International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3

ISSN: 2454-4248
223 – 229

Tools Used in Big Data Analytics

Hardeep Singh
Deptt. of Computer Science & Engg.
GNDU RC, Sathiala- 143205,
Amritsar, Punjab
hardeepsinghcse12@gmail.com

Satveer Kour
Deptt. of Computer Science & Engg.
GNDU RC, Sathiala- 143205,
Amritsar, Punjab
rattansatveer1985@gmail.com

Sandeep Kaur
Deptt. of Computer Science & Engg.
GNDU RC, Sathiala- 143205,
Amritsar, Punjab
sandeepkaurcse@gmail.com

Abstract: Big data is the current state of the art topic creating its unique place in the research and industry minds to look into depth of topic to get valuable results needed to meet the future data mining and analysis needs. Big data refers to enormous amounts of unstructured data created as a result of high performance applications ranging from scientific to social networks, from e-government to medical information system and so on. So, there also prevails the need of to analyze the data to get valuable data results from it. This paper deals with analytic emphasis on big data and what are the different tools used for big data analysis In this paper, different sections through an overlook on different aspects on big data such as big data analysis, big data storage techniques and tools used for big data analysis.

Keywords: *Big data analysis, Internet of Things, OLAP, Multidimensional, NoSQL.*

*****

I. INTRODUCTION

Big Data is the term came into existence when we talk about petabytes, terabytes of data flowing at high speed in various formats generated from different things connected to each other or in networks.Initially, the term big data appeared in 1998 by John Mashey with the title "Big Data and the Next Wave of Infra Stress" a book written by him. However, the first academic paper with title big data was appeared in 2000 by Diebold **[1]**.Big data growth can be seen through daily life examples or through the following facts: 72 hours of videos are uploaded to YouTube in every minute, the web pages indexed by Google were around one million in1998, which reached 1 billion in 2000 and have already exceeded 1 trillion in 2008.On an average Google have more than 1 billion queries per day. In 2003, it was recorded that around 5 million of data is created whereas the same amount of data is created every two days. In 2013, the same amount of data is created every 10 minutes. It is also estimated that 90%of the world data today has been created in the last two years which is growing at the rate of 40% every year **[2]**. According to the Gartner IT dictionary, big data can be defined as in terms of Three V's: Volume, Variety and Velocity. Where volume can be defined as the generation and collection of masses of data, it implies of large volume of data. Variety can be defined as different types of data which includes structured, unstructured, semi-structured which can include documents, emails, text messages, audio, graphs etc. Velocity can be stated as data arrival continuously as streams of data. Velocity is applied to data in motion. Another important leader in big data is IDC define big data as "big data technologies describe a new generation of technology and architectures, designed to economically extract value from very large volumes of wide variety of data by enabling the high velocity capture, discovery and analysis".**[7]**

Big Data Analytics represent the challenges of data that are vast, unstructured, and fast moving to be managed by traditional methods. From businesses and research institutions to governments, organizations now routinely generate data of unpredicted scope and complexity. To efficiently extract the meaningful data from such Big Data to improve their business performance and increase their market share. The tools available to data sources quickly and easily are challenging. Thus, analytics helps in realizing the full value used to handle the volume, velocity, and variety of big data which have improved greatly in recent years. In general, these technologies are not prohibitively expensive, and much of the software is open source. Hadoop, the most commonly used framework, combines commodity hardware with open source software. It takes incoming data and distributes them onto cheap disks; also providing tools for analyzing the data. However, these technologies do require a skill set that is new to most IT departments, which will need to work hard to integrate all the relevant internal and external sources of data. Although having attention to technology isn't sufficient, it is always a necessary component of a big data approach. This paper discusses some of the most commonly used big data technologies much of them are open source that work together as a big data analytics system. **[5]**

 ISSN: 2454-4248
223 – 229

II. BIG DATA STORAGE

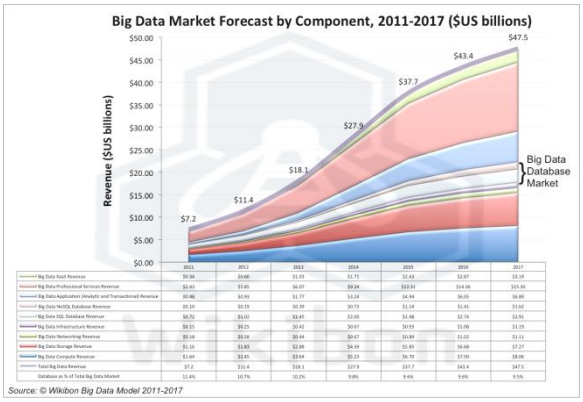

Figure1: Growth in database revenue **[8].**

The massive growth of data has more requirements on storage and management. Big data storage refers to the storage and management of large-scale datasets while achieving reliability and availability of data accessing. The storage infrastructure needs to provide information storage service with reliable storage space and it must provide a powerful access interface for query and analysis of a large amount of data. Traditionally, as secondary equipment of server, data storage device includes storing, managing, look up, and analyze data with structured RDBMSs. With the tremendous growth of data, data storage device is becoming more important, and many Internet companies seek big capacity of storage to be competitive. Therefore, there is a compelling need for research on data storage. **[6]**

The database technology has been evolving for more than 30 years. Various database systems are developed to handle datasets at different scales and support various applications. Traditional relational databases are unable to meet the challenges on categories and scales due to big data. NoSQL databases (i.e., non traditional relational databases) are becoming more popular for big data storage. NoSQL databases include flexible modes, support for simple and easy copy, simple API, eventual consistency, and support of large volume data. NoSQL databases are becoming the core technology for of big data. We will examine the following three main NoSQL databases in this section: Key-value databases, column-oriented databases, and document-oriented databases, each based on certain data models.

Key-value Databases: Key-value Databases are constituted by a simple data model and data is stored corresponding to key-values. Every key is unique and customers may input queried values according to the keys. Such databases feature a simple structure and the modern key-value databases are characterized with high expandability and shorter query response time than those of relational databases. Over the past few years, many key-value databases have appeared as inspired by Amazon's Dynamo system.

***Dynamo*:** Dynamo is a highly available and expandable distributed key-value data storage system which is used to store and manage the status of some core services realized with key access, in the Amazon e-Commerce Platform. The public mode of relational databases may generate incorrect data and can limit data scale and availability, while Dynamo help to resolve these problems with a simple key-object interface, which contains simple reading and writing operation. It helps in achieving elasticity and availability through the data partition, data copy, and object edition. Dynamo partition plan mainly relies on consistent hashing, providing advantage that node passing only affects directly adjacent nodes and do not affect other nodes, to divide the load for multiple main storage machines. Dynamo copies data to N sets of servers, in which N is a configurable parameter in order to achieve high availability and durability. Dynamo is also used in providing eventual consistency, so as to conduct asynchronous update on all copies. **[6]**

***Voldemort*:** Voldemort is also a key-value storage system, which was initially developed for and is still used by LinkedIn. Key words and values in Voldemort are composite objects maintained by tables and images. Voldemort interface includes three simple operations such as reading, writing, and deletion, all of which are confirmed by key words. Voldemort provides asynchronous updating concurrent control of multiple editions but does not ensure data consistency. However, Voldemort supports optimistic locking for consistent multi-record updating. When conflict arises between the updating and any other operations, the updating operation will quit. The data copy mechanism of Voldemort is the same as that of Dynamo. Voldemort not only stores data in RAM but allows that data to be inserted into a storage engine. Especially, Voldemort supports two storage engines including Berkeley DB and Random Access Files.

***Column-oriented Database*:** The column-oriented databases store and process data according to columns other than rows. Both columns and rows are segmented in multiple nodes to realize expandability. The column oriented databases are mainly inspired by Google's BigTable. In this Section, we first discuss BigTable and then introduce several derivative tools.

***BigTable*:** BigTable is a distributed, structured data storage system, which is designed to process the large-scale (PB class) data among thousands commercial servers. The basic data structure of BigTable is a multi-dimension sequenced mapping with sparse, distributed, and persistent storage. Indexes of mapping are row key, column key, and timestamps, and every value in mapping is an unanalyzed byte array. Each row key in BigTable is a 64KB character

 ISSN: 2454-4248
223 – 229

string. By lexicographical order, rows are stored and continually segmented into Tablets (i.e., units of distribution) for load balance. Hence reading a short stream of data can be highly effective, since it only involves communication with a small portion of machines. The columns are grouped according to the prefixes of keys, and thus forming column families. These column families are the basic units for access control. The timestamps are 64-bit integers used to differentiate various editions of cell values. Clients may flexibly determine the number of cell editions stored. These editions are sequenced in the descending order of timestamps, so the latest edition will always be read. The BigTable API helps in the creation and deletion of Tablets and column families along with modification of metadata of clusters, tables, and column families. Client applications may insert or delete values of BigTable, query values from columns, or browse sub-datasets in a table. BigTable too supports transaction processing in a single row. Users use such features to accomplish more complex data processing. Every procedure executed by BigTable includes three main components: Master server, Tablet server, and client library. BigTable only allows one set of Master server be distributed to be responsible for distributing tablets for Tablet server, detecting added or removed Tablet servers, and conducting load balance. In addition, it can also modify BigTable schema, e.g., creating tables and column families, and collecting garbage saved in GFS as well as deleted or disabled files, and using them in specific BigTable instances. Every tablet server manages a Tablet set and is responsible for the reading and writing of a loaded Tablet. When Tablets are too big, they will be segmented by the server. The application client library is used to communicate with BigTable instances. BigTable is based on many fundamental components of Google, such as GFS, cluster management system, SSTable file format, and Chubby. GFS is use to store data and log files. The cluster management system is responsible for task scheduling, resources sharing, processing of machine failures, and monitoring of machine statuses. SSTable file format is mainly needed to store BigTable data internally, and provides mapping among persistent, sequenced, and unchangeable keys and values as any byte strings. BigTable utilizes Chubby for the following tasks in server: 1) ensure there is at most one active Master copy at any time; 2) store the bootstrap location of BigTable data 3) look up Tablet server; 4) conduct error recovery in case of Table server failures; 5) store BigTable schema information; 6) store the access control table. **[6]**

Cassandra: Cassandra is a distributed storage system to manage the huge amount of structured data distributed among multiple commercial servers. The system was developed by Facebook and became an open source tool in 2008. Cassandra mimics the ideas and concepts of Amazon Dynamo and Google BigTable, especially integrating the distributed system technology of Dynamo with the BigTable data model. Tables in Cassandra are distributed four-dimensional structured mapping in which the four dimensions include row, column, column family, and super column. A row is distinguished by a string-key with arbitrary length. No matter the amount of columns to be read or written, the operation on rows is an auto. Columns may constitute clusters, which is called column families, and are similar to the data model of BigTable. Cassandra provides two kinds of column families: column families and super columns. The super column includes arbitrary number of columns related to same names. A column family includes columns and super columns, which may be continuously inserted to the column family during runtime.

Derivative tools of BigTable: As we know that BigTable code cannot be obtained by open source license but some of the open source projects compete to implement the BigTable concept to develop similar systems, such as HBase and Hypertable. HBase is a BigTable cloned version programmed with Java and is a part of Hadoop of Apache's Map Reduce framework. HBase replaces GFS with HDFS. It is used to write updated contents into RAM and also regularly writing them into files on disks. The row operations are atomic operations containing row-level locking and transaction processing, which is optional for large scale. Partition and distribution are transparently operated and have space for client hash or fixed key. HyperTable was developed to obtain a set of high-performance, expandable, distributed storage and processing systems for structured and unstructured data. HyperTable relies on distributed file systems, e.g. HDFS and distributed lock manager. Data representation, processing, and partition mechanism are similar to BigTable. HyperTable contain its own query language, which is called HyperTable query language (HQL), allowing users to create, modify, and query underlying tables. Since the column-oriented storage databases mainly mimic BigTable and their designs are all similar, except concurrency mechanism and several other features. For example, Cassandra mainly focuses on weak consistency of concurrent control of multiple editions while HBase and HyperTable emphasis on strong consistency through locks or log records. **[5]**

***Document Database*:** As Compared to key-value storage, document storage help to support more complex data forms. Since documents do not follow strict modes, there is no need to conduct mode migration. In addition, key-value pairs can still be saved. We will examine three important representatives of document storage systems, i.e., MongoDB, SimpleDB, and CouchDB.

 ISSN: 2454-4248
223 – 229

MongoDB: MongoDB is open-source and document-oriented database. MongoDB stores documents as Binary JSON (BSON) objects, which is similar to object. Every document has an ID field as the primary key. Query in MongoDB is expressed with syntax similar to JSON. A database driver sends the query as a BSON object to MongoDB. The system allows query on all documents, including embedded objects and arrays. To enable rapid query, indexes can be created in the query able fields of documents. The copy operation in MongoDB can be executed with log files in the main nodes that support all the high-level operations conducted in the database. During copying, the slavers query all the writing operations since the last synchronization to the master and execute operations in log files in local databases. MongoDB supports horizontal expansion with automatic sharing to distribute data among thousands of nodes by automatically balancing load and failover. **[6]**

SimpleDB: SimpleDB is a distributed database and is a web service of Amazon. Data in SimpleDB is organized into various domains in which data may be stored, acquired, and queried. Domains include different properties and name/value pair sets of projects. Date is copied to different machines at different data centers in order to ensure data safety and improve performance. This system does not support automatic partition and thus could not be expanded with the change of data volume. SimpleDB allows users to query with SQL. It is worth noting that SimpleDB can assure eventual consistency but does not support to Muti-Version Concurrency Control (MVCC). Therefore, conflicts therein could not be detected from the client side.

CouchDB: Apache CouchDB is a document oriented database written in Erlang. Data in CouchDB is organized into documents consisting of fields named by keys/names and values, which are stored and accessed as JSON objects. Every document is provided with a unique identifier. CouchDB allows access to database documents through the Restful HTTP API. If a document needs to be modified, the client must download the entire document to modify it, and then send it back to the database. Once document is rewritten, the identifier will be updated. CouchDB utilizes the optimal copying to obtain scalability without a sharing mechanism. Since various CouchDB may be executed along with other transactions simultaneously, any kinds of Replication Topology can be built. The consistency of CouchDB relies on the copying mechanism. CouchDB supports MVCC with historical Hash records. Big data are stored in hundreds and thousands of commercial servers. Thus, the traditional parallel models, such as Message Passing Interface (MPI) and Open Multi-Processing (OpenMP), may not be adequate to support such large-scale parallel programs. Recently, few proposed parallel programming models help to improve the performance of NoSQL and effectively reduce the performance gap to relational database.

***Dryad*:** Dryad is a general-purpose distributed execution engine for processing parallel applications of coarse-grained data. The operational structure of Dryad is a directed acyclic graph, in which vertexes represent programs and edges represent data channels. Dryad executes operations on the vertexes in clusters and transmits data via data channels, including documents, TCP connections, and shared-memory FIFO. During operation, resources in a logic operation graph are automatically mapped to physical resources. The operation structure of Dryad is coordinated by a central program called job manager, which can be executed in clusters or workstations through network. A job manager consists of two parts: 1) application codes which are used to build a job communication graph, and 2) program library codes that are used to arrange available resources. All of the data kinds are directly transmitted from one vertex to another. Therefore, the job manager is only responsible for decision-making, which does not obstruct any data transmission. In Dryad, application developers can flexibly choose any directed acyclic graph to describe the communication modes of the application and express data transmission mechanisms. Dryad also allows vertexes to use any amount of input and output data. DryadLINQ is the advanced language of Dryad and is used to integrate the aforementioned SQL-like language execution environment.

All-Pairs: All-Pairs is a database technology mainly designed for biometrics, bio-informatics, and data mining applications. It focuses on comparing element pairs in two datasets by a given function. All-Pairs can be expressed as three-tuples (Set A, Set B, and Function F), in which Function F is utilized to compare all elements in Set A and Set B. The comparison result is an output matrix**M**, which is also called the Cartesian product or cross join of Set A and Set B. All-Pairs is implemented in four phases: system modeling, distribution of input data, batch job management, and result collection. In Phase I, an approximation model of system performance will be built to evaluate how much CPU resource is needed and how to conduct job partition. In Phase II, a spanning tree is built for data transmissions, which makes the workload of every partition retrieve input data effectively. In Phase III, after delivering of flowing data to proper nodes, the batch-processing submission is build by All-Pairs for jobs in partitions, along sequencing them in the batch processing system, and also formulating a node running command to acquire data. In the last phase, collection of results is done by extraction engine which combines them in a proper structure, which is generally a single file list, in which all results are put in order.

 ISSN: 2454-4248
223 – 229

***Pregel*:** The Pregel system of Google facilitates the processing of large-sized graphs, e.g., analysis of network graphs and social networking services. A computational task is expressed by a directed graph constituted by vertexes and directed edges. Every vertex is related to a modifiable and user-defined value, and every directed edge related to a source vertex is constituted by the user-defined value and the identifier of a target vertex. When the graph is built, the program conducts iterative calculations, which is called superstep among which global synchronization points are set until algorithm completion and output completion. In every superstep, vertex computations are parallel, and every vertex executes the same user-defined function to express a given algorithm logic. Every vertex may modify its and its output edges status, receive a message sent from the previous superstep, send the message to other vertexes, and even modify the topological structure of the entire graph. Edges are not provided with corresponding computations. suspension is used to remove the function of every vertex. When all vertexes are in an inactive status without any message to transmit, the entire program execution is completed. The Pregel program output is a set consisting of the values output from all the vertexes. The input and output of Pregel program are isomorphic directed graphs.

III. TECHNOLOGIES OF BIG DATA

Figure2: Tools used in big data analysis**[9].**

parallelism. The reduce tasks use the output of the maps to obtain final results of the job. It basically follows the divide and aggregate rules.

Apache S4: It is a platform developed for processing continuous data streams.S4 is designed specifically for managing data streams.S4 applications are made by combining streams and processing elements in real time.

Apache Mahout: It is a scalable machine learning and data mining open source application. It has implantation of a wide range of machine learning and data mining algorithms. It is basically based on a hadoop paradigm.

R [5]: It is an open source programming language and software made for statistical computing and visualization. It was designed by Ross Ihaka and Robert Gentleman. It is basically meant for statistical analysis of huge data.

As we know, Internet of things (IOT) enables big data and we can see their implementation in various functions and operations. Both IOT and big data extend their capabilities to wide range of areas or we can say that the sources of big data are correlated to each other in one or another form. The data comes from different sources like sensors, social media, mobile phones, GPS signals, industrial automation, agriculture, digital pictures, audios, videos, intelligent transportation, intelligent building construction, climate forecasting, real time viewing of programs and a lot more examples can be seen in our day to day life through which we producing large amount of big data through different connected devices to each other over the network **[3]**.This progress and innovation is gradually getting hindered by the problem of collecting it, managing it, processing it and analyzing it in a scalable fashion. But to overcome the problem of handling big data, various technologies are developed and some others are in developing phase to get best of best out of it. Big data architecture deals with hadoop and other software as follow to meet big data analyzing needs as follows:

Apache Hadoop [5]:Hadoop is a processing engine or a framework that is designed for distributed processing of large datasets across large clusters of computers. It is an open source implementation for Google Map reduces.

Hadoop basically has two components **[4]**:

1. Hadoop distributed file system (HDFS) which can support data in structural, unstructured or in any form in between. It is a distributed system that is reliable in providing high access to data.
2. The Map Reduce programming algorithm divides the input dataset into independent subsets that are being processed by map tasks in

MOA: It is an open source programming tool designed for performing data mining in real time. It is implemented for classification, regression, clustering and frequent item set and graph mining. The framework provides an environment for defining and running stream processes using simple XML based definitions whereas the SAMOA is the new upcoming software project for distributed stream mining that will combine S4 and storm with MOA.

Pegasus: It is a big graph mining tool that is built on the top of Map Reduce. It allows finding the patterns and anomalies in massive real world graphs.

Graph lab: It is a high level graph parallel system built without using the Map Reduce. It computes over dependent records over which are stored as vertices in a large distributed data-graph. Computations and Algorithms in graph lab are expressed as vertex–program which are

 ISSN: 2454-4248
223 – 229

executed in parallel on each vertex and can interact with adjacent vertex. Many tools for big data mining and analysis are available, which comprises of professional and amateur software, expensive commercial software, and open source software. In this section, we briefly review the top five most widely used software, according to a survey of "What Analytics, Data mining, Big Data software that you used in the past 12 months for a real project?" of 798 professionals made by KDNuggets in 2012.

***R* (30.7 %)**: *R*, an open source programming language and software environment, is designed for data mining/analysis and visualization. While computing intensive tasks are executed, code programmed with C, C++ and FORTRAN may be called in the R environment. In addition, skilled users can directly call *R* objects in C. Actually, *R* is a realization of the *S* language, which is an interpreted language developed by AT&T Bell Labs and used for data exploration, statistical analysis, and drawing plots. Compared to *S*, *R* is more popular since it is open source. as it ranks top in the KDNuggets 2012 survey. Furthermore, in a survey of "Design languages you have used for data mining/analysis in the past year" in 2012, *R* was also in the first place, defeating SQL and Java. Due to the popularity of *R*, database manufacturers, such as Teradata and Oracle, have released products supporting *R*.

***Excel* (29.8 %)**: Excel, a core component of Microsoft Office, provides powerful data processing and statistical analysis capabilities. When Excel is installed, some advanced plug-ins, such as Analysis ToolPak and Solver Add-in, with powerful functions for data analysis are integrated initially, but such plug-ins can be used only if conversion, filtering, statistics, mining, and finally data visualization. The entire development process is conducted under a visualized environment. KNIME is a module-based and expandable framework. There is no dependence between its processing units and data containers, making it adaptive to the distributed environment and independent development.

***Weka/Pentaho* (14.8 %):** Weka, abbreviated from Waikato Environment for Knowledge Analysis, is a free and open-source machine learning and data mining software written in Java. Weka provides such functions as data processing, feature selection, classification, regression, clustering, association rule, and visualization, etc. Pentaho is one of the most popular open-source BI software. It includes a web server platform and several tools to support reporting, analysis, charting, data integration, and data mining, etc., all aspects of BI.Weka's data processing algorithms are also integrated in Pentaho and can be directly called.

users enable them. Excel is also the only commercial software among the top five. **[6]**

***Rapid-I Rapid miner* (26.7 %)**: Rapid miner is an open source software used for data mining, machine learning, and predictive analysis. In an investigation of KDNuggets in 2011, it was more frequently used than R. Rapid Miner include Extract, Transform and Load (ETL), data pre-processing and visualization, modeling, evaluation, and deployment in their data mining and machine learning program . The data mining flow is described in XML and displayed through a graphic user interface (GUI). Rapid-Miner is written in Java. It integrates the learner and evaluation method of Weka, and works with R. Functions of Rapid miner are implemented with connection of processes including various operators. The entire flow can be deemed as a production line of a factory, with original data input and model results output. The operators can be considered as some specific functions with different input and output characteristics.

***KNMINE* (21.8%):** KNIME (Konstanz Information Miner) is a user-friendly, intelligent, and open-source rich data integration, data processing, data analysis, and data mining platform. It helps users to create data flows or data channels in a visualized manner, and then selectively run some or all analytical procedures. It also provides analytical results, models, and interactive views. KNIME was written in Java provides more functions as plug-in. Users can insert processing modules for files, pictures, and time series for integrating them into various open source projects through plug-in files e.g., R and Weka. KNIME controls data integration, cleansing,

IV. CONCLUSION

This paper is basically reviewed about the different aspects of big data analysis. What is big data analysis and how it can be done. The big data analytic tools including big data storage technologies used traditionally and currently available. Yet it is not over, there are still new future aspects and challenges of big data are discovered day by day to improve analysis and data mining to get best fit tools to handle it and get maximum results out of it.

V. OBJECTIVES

The main objectives of this paper are

- The research on Big Data technology is very mindful and brilliant.
- This makes us aware and enlightened our knowledge.
- Moreover, we researched about theory and this theory will definitely help in our practical implementation of Big Data Analytics.

 ISSN: 2454-4248
223 – 229

- Theory is important in practical purposes.

REFERENCES

[1] Fan,W & Bifet,A . Mining big data: Current status and forecast to the future. SIGKDD Explorations, 14(2), 1-5.
[2] Snijders,C, Matzat,U & Reips,U. Big Data: Big Gaps of Knowledge in The Field of Internet. International Journal of Internet Science,7, 1-5.
[3] Cuzzocrea,A, Song,Y & Davis,K.(2011). Analytics of Large Scale Multidimensional Data: The Big Data Revolution!. DOLAP.11.
[4] Aggarwal,C.(2013).Managing and Mining Sensor Data. International Journal of Advances In Database System, Springer.
[5] Bifet,A, Holmes,G, Kirkkby,R, & Pfahringer,B (2010).MOA: Massive Online Analysis. International Journal of Machine Learning Research.
[6] Chen.M, Mao.S, &Liu.Y(2014). Big Data: A Survey. International Journal of science and business media,Springer.
[7] Russom.P (2011). Big data Analytics.TDWI Best practices Report,TDWI Research
[8] wikibon.org.(n.d.).Google, Growth in database revenue.[Infographic].Retrieved from http://wikibon.org/wiki/v/Big_Data_Database_Revenue_and_Market_Forecast_2012-2017
[9] datanao (n.d Tools used in big data analysis [Infographic].Retrieved from http://www.datanao.com/mdm-big-data-tools