- [5] N. F. H. M. K. Judith Hurwitz, Big Data for Dummies, New York: Willey, 2013.

- [6] Global Pulse, "Big Data for Development: Challenges and Opportunities," Global Pulse, 2012.
- [7] General Daynatics, "Current Data Security Issues of NoSQL Databases," General Daynatics, 2014.
- [8] "Hadoop Wiki," Apache, [Online]. Available: <http://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support>. [Accessed 11 07 2014].
- [9] "Hadoop," Hortonworks, [Online]. Available: <http://hortonworks.com/hadoop>. [Accessed 10 07 2014].
- [10] "Navicat Premum," Navicat Premum, [Online]. Available: <http://www.navicat.com/products/navicat-premium>. [Accessed 25 07 2014].
- [11] "Virtualization," Tech Target, [Online]. Available: <http://searchservervirtualization.techtarget.com/definition/virtualization>. [Accessed 26 07 2014].
- [12] "Using Windows Performance Monitoring Tool," TechNet, [Online]. Available: <https://technet.microsoft.com/en-us/library/cc749115.aspx>. [Accessed 29 07 2014].
- [13] "Big Data," SAS, [Online]. Available: <http://www.sas.com/big-data/>. [Accessed 10 06 2014].
- [14] "Data Visualization what it is and why it is important," SAS, [Online]. Available: http://www.sas.com/en_us/insights/big-data/data-visualization.html. [Accessed 9 08 2014].
- [15] "usage share of OS," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Usage_share_of_operating_systems. [Accessed 13 08 2014].
- [16] V. Beal, "security (computer security)," Webopedia, [Online]. Available: <http://www.webopedia.com/TERM/S/security.html>. [Accessed 23 8 2014].

AUTHORS PROFILE

Alazar Baharu: Alazar Baharu is academic staff in AMiT CS&IT Department at AMU. He is having additional charge of Library Director in order to his benevolent dedication and expertise for digitalization of library operations and services. He has developed many significant automation systems and services which are successfully deployed at his working organization like Demographic Heath surveillance system for which he is actively supporting the system as Data Manager. He has sound knowledge of Big Data analytics for decision support services and Information System Science.

Durga Prasad Sharma: Confluence of the three, an Eminent Academician, Technology Expert and Rehabilitation Activist, Prof. Durga Prasad Sharma has an excellent record of achievements and contributions in India and abroad. His contributions to the field of technology based rehabilitation under UN Convention for PWDs, ILO & UNDP schemes have been incomparable in the world. He is associated with Computer Science Teachers Association (CSTA-ACM) for USA & Canada and International Fellow of several organizations like SIE-Singapore and IACSIT-China & USA. He is recipient of 46 National and International Awards and wide range of appreciations. For his notable contributions to the research and academia, twenty nine world scientific organizations /societies have offered him various types of Membership and Fellowships. He has written 21 Computer Science and IT books and published more than 100 research papers/ article. He has guided 9 PhD scholars from USA, Fiji, Saudi Arabia and India.