BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)
ISSN (P): 2348-0513, ISSN (E): 2454-471X
Vol. 5, Issue 02, Feb 2017, 1-6

TOOLS USED IN BIG DATA

AVNEET KAUR

Department of Computer Science & Electronics Engineering, G.N.D.U Amritsar, India

ABSTRACT

The term “Big Data” was coined to address the massive volume of data storage and processing. It is increasingly becoming imperative for organizations to mine this data to stay competitive. Big Data means to manage large scale information and analyze the technologies that go beyond the capability of traditional data processing technologies. Big Data is being differentiated from traditional technologies in three ways:

- The volume of data,
- The rate of data generation and transmission (velocity), and
- The types of structured and unstructured data (variety).

KEYWORDS: Big Data, Analytics of Big Data, Tools

INTRODUCTION TO BIG DATA

Today we are living in the digital world. With the increased conversion of analog information to digital information the amount of structured data & unstructured data which is being created and stored is exploding. The data is being generated from various sources such as transactions, social media, sensors, digital images, videos, audios click & streams for domains including healthcare, retail, energy and utilities. The term “Big Data” [1] was coined to address the bulk volume of data storage and processing. It is increasingly becoming need of hour for organizations to mine this big data to stay competitive. Analyzing of this data can provide significant competitive edge for an enterprise. The data when analyzed properly leads to a wealth of information which will help the businesses to reformulate the strategies.

Big Data refers to data that is too big to be adjusted on a single server or too unstructured to accommodate it into a row and column database. *"Big Data [2] also refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies."* Big Data is being differentiated from traditional technologies in the three ways that are: the amount of data (volume), the rate or speed of data generation and transmission (velocity), and the various types of structured and unstructured data (variety).

Characteristics

The following are the characteristics of Big data:

Volume – The quantity of data that is generated is very important in this domain. It is the size of the data which determines the value and potential of the data that is under consideration and whether it can actually be considered as Big Data or not.

Variety - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is

also a very important or essential fact that needs to be known by the data analysts.

Velocity - The term 'velocity' in the context refers to the speed of data generation or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

Variability - This is a factor which can be a problem for those who analyse the data. This refers to the inconsistency in the data at times, thus hindering the process of being able to handle and manage the data effectively.

Veracity - The quality of the data being captured can vary a lot. Accuracy of analysis depends on the veracity of the source data.

Complexity - Data management can be a very complex process, especially when large volumes of data is generated from multiple sources. These data need to be linked, connected and correlated in order to easily able to grasp the information that is supposed to be conveyed by these data.

ANALYTICS OF BIG DATA

Big data analytics [3] is defined as advanced analytic techniques against very large, diverse data sets that include different types of data such as structured/unstructured and streaming/batch, and different data sizes from terabytes to zettabytes. Using various advanced analytics techniques such as text analytics, machine learning [4], predictive analytics, data mining [5], statistics, and natural language processing [6], businesses can analyze previously unused data sources independent or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions. Large data sets which are considered as information overload are difficult for humans to analyze manually. Big data tools are used because these tools have the ability to run ad-hoc queries against the large data sets in less time with an average or good performance. Analysis of Big data enables the executives to get the relevant data in less time for making faster decisions. Big data can pave way for fraudulent analysis, customer segmentation based on the store behavior analysis, loyalty programs that identifies and targets the customers.

Various Techniques Used in Big Data Analysis Are

- **Machine learning:** A special area of computer science (with a field historically called "artificial intelligence") which is concerned with the design and development of algorithms that allow computers to acquire or develop behaviours based on empirical data. A major focus of machine learning research is to automatically able to recognize complex patterns and make refined or sophisticated design based on data. Natural language processing is an example of machine learning.
- **Natural language processing (NLP):** It is a set of techniques from a subspecialty of computer science. Some of the NLP techniques are categorised under machine learning. One of the applications of NLP is using sentiment analysis technique on social media to determine how related customers are reacting to a branding campaign.
- **Predictive modelling:** It is a set of techniques in which a mathematical model is chosen or created to b predict the best probability of an outcome. Regression is also one of the examples of the many predictive modelling techniques.
- **Data mining:** A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis,

classification and regression. Applications include mining customer data to determine segments more likely to respond to an offer, mining human resources data to identify characteristics of most successful employees, or market basket analysis to model the purchase behaviour of customers.

- **Regression:** A set of statistical techniques to determine how the value of the dependent variables changes when one or more independent variables are modified. This technique is often used for forecasting or prediction. Examples of applications include forecasting sales volumes base on various market and economic variables or determining what measurable manufacturing parameters most influence customer satisfaction. Use for data mining.

VARIOUS TOOLS USED IN BIG DATA UNDER DIFFERENT AREAS ARE

Database/Data Warehouse

Cassandra

It was developed by Facebook. It is NoSQL database and is managed by the Apache Foundation. It is used by many organizations like Netflix, Twitter, Cisco and Digg with large, active datasets. Third-party [7] vendors provide services and commercial support.

Operating System: OS Independent.

MongoDB

It was designed to support humongous databases. It is also a NoSQL database with various features like document-oriented storage, full index support, and high availability, etc.

Operating system: Windows, Linux, OS X, Solaris.

Riak

Riak is the most powerful open-source, distributed database. Different Users have Comcast, Yammer, Voxer, Boeing, SEOMoz, Joyent, and DotCloud, Formspring, the Danish Government and many others.

Operating System: Linux, OS X.

Features of Riak are

- **Low-Latency:** Riak is designed to store data and serve requests predictably and quickly, even during peak times.
- **Availability:** Riak replicates and retrieves data intelligently, making it available for read and write operations even in failure conditions.
- **Operational Simplicity:** Riak allows you to add machines to the cluster easily, without a large operational burden.
- **Big Data:** This distributed database can run on a single system or scale to hundreds or thousands of machines.

Operating System: OS Independent.

Data Mining

Rapid Miner/Rapid Analytics

Rapid Miner is the world-leading open-source system for data and text mining. In addition to the open source versions of each, enterprise versions and paid support are also available from the same site.

Operating System: OS Independent.

Mahout

It is Apache project and offers algorithms for clustering, classification and batch-based collaborative filtering that run on top of Hadoop [11]. The project's goal is to build scalable machine learning libraries. **Operating System:** OS Independent.

Weka

It stands for "Waikato Environment for Knowledge Analysis". It [12] offers a set of algorithms for data mining that you can apply directly to data or use in other Java application. It is part of a larger machine learning project and sponsored by Pentaho. **Operating System:** Windows, Linux, OS X.

File Systems

Gluster

This file system tool is sponsored by Red Hat. This tool offers unified file and object storage for very large datasets.

Operating System: Linux.

Hadoop Distributed File System

It is known as HDFS [13], this is the primary storage system for Hadoop. It efficiently replicates data onto several nodes in a cluster in order to provide reliable, fast performance.

Operating System: Windows, Linux, OS X.

Big Data Search

Lucene

The self-proclaimed "de facto standard for search libraries," Lucene offers very fast indexing and searching for very large datasets. In fact, it can index over 95GB/hour when using modern hardware. **Operating System:** OS Independent.

Solr

Solr is an enterprise search platform based on the Lucene tools. It powers the search capabilities for many large sites, including Netflix, AOL, CNET and Zappos.

Operating System: OS Independent.

Freely Available Big Data Tools

- **R: -** R is a tool that is being used in programming languages. It is a free software environment for free statistical

computing and graphics. It compiles and runs on wide variety of UNIX platforms, Windows & Mac OS. It is an integrated suite of software facilities for calculation, data manipulation and graphical display. It also possesses some more functionalities like effective data handling and storage facility, a group of operators for calculation on arrays in particular matrices. R language is widely used by statisticians and data miners for statistical software and data analysis.

- **WEKA: -** It is a tool used in data mining tasks. WEKA is a collection of machine learning algorithms. WEKA features include machine learning, data mining, pre-processing, classification, regression, clustering and visualisation. Its main user interface is the explorer; the same functionality can also be accessed through the component -based knowledge flow interface and from the command line.

BIG Data Tools Comparison

Table 1: Comparison of Tools

S. No	Attributes	MongoDb	WEKA	Cassandra	Riak
1.	**DEVELOPER**	MongoDb Inc.	University of Waikato	Apache Software foundation	Basho Technologies
2.	**PROGRAMMING LANGUAGE**	C++, C, JavaScript	Java	Java	Erlang
3.	**SOURCE CODE**	Open source & Free	Open source & free	Open source	Open source
4.	**LICENSE**	GNU AGPL V3.0	GNU General public	Apache license 2.0	Apache license 2.0
5.	**OPERATING SYSTEM**	Windows, Linux, OSX, Solaris	Windows, OSX, Linux	OS independent	Linux, BSD, Solaris, Mac OSX
6.	**TYPE**	Document oriented database	Machine learning, Datamining	Database	NoSQL Database, Cloud storage

Big Data Tools Comparison

Table 2: Comparison of Tools

S. No	Attributes	Couch DB	Hadoop	R	Rapid Miner
1.	**DEVELOPER**	Apache software foundation	Apache software foundation	R Development core team	Rapid miner
2.	**PROGRAMMING LANGUAGE**	Erlang	Java	Implementation S programming language	
3.	**SOURCE CODE**	Open source	Open source	Closed & free under GNU	Open source
4.	**LICENSE**	Apache license	Apache license 2.0	GNU General public license	AGPL
5.	**OPERATING SYSTEM**	Cross platform	Cross Platform	Cross Platform	Cross Platform
6.	**TYPE**	Document oriented database	Distributed file system	Programming language	Statistical Analysis, Predictive analysis

CONCLUSIONS

Big Data is an emerging problem for large companies and organizations, as massive volumes of data are being generated, examined, stored and analysed in our day to day life. The demanding problems of big data can be categorized as issues related to data variety, velocity (speed), volume, and veracity. This need to process large quantities of data has never been too large. In the commercial sphere, business intelligence, driven by the ability to gather data from a dizzying array of sources. Big Data analysis tools like Map Reduce over Hadoop and HDFS, promises to help organizations better understand their customers and the marketplace, hopefully leading to better business decisions and competitive advantages.

REFRENCE

1. C. A. :. A. C. a. I. Ltd., "Big Data Spectrum," p. 61, 2012.
2. "Big Data Analytics for Security Intelligence," 2013. [Online]. Available: www.cloudsecurityalliance.org/research/big-data.
3. IBM, [Online]. Available: http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html.
4. "TEXATA 2014,the Big Data World Analytics Championship," 2014. [Online]. Available: http://www.texata.com/resources/definitions/.
5. "TEXATA 2014,the Big Data Analytics World Championship," 2014. [Online]. Available: http://www.texata.com/resources/definitions/.
6. "TEXATA 2014,the Big Data Analytics World Championship," 2014. [Online]. Available: http://www.texata.com/resources/definitions/.
7. AlekseyYeschenko, Third Party support, 2014. [Online]. Available: http://wiki.apache.org/cassandra/ThirdPartySupport.
8. G. Kunz, 2013. [Online]. Available: http://wiki.apache.org/cassandra/Durability.
9. "MONGODB Mannual3.0," 2013-2015. [Online]. Available: http://docs.mongodb.org/manual/core/data-modeling-introduction/?_ga=1.148671095.1560745460.1426416514.
10. "The easiest way to run MONGODB," [Online]. Available: https://mms.mongodb.com/?_ga=1.48588518.1560745460.1426416514.
11. P. Warden, Big Data Glossary, SHROFF PUBLISHERS & DISTRIBUTORS PVT.LTD.
12. P. Warden, Big Data Glossary, SHROFF PUBLISHERS & DISTRIBUTORS PVT.LTD..
13. P. Warden, Big Data Glossary, SHROFF PUBLISHERS & DISTRIBUTORS PVT.LTD .
14. "Revolution Analytics," [Online]. Available: http://www.revolutionanalytics.com/.
15. MAVNE, The Apache Software Foundation, 2011-2014. [Online]. Available: http://oozie.apache.org/.