Comprehensive Study of Big Data with Their Relative Analytics Tools and Technologies

Diksha Soni[1]*, Yogesh Kumar Gupta[1], Savita Dubey[2]

Abstract−In the period of Big Data war, primarily data is stored in the form of unstructured or semi structured form generated from the organizations/ industries and semantic web areas. However it is not easy to manage this flood of complex data, as big data is much approximating the Internet with its drawbacks along with its valuable and valid positive aspects, still there is need to have an optimistic discuss over analytics technologies such as Hadoop to handle the loch of Big Data. In evolutionary Big Data environment, the analytics of big data has been entered in a form of well-liked culture where MapReduce model is responsible for efficient analysis and for storing the approaching data; HDFS is used as storage layer. A consolidated discussion on analytics tools for big data handling, along with its process model, characteristics, challenges and issues are comprehensively mentioned in this paper. Some widely used statistical software packages are also demonstrated for imperially handling the data in real time that is probably found useful for the academics and scholars.

Keyword−Big Data, Hadoop, Hadoop Distributed File System, MapReduce, Statistical Software Package.

1 Introduction:

BIG Data is a gathering of large data sets containing array of data types, which may be in Kilobytes, Gigabytes, Megabytes, Terabytes, Petabytes, Exabytes and Zettabytes generated from various sources such as social media, stock market, medical science, satellite data etc with very high velocity [1]. The Big data referred to the enamours flock of structured, unstructured and semi structured data which may be analysed for getting meaningful information, pattern, trends and association especially related to human behaviour and interaction that intonates the business on a day to day basis.

From the beginning of evolution to 2003,only 5 Exabytes of information has created, currently by 2012 we generate that equivalent amount in just two days, if we consider data of digital universe that will cultivate to2.72 Zettabytes and by 2015 that will twice over every two years to reach 8ZB[17]. To store, access, manage, process and analyze the big data is more challenging for traditional techniques that involve large distributed file systems in commodity hardware for storing, which should be more flexible, fault tolerant, redundant, scalable and reliable. The tools and techniques used for big data analysis to manage massive amount of huge data are Hadoop, Map Reduce, NoSQL database, HPCC and Apache Hive [21].

To analyze voluminous amount of transaction data and to make extra well versed commerce decisions, big data analytics helps organizations by enabling analytical modellers, data scientists and other analytics professionals, as well as through analytics technologies it organizes others forms of structured and unstructured data that possibly will updated by conservative business intelligence (BI) programs, all these considered into primary goals of Analytical Big Data.

But big data is not the amount of data that's significant. It's what organisation does with the data only that matters [22]. It takes too much money, time and costs to load big data sets into a traditional relational database. So some of the new computational and analysing technologies based on massively parallel processing databases have emerged, which can concurrently distribute the processing of very large data sets of data across many servers like Hadoop.

2 Six Pillars of Big Data Analytics:

As we all know the 6V's characterize what Big Data is all about. Here is the concept of 6V's characterized the Big Data analytics are outlined as given in figure.

Volume: Volume related to Big Data refers to degree of vast data that continuously reproduce the amount of datasets in kilobytes, megabytes, terabytes, Petabytes and in Zetabytes that outplace conventional storage space and logical solutions.

Variety: Data generated from different sources are exceptionally heterogeneous [6]. The heterogeneity and variable nature of unstructured data can't be sort out by conventional data managing processes, as it comes from different sources in different formats and data types like text, images, audios, videos etc.

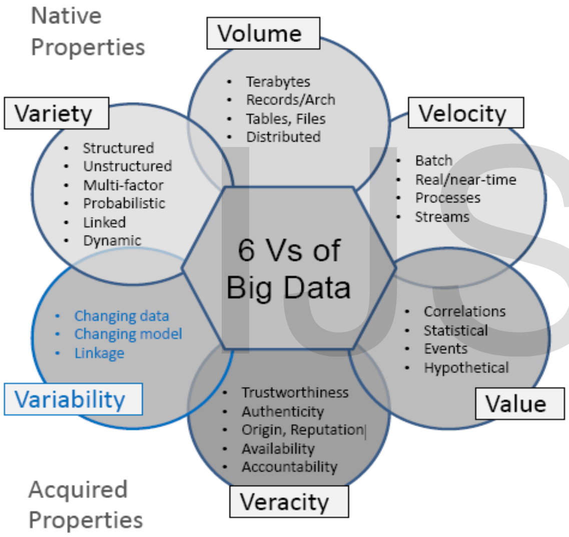

Fig 1_Demchenko: Concept of 6V's in Big Data [21]

Velocity: Velocity describes the rate at which large amount of data are generated in real time with demands and also deals with the pace at which it should be analyzed and act upon.

Veracity: The fourth V veracity which represents the confidence, accuracy or uncertainty in your data and how it might communicate to business significance. Comparatively veracity is very critical challenge for big data that need to be addressed using analytical tools and mining techniques [5].

Variability (and complexity): Variability deals with the term deviation in data flood rate. And the fact, big data are generated through an innumerable sources refers to the complexity [7]. Both are considered as the two additional dimensions of big data.

Value: In the terms of big data the value referred as a defining aspect such as correlations, statistical data, events, hypothetical data etc. [6].

3 Literature Review:

Since 2011 it has been considered that the interest area regarding big data has been exponentially increased [10]. For revealing the hidden patterns, the process of research into the complex data analytics basically concerned with statistical analytics tools [11].

Jifu G, LingLing Z. (2014). "Data, DIWK and Big Data and Data Science", mentioned about the data, Information, Knowledge and Wisdom concept for the knowledge discovery from the big data. This paper describes the relationship among the DIWK with respect to big data for retrieving the meaningful information because the valuable data evolving the knowledge is the only one that lie in the complex and big data. According to author it's not big to judge the valuable data, but sometimes it is more important to deal with small data also where Data Science plays their challenging role which might judge more innate value from big data. The term Data Science focus on vast storage of big data which is responsible for designate a new profession aims to solve big data problems [8].

Bhatia A, Vaswani G. (2013). "Big data: A Review", described that, to control manage and better analysis of data, there is potential for improving the scalability, portability and success of many organizations. Before this prospective potential can be fully realized, several technical challenges must be addressed, as described in this paper that evolves some common and obvious issues across a huge multiplicity of application and function domains, to deal with scalability, heterogeneity, timeliness, privacy, lake of structure and visualization [12]. They suggested the idea of adapting some transformative solutions for handle the big data challenges.

Chen M, Mao S, Liu Y. (2014). "Big Data: A Survey", in this paper researchers reviewed a comprehensive study of general background, status of art of Big data with some related technologies over Internet of things (TOI), Cloud Computing, Data Centers and Hadoop [13]. They also undertook the cover area on four of its main modules i.e. data collection, acquisition, analysis and integration, finally concluded with some technical challenges and applications.

Gandomi A, Haider M. (2014). "Beyond the Hype: Big Data concepts, methods and analytics", a consolidated picture of big data is given in this paper by integrating definitions from academics and practitioners. Usage of analytics methods and tool techniques for unstructured data are the primary focus of this paper to gain valuable and valid insights form complex data. To leverage huge amount of heterogeneous data in audio, video and text formats, the basic need of developing efficient and appropriate analytical methods are tinted in this paper by the researchers. For structured big data, the need to devise emerging tools and methods for predictive analysis is also reinforced by them.

Jha A, Dave M, Madan S. (2016). "A Review on the Study and Analysis of Big Data using Data Mining Techniques" mentioned about the data analytics using data mining techniques that provides the better aid in this area. This paper reviewed on different approaches for analysis, big data challenges, applications and also its importance in several fields using the mining techniques. They concluded with some text, video, images and audio analytics techniques that has scaled with advances in mainstream analysis, machine vision and speech recognition correspondingly, which are probable key of social, economical issues [16].

Hu H, Wen H, Wen, Y, Chua T, LI X. (2014). "Toward Scalable Systems for Big data Analytics" presented the system tutorial along with literature review on Big Data Analytics. The paper summarized the Big Data challenges and issues with respect to business analytics and a potential solution has been suggested that decompose the big data systematic framework into 4 chronological phases: data collection (rich big data attributes and sources are listed), acquisition (data gathering techniques are investigated followed by preprocessing and transmission), Storage (cloud based NoSQL) and analytics (analytics tools and methods) [6]. They finally concluded the paper by providing the explanation of Hadoop framework to which eventually solves the big data issues.

Kapdoskar R, Gaonkar S, Shelar N, Surve A, Gavhane S. (2015). "Big Data Analytics", followed up by having a main goal of good understanding of data mining process for big data analytics by implementing a web crawler. They proposed an algorithm for designing a web crawler for handling the big data. Data cleaning is suggested for removing the inconsistencies from the collected data by the web crawler that will be further migrated to database for integration, analysis and visualization.

Raghupathi W, Raghupathi V. (2012). "Big Data Analytics Architectures, Frameworks, and Tools" presented uniqueness of Big Data by considering the Volume/amount, Variety and Velocity and Veracity (and complexity) as the 4 basic pillars of big data. Big data analytics techniques also introduced by defining their scaling capability to handle the complex and sophisticated big data problem. Simultaneously framework architecture, methodologies and tools for analytics like hadoop/ Mapreduce/ HDFS are discussed which have also leveled up the appearance and granularity of big data demands. Several challenges and issues need to be highlighted that guaranteeing the safeguarding security and privacy which need to be handled.

4 Process Model of Big Data Architecture:

From the voluminous amount of data, for retrieving the significance information, trends, patterns and association especially related to human behavior, Big data required a centric model to achieve data gathering, storage, filtration, visualization and computation [4]. For big data analytics in real time some tools and techniques are available that follows up the process model. The computational model of big data analytics can be summarized as given by following components:

Data acquisition: Big Data refers to immense quantity of data. Thus, this becomes the problem when the information coming from the variety of sources. That whole information requires be gathering, storing and processing by the tools. Internet based navigation sites are perhaps most publicized collector of consumer data, such as yahoo, Amazon, Google or Bing [6]. For this purpose these organizations have greatly broader trade models. For example: Google offers Gmail free for each and every individual.

Data Filtration/Cleaning: data collected from the various sources are in structured, unstructured or in semi structured form which may have some inconsistencies, inaccuracy that can't be directly used for analysis. For cleaning up such data to the point some generalized cleaning tools, error detection tools, batch processing and data wrangling tools are used like open refine tool etc. that helps in identifying incomplete, irrelevant, incorrect, inaccurate data by replacing or modifying the rude data [4]. Maintaining the data as close as original, is the crucial task of these tools.

Data Analytics, Modeling and Prediction: Data analytics helps the organizations for cost reduction, to get faster and better decision making, for creating new products and services and for harness their data with analytical perspectives. Predictive modeling analytics is all about what is going to happen is future? This is done by analyzing historic data. It is procedure of generating; testing, validating and evaluating a model to finest foretell the probabilities of outcomes for analysis [6]. Predictive analytics includes preprocessing, data mining, modeling, deployment etc. by some available freeware software's like H2O; Apache spark MLlib, DMWay, and RapidMiner etc.

Data Visualization: Visualization of data meaning presentation of significant information (that is important than ever) in a graphical or pictographic form, which facilitates decision makers, to observe analytics presented visually, so that they can identify new patterns or grab difficult concepts that need attention or importance. Visualization of big data leads to assurance or certainty [12]. On Hadoop platform a number of visualization tools of big data can be executed. Some general modules in analytics tool Hadoop are: Hadoop MapReduce, Common utilities, HDFS- Hadoop Distributed File System and Hadoop YARN [24].

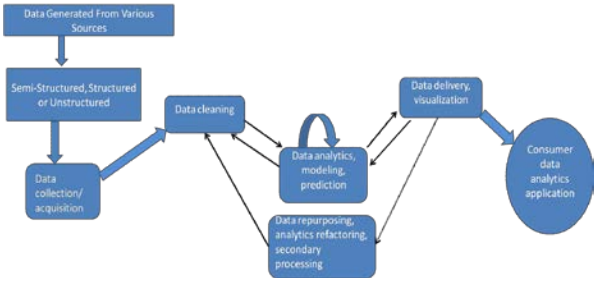
Data Generated From Various Sources
Semi-Structured, Structured or Unstructured
Data collection/ acquisition
Data cleaning
Data analytics, modeling, prediction
Data repurposing, analytics refactoring, secondary processing
Data delivery, visualization
Consumer data analytics application

Fig 2: Process Model of Big Data Lifecycle

5 Tools and Techniques for Big Data Analytics:

So as to deep considerate the concept of Big Data, this segment of paper describes about several elementary techniques including Hadoop, Map reduce etc. that are directly associated to big data.

With an era of evaluated technologies, the flow of tremendous amount of data is rapidly mounting, and having such cluster of data can no longer be possible to store and analyze with traditional analysis techniques. Therefore, to improve operational competence of piles of data, increase competitive advantages and to drive new profits over business rivals some new emerging techniques have been developed to make this unstructured data analytics likely possible [12]. By leveraging the authority of scattered computing resources, new approaches redefine the way of analyzing and the organization of unstructured data.

In real meaning we can summed up the analytics tools like that: tools that facilitate users to frequently and quickly analyze huge amount of data in real time and which provides a better framework for data mining techniques and infrastructures. Such analytic tools incorporate into various phases of decision making processes [17]. Data composed from the different sources residing at different locations are in different formats, has to be pooled together for analytics process via extraction, transformation and loading of data. After that Hadoop/Mapreduce platform takes that transformed data (which may be structured or unstructured) as an input, where decisions are taken regarding that processed data [19].

5.1 Hadoop:

Hadoop is java written, Apache open source distributed framework that stores extremely large amount of various type of data by unstable structures (or no structure at all) and using single programming model this framework allows running distributed processing applications on clusters of computers [11]. Hadoop solves the big data problem; it is a platform having enormous processing capability that can handle number of tasks and jobs. Until now the data that was difficult to analyze and manage, for this purpose Hadoop offers a tremendous deal of facility in enabling enterprises to tie together the big data.

5.1.1 Architecture of Hadoop Framework:

Hadoop has mainly two components of modules to work with data:

1) MapReduce as processing layer
2) HDFS (Distributed file system) as storage layer

Hadoop architecture is applied to process high amount of data in any structure. It is used for distributed storage and analysis in real time on a single machine. MapReduce programming layer is a parallel processing model on which Framework of Hadoop relays and HDFS is a Hadoop distributed file system used to store data [17]. YARN framework and common utilities are two additional aspects of the framework where YARN framework is used for resource management and job

scheduling. And apart from that Hadoop Common contains some utilities and java libraries helpful for other components of framework.

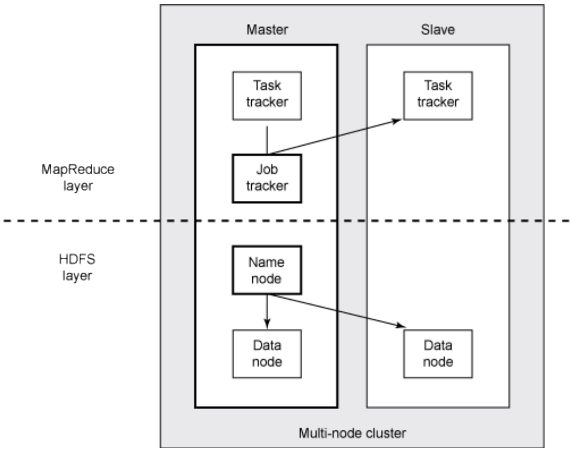

Fig 3_Fadnavis: Architecture Model of Hadoop Framework [11]

5.1.2 How Does Hadoop works?

Over a clusters of commodity computers hadoop runs the code to process. Hadoop performs the following core steps given as:

Firstly the data coming from various sources is divided into files and directories of uniformed block size of 128m of 64M (probably 128M). For further processing these chunks of files are then scattered across nodes or clusters. Here HDFS takes the responsibilities of processing, as it is local file system where blocks or chunks are replicated to avoid hardware failure. After then checking of code execution is done to perform sorting process between the MapReduce implementation. This sorted data then sent over certain computers to write the debugging logs for each and every job [14].

5.2. MapReduce Programming Model:

MapReduce is a software or programming structure or framework created by Google for distributed computing and processing of massive data. It is motivated by two terms “Map” and ”Reduce” (performed by specified functions: Mapper and Reducer) which are used for processing by adapting the concept of divide and conquer method [31]. The problems of big data are broken down into small units of work and processed in a parallel way. Parallel processing is necessary to achieve scalable solutions of big data problem.

MapReduce runs the job into two phases:-

By using Map and Reduce functions (which are written by users), a developer can write parallel distributed programs. In Map function key/values are taken as input pair, after performing some computation over it, intermediate results are produced which is merge-sorted and exchanged that is time consuming task depends on available resources [4]. This processed data is then computed by Reduce function for generating desired output. Elaborately, the coming data is processed as according to the illustrated figure in following 6 steps:

1) Input Reader: Files are taken or retrieved from the database as a input in a basic form of this stage and converted into tuples (key/value pairs). Finally processed by Map task.
2) Map function: Key/ value pairs generated from the above input reader stage, is taken

as input and logic is performed over it by the map function to get the new key/value pair as output. This output is in scattered forms that are merged into single file in the end.

3) Combine function: This is optional function provided for general cases like: 1) when the repetition of intermediate keys is occurred by each map job 2) when the reduce function written by user is associative and commutative.
4) Partition function: From the map task to reduce task to partition the intermediate keys output, the hash function is used (it is good for balancing but still another partition functions can be used here).
5) Reduce function: Here by the reduce function pair with the same key will be processed i.e. on set of values, reduce function is applied with the associated key. Possibly for each reduce job all the process is done in an increasing order of key value.
6) Output writer: To get the stable storage output is written by the output writer, in order to store the data in database.

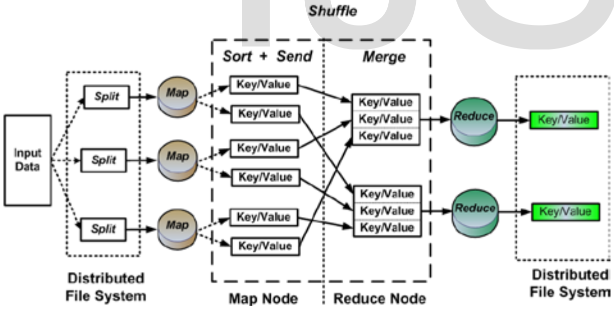

Fig 4: Process Flow Diagram of MapReduce [30]

5.3 HDFS (Hadoop Distributed File System):

HDFS is a portable java based dispersed file system that gives more reliability, scalability and which is highly fault tolerant than existing file system to store very large files in streaming access patterns redundantly. It is based on Google file system (GFS) and optimized or designed to run on commodity cluster hardware [17]. HDFS adopts the replication of data across the scattered nodes in order to achieve the fault tolerance i.e. HDFS is highly resilient. It follows the concept of master slave architecture. A single Hadoop cluster is a combination of one Name node (as a Master node) and numeral of Data nodes (as slave nodes). Name node is responsible to store the metadata like: file attributes, name, locations of each and every block address and replicas etc. whereas the actually data is stored in the Data nodes, i.e. divided large files which are in the form blocks or chunks, stored in Data nodes. It is considered that each chunk of data is replicated over other server node.

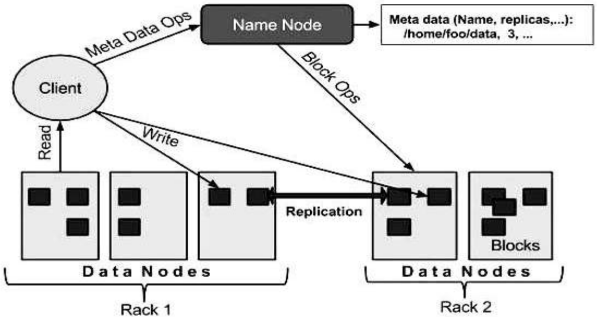

Fig 5: HDFS Architecture with Load Balancing [31]

6 Common Statistical Software Packages for Big Data Analytics:

6.1 SPSS:

Statistical Package for Social Sciences as its name suggest generally it is a window based software program used in social science and in business world, that is responsible for data entry, data editing, charting, data manipulation (by creating tables and graphs), survey and predictive analysis [23]. Here for the purpose of analysis data may be collected from any conceivable source: scientific research, a consumer database, or even the server log records of websites [23]. Through this software any sort of data file formats can be opened and displayed which are usually used for ordered data such as plain text files, Excel files and relational (SQL) databases. Some of its lofty end capabilities shift into the dominion of appliance culture to increase the proceeds, conduct research, and outperform competitors to make enhanced assessments. IBM is considered as one of the giant players in dome of big data that provides its considerable participation in analytics of big data. To effort with big data goods (such as cloudera, Hortonworks, Apache Hadoop and IBM

InfoSphere) IBM SPSS has been designed with some components or modules i.e. Data modeler [22]. The IBM SPSS Packages collection looks similar to this:
IBM Analytics Answers, IBM Social Media Analytics, IBM Analytical Decision Management, Customer analytics, IBM SPSS Statistics, IBM SPSS Modeler, Thread and fraud analytics and Operational analytics.

6.2 R:

R implements the statistical programming language commonly known as S language, in any software environment for the purpose of statistical data analysis. The S language comes into picture for implementing statistical actions or methods with restricted efforts and for well-organized statistical data analysis graphical representation, data handling and reporting [26]. It is a free and powerful software deals well with analytics tool such as Hadoop. Some special libraries are defined in R for data manipulation "Front ends" for R has been developed by some enterprising souls to make simple and elegant use for users who are unfriendly with command line interface [22]. So for this reason an unofficial menu based interface is recognized containing a gathering of fundamental statistical methods, best known as R Commander. R is maintained for three platforms: windows, Macintosh and as well as for Linux. The interpreted computer language is the core of R which allows looping and branching and modular programming also using some of its basic functions. R have number of facilities that facilitates a collection of operators for calculations on lists, vectors, lists and as well as for matrices. As one of its capabilities R has efficient storage facility, here one question arises regarding data handling that how much data can be handled by R? and the answer is R is exists in 64 bit version. And also an address space of 18,446,744,073,709,551,616 or greater than 18 quintillion bits is suggesting by R [22]. A commercial company, Revolution analysis supports for big data harvest in recital with IBM, which facilitates a coherent, incorporated and great sets of tools for data analysis.

6.3 SAS:

It is a platform independent extensive programming approach that provides a bright standpoint on business intelligence, transforming datasets into insight. SAS is considered as a comprehensive statistical tool for a broad range of statistical, evolving multivariate analysis, cluster analysis, regression analysis, nonparametric analysis, survey data and survival analysis [28]**.** In business analytics SAS comes out as a leader that extensively provides the data transformation and business analysis. It caters data management software and services in business intelligence through its pioneering analytics process. Power and sample size application provided by the Software is an interface for power and sample size computations [24]. As SAS is platform independent so it supports any operating systems like either Linux or Windows. Mainly two of its components are considered as most used components: data step and procedure step. In data step, data is read from the exterior source for manipulation and printing reports that is further combined with other type of data while in procedure step analysis is performed over data and finally the output is produced in the end [22]. With the help of SAS software one can perform several operations on data like IBM SPSS performs: Data Management, Business planning, Statistical Analysis, Operation research and project management, Data extraction/ transformation/ updating and modification and quality management [22]. Numbers of components comes under the SAS such as: Base SAS, SAS/ETS (Econometrics and Time series Analysis), SAS/GRAPH, SAS/OR (operation research), SAS/IML (Interactive matrix language) etc.

7 Challenges/ Opportunities:

Now a day we are waterlogged in an overflow of data. In big data analytics only the analysis phase is the one on which we focus, this analytics involves number of discrete stages or phases, each of which introduces some challenges. The problem with big data starts away during one of the distinct phases of analytics like data acquisition [18]. Currently it requires us to make decisions about the data in an ad hoc behavior (i.e. how to store meaningful data reliably along with metadata, what data need to be store and what is to discard etc.) Regarding this we have some common challenges and opportunities both.

Piles of data generated from different sources like sensors, mobile devices, blogs, tweets, mails, social network etc. which are basic reason of explosion of big data rapidly, however there is need to organize (to collect, store and drive value out of flood of data) the unscramble data having some obstacles or challenges[18].Timeliness,incompleteness/heterogeneity, scalability, privacy and security concerns and collaboration are some common challenges occurred during analysis of big data. Likewise big data provides some powerful opportunities to the organizations for taking the business profits to the higher level in market [15]. Particularly with the arrival of coming generation as the trends of technology advances, the amount and the number of available experimental data sets is rising exponentially and just because of this reason big data has the competence to modernize, not just research but also edification culture. In the period of big data, there are influential trends in creation the value of big data for intelligent transportation, computational social science and research, computational security, information technology, urban planning, energy saving and environmental modeling etc.

8 Conclusion:

This paper elaborated some fundamental concepts and broad overview of big data with its tools and technology along with some common statistical software packages such as SPSS, R and SAS. As this is the Information epoch so the valid and essential data need to be extracted, emphasized, analyzed and utilized as of the heterogeneous facts, significantly. In the direction of solving the Big Data problem the potential and prominent solution parameters are achieved consequently through better analysis of huge data in order to get advances in various methodical disciplines. Now it's the time to organize and planned up the Hadoop based big data lake by the available data analytics tools used in organizations for the better unforeseen insights and decision making.

Hence the paper was reviewed to deliver a study of analytics of big data to improve decision making capacity along with its phases: data acquisition, storage, cleaning/filtering, analysis processing and visualization respectively. Moreover some available software packages in the world of Big Data analytics particularly are also mentioned in our paper. This paper gives the comprehensive idea to the people of industries/ organizations and also to the developers an enhanced solution with various examples of big data tools that can be exploited and applied properly.

9 References:

[1] Elgendy N, Elragal A. Big Data Analytics: A Literature Review Paper. Springer International Publishing Switzerland, Research gate. 2016; doi: 10.1007/978-3-319-08976-8_16, PP.214-227.

[2] Agneeswaran V.S. Big Data – Theoretical, Engineering and Analytics Perspective. Proceeding of Lecture Notes in Computer Science (Springer). 2012; Vol 768: PP 8-15.

[3] Kapadoskar R, Gaonkar S, Shelar N. Big Data Analytics. International Journal of Advanced Research in Computer and Communication engineering (IJARCCE). 2015; Vol.4, Issue 10, October, 2278-1021.

[4] Gupta Y.k, Jha C.K. Study of big data with medical imaging communication. International Conference on Communication systems(ICCCs-16). CRC-press (Tylor &Fransis group). 2016; 6(1): 443-445.

[5] Bhoola K, Kruger K, Peick J. Big Data Analytics. Cape town international convention centre. 2014; 23-24 October.

[6] Hu H, Wen H, Wen Y, Chua T, LI X. Toward Scalable Systems for Big data Analytics. IEEE Access. 2014; Vol-2: 2169-3536.

[7] Gandomi A, Haider M. Beyond the Hype: Big Data concepts, methods and analytics. International Journal of information Management (ELSEVIER). 2014; Vol-35: 0268-4012.

[8] Jifu G, LingLing Z. Data, DIWK and Big Data and Data Science. International conference on Information Technology and Quantitative Management (ITQM). 2014; 814-821.

[9] Paokkonan P, Pakkala D. Reference Architecture and Classification of Technologies,

Products and Services for Big Data Systems. ELSEVIER. 2015; 2214-5796.

[10] Ward J.S, Basker A. Undefined by Data: A Survey of Big Data Definitions. University of Andrew, UK. 2013; 1309.5821v1, September.

[11] Fadnovis R.A, Tabhane S. Big Data Processing Using Hadoop. International Journal of Computer Science and Information Technologies. 2015; 6 (1): 0975-9646, 443-445.

[12] Bhatia A, Vaswani G. Big Data: A Review.International Journal of Engineering science and Research Technologies. 2103; ISSN: 2277-9655, August.

[13] Chen M, Mao S, Liu Y. Big Data: A Survey. Mobile Networking Appl.2014; doi 10.1007/511036-013-0489-0.

[14] Das T.K, Kumar P.M. Big data Analytics: A Framework for unstructured Data Analytics. International Journal of Engineering and Technology (IJET). 2013; Vol-5, ISSN: 0975-4024, March.

[15] Gupta Y.K, Jha C.K. A Review on the Study of Big Data with Comparison of Various Storage and Computing Tools and their Relative Capabilities. International Journal of Innovative in Engineering and Technology (IJIET). 2016; Vol. 7,ISSN: 2319-1058, Issue 1 June.

[16] Dave M. A Review on the study and Analysis of Big Data Using Data Mining Techniques. International Journal of Latest Trends in Engineering and Technology (IJLTET). 2016; Vol.6, ISSN: 2278-621X, Issue 3 January.

[17] vivekananth P, Baptist L.J.A. An Analysis of Big Data Analytics Techniques". In International Journal of Engineering and Management research (IJEMR). 2015; Vol.5, ISSN:2394-6962, Issue – 5, October.

[18] Kaur A. Big Data: A Review of Challenges, Tools and Technologies. International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET). 2016; vol.2,ISSN: 2395-1990, Issue 2.

[19] Raghupathi W, Raghupathi V. Big Data Analytics in Healthcare: Promising and Potential . Health Information Science and Systems. 2014; doi: 10.1186/2047-2501-2-3.

[20] Zakir J, Seymour T, Berg K. Big data Analytics. Issue In Information Systems. 2015; Vol.16, Issue 2, PP.81-90.

[21] Demchenko Y, Oprescu A, Ngo C, Grosso P, Laat C. Towards Defining Big Data Architecture Framework. University Van Amsterdam

[22] Pries, Kim H, Dunnigan, R. Big Data Analytics: A Practical Guide for Managers CRC-Press (Taylor & Francis Group): Baco Raton, 2015.

[23]Articletitle.-http://www.predictive_analyticstoday.com/top_statistical_software. Date accessed: 15/08/2016.

[24]Articletitle.-www.stat.ufl.edu/~aa/sta6127/appendix.pdf. Date accessed: 15/08/2016

[25]Articletitle.-www.uvm.edu/~dnowell/fundamentals/SPSSManual/SPSSLongerManual/SPSSchapter1.pdf. Date accessed: 21/8/2016

[26]Articletitle.-http://cran.rproject.org/doc/manuals/R-intro.pdf. Date accessed: 22/08/2016

[27]Articletitle.-http://Springer.com/article/10.1007/S10708-013-9516-8. Date accessed: 18/07/2016

[28]Articletitle.-http://linkspringer.com/chapter/10.1007/978-4614-4854-9_1#page-1. Date accessed: 22/08/2016

[29]Articletitle.-http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics. Date accessed: 19/07/2016.

[30]Articletitle.-www.stackoverflow.com/Questions/29769356/does-execution-of-map-and-reduce-phase-happen-inside-each-datanode-by-node-manag. Date accessed on 05-08-2016.

[31]Articletitle.-www.tutorialspoint.com/hadoop/hadoop_hdgs_overview.htm