Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org
Volume 5, Special Issue 2, December (2015)

Analysis of Big Data Tools and Approaches

Amrita Ticku
CSE. ,BSAITM, Faridabad

Mohammad Danish
CSE , AFSET , Faridabad

Abstract – When keeping old data becomes hectic we generally try to find some solutions for storing data in a particular format so that we can easily find the required data .This concern turns to be further compulsory for producing various tools in big data. The main goal of big data analytics is to utilize the advanced analytic techniques besides very big, different datasets which contain multiple sizes from terabytes to zettabytes and multiple types like structured or unstructured and batch or streaming. Without having proper knowledge of big data problem we upload data on internet and that creates big problem for storage that too the data is scattered so we have to store and also in a particular format so that we can easily access the particular data in right time. In this research paper, a various collection of big data tools are illustrated and also compared with their essential features.

Index Terms – Big data, Big data approaches, Big data analytics, Data analysis, Data visualization.

This paper is presented at International Conference on Recent Trends in Computer and information Technology Research on 25th & 26th September (2015) conducted by B. S. Anangpuria Institute of Technology & Management, Village-Alampur, Ballabgarh-Sohna Road, Faridabad.

1. INTRODUCTION

Big Data technologies are transforming the way data is used to be analyzed. One reason is the massive amount of data that is being generated from different sources such as Facebook, orkut, sensors; google and other search engines, banks, telecommunication and web, medical science handling this massive amount of data takes us in the era of Big Data. According to the study of you tube statistics 200 hours of video are being uploaded to YouTube servers in every minute. Facebook is dealing with more than 700 terabytes of data daily, companies such as a Google and Yahoo are recording search engine results for analyzing the searching trends; crawling different web sources to analyze for any important events; gathering marketing data for analyzing the current and future trends which all results the generation of large data sets also referred to as Big Data. Data is everywhere, from social sciences to physical science, business and commercial world, for example, digitizing the past fifty year's newspapers will results the massive amount of data, in astronomy storing billions of astronomical objects, in biology storing genes, proteins and small molecules results in massive amounts of data. In business world such as handling millions of call data records in telecommunication, handling millions of transactions in banking and handling millions of transactions for multinational grocery store results in large data sets. Analysing these large datasets and getting out meaningful information from it is a challenging in it.

1.1 Data Analysis

Nowadays the data uploaded on web is increasing day by day and also different types of information are provided by n-number of user. The variety of data is also increasing that can easily view in any search engine. The big data is defined as datasets whose size is beyond the ability of typical database tools to store, capture, manage and analyze. The best way to define big data is via five V’s which is data volume, data velocity and data variety or variability [1]. The data volume is regarding size (terabytes and petabytes), records, transactions and tables, files. The data velocity is about how frequently data is generated by an application in real time or in streaming. The data variety includes different types of data that is structured (simply RDBMS), semi-structured (XML, RSS feeds) and unstructured (text, human languages). Big data is remarkably diverse in terms data types, sources and entities represented.

1.2 Building block of big data

There are certain characteristics of big data which are listed below [2]:

- **Volume** – The volume is related to the size of data. At present data is in petabytes and in near future it will be of zettabytes.

- **Variety** – The data is not coming from single source it includes semi structured data like web pages, log files etc, raw, and structured and unstructured data.
- **Variability**–The variability considers inconsistencies in data flow.
- **Value** – The value is importance of data used by the user. The user queries against certain data stored, obtains result, rank them and can store for future work.
- **Velocity** – The velocity is defined as the speed of data coming from different sources. The speed of uploading the data is not limited and is also not constant.
- **Complexity** – The data is coming from various resources in huge amount thus it is difficult to link or correlate multiple data

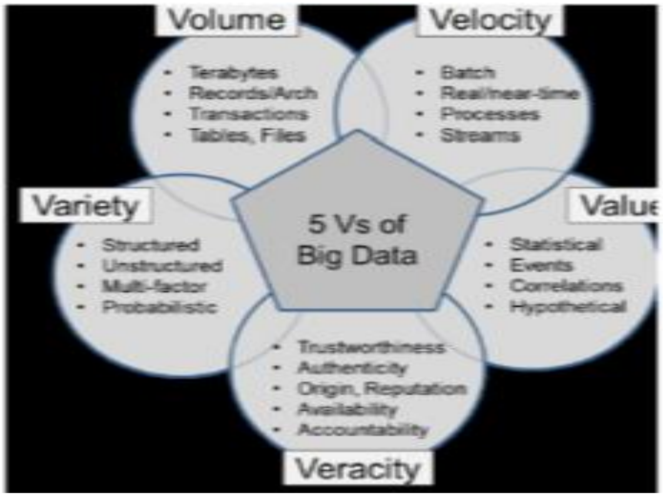

Figure 1 Building Blocks of Big Data

1.3 Analysis of big data

In today’s era everybody is having an electronic device is producing data and lots of it. Figure 1 gives a broad overview of possible sources. Whenever we want do something new we upload it whether it is necessary or not and by default we are increasing the size of data. We’d like to show what is possible and what is acceptable in data collection. Tracking electronic devices is just one of many approaches. [3]

We will provide a quick overview of the architecture behind Big Data, but we will not discuss data analysis algorithms. Online processing software will be discussed as it will explain later how data and information transfer is possible. Through showing the possibilities of Big Data we’d like to advertise data sharing to governments. Towards this goal we will provide a discussion about the public trust. [2]

1.4 Components of big data

There are mainly five types of big data which are discussed below.

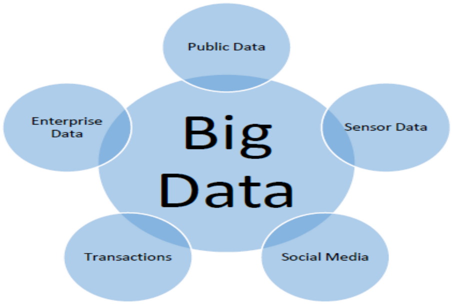

Figure 2 Components of Big Data

2. RELATED WORK

2.1 HIVE web based interface

Hive has an interactive web interface which is designed for the administration purpose of the hive and as well as for querying the database. User can create and delete tables and browse the database schema. In addition, users can execute the queries by supplying it from the web based interface. it is not really an analytical platform, but it was developed to work with hive easily and interactively by using the web interface. The web interface is easy to use, but it is too technical for non-technical users to create and browser schema and other tasks such as starting and stopping the hive. Also for hive web interface to work, the user should have hive configured and deployed on the computer system which requires hadoop as well for processing, hence makes it very complex for non-expert users with no support of analytics.

2.2 IBMInfoSphere BigInsights

InfoSphere BigInsights is a Big Data analytics platform from IBM, which support different type of analytics under one roof. InfoSphere is built on top of Hadoop to enhance its capabilities and provides an interactive interface on it for analyzing the Big Data. InfoSphere has built-in analytics capability, including text analytics for getting insights from large textual data, social data analyzer for analyzing social media data, machine data analytics for analyzing machine data such as data from sensors and GPS and InfoSphere also supports the integration with other Big Data technologies. In addition InfoSphere provides a SQL interface, namely as BigSQL and also a spreadsheet like interface called BigSheets for analyzing and exploring Big Data easily with development tools analytics and security for Big Data operations. BigSheets and BigSQL modules of InfoSphere are some of the core components of the system [7,4].

2.3 SAP BigData Analytics:

SAP is one of the leading providers of Big Data analytics platforms. It is one of the first companies to introduce in-memory database for analytics. The aim of building in-memory database was to provide a single database for both transactional and analytical data processing, commonly refer to as OLTP4 and OLAP5 systems. However, any applications can be built on top of SAP HANA6 in-memory database for analytics. However SAP HANA back-end engine is available for use in the cloud as well, but there are less application exists, which can be used by the non - expert user for analytics. HANA has a strong library called Predictive analytics library (PAL) which can be used by the HANA developers to perform predictive calculations by just invoking the library functions hence providing ease for Big Data analytics [5][6].

2.4 Teradata BigData Analytics

The Aster Big Analytics Appliance is solution from TERADATA for Big Data Analytics. Aster is basically a database developed by TERADATA supporting row based and column based storage and it's the key component of their Big Data analytical platform which consists of Aster database, Aster SQL-MapReduce, which is basically an interfacing SQL language for Aster 4 Online Transaction Processing database but include support for built-in functions for MapReduce using SQL language similarly to Hive which sits on top Hadoop along with their own supplied and tested hardware. We will discuss in detail about the architecture and working of Hive and Hadoop in the next chapter.

The Aster Big Analytics Appliance is a combination of software and hardware and the basic idea is to run Big Data technologies on a dedicated and specially designed hardware by connecting multiple nodes with InfiniBand [7] instead of running Hadoop on commodity hardware with traditional network connections. New nodes can be added as needed by using separate dedicated hardware machines [8].

2.5 Cloudera BigData Solutions

Cloudera founded in 2008 and was the first enterprise distributors of Apache Hadoop and other Big Data technologies. The contributions of Cloudera to Big Data world is the abstraction over different Big Data technologies such as Hadoop, Hive, HCatalog [9] etc., which provides easiness to users to use these technologies without going into technical details.

Cloudera solutions are ready to install on any commodity hardware which hides the technical details of compiling and configuring the Big Data technologies and provides the system management such as configuration, deployment, security management, diagnostics, operational reports generation etc. Cloudera is at forefront of providing back-end solutions for Big Data exploration and analysis but does not provide any tool or framework which actually can be used on top of these technologies especially for non-expert users [10].

2.6 HortonWorks

HortonWorks founded in 2011 by the ex-Yahoo engineers. The concept is same as Cloudera to provide different Big Data technologies for enterprise computing and as a result the company developed HortonWork Data Platform for enterprise computing which builds on top of Hadoop and have the capability to provide different type of analysis such as batch, real time and interactive. The platform also supports data management, the facility which is built on top of the Hadoop file system and resource manager. In addition to multiple analysis capabilities of the Horton data platform, it supports different technologies for data integration and data flow control. Technologies such as Apache Falcon [11], Sqoop [12]. and Flume [13] are part of the platform and provide easy and systematic access for handling data in and out of Hadoop [14].

2.7 Amazon Big Data Analytics Platform

Amazon is a pioneer in cloud computing and provides technological services through web interface usually referred as web services. Amazon elastic map reduce (EMR) is a web service from amazon where we can access Big Data technologies such as Hadoop and Hive via Amazon exposed API's or via launching an EC29 instance, where users can deploy map and reduce script for processing on Hadoop and Hive. Hive can be accessed via the amazon exposed API's such as JDBC or by using a Hive compatible client such as Hive command line utility and beeline.

Non-expert users which do not have technical grounds to write map and reduce scripts are not able to use this technology and also non-programmer cannot work with the exposed API's as they have to be proficient in programming in order to work with them. In addition, users need to build an application on top of Amazon web services to provide the access to non-expert users for big data exploration.

2.8 Oracle Big Data Platform

Oracle is not behind from any other company for Big Data solutions. Oracle has introduced the in memory database option in oracle database version of 12c. Oracle partners with Cloudera to provide Hadoop for Big data analytics. In addition, Oracle has developed its own analytical tools using R [15] for Hadoop together with In-memory and NoSQL databases. In addition the system has a data management facility for moving data in and out of Hadoop and to and from the database with dedicated and tested hardware from Oracle. Despite the powerful hardware and software capability to deliver real insights into the Big Data there is a need to build a simple application on top of this technology to support non-expert user for exploring the Big Data [16].

2.9 Hewlwtt Packward Big Data Platform

HP provides analytics through its HP Vertica analytics platform. The product is basically based on Vertica database which is the backbone of the platform. The database provides fast execution of queries and especially for set type queries and for warehouse applications due to its columnar storage structure. In addition database uses advanced compression techniques to compress the data to be stored on the disk along with in-database analytics. Furthermore, the database has integration with Hadoop for processing stored data in the Hadoop File System Related Work which can be loaded into the database through Hadoop-Vertica connector. However system does not provide the flexibility to execute MapReduce jobs into the Hadoop instead platform is more interested to get only the data from the Hadoop distributed file system for processing by using rich in-database analytical functions [17].

3. BIG DATA ANALYTICS TECHNIQUES

Several techniques are drawn on disciplines such as computer science and statistics that can be used to analyze huge datasets. This section takes a brief look at some of categories of commonly used techniques for analyzing big data. All the techniques listed can be applied to diverse and larger datasets [18].

3.1 A/B Testing

This technique is also known as bucket testing or split testing. In this technique control group is compared with a variety of test groups in order to determine what changes will improve a given objective variable. Large number of tests are analyzed and executed to ensure that group is of sufficient size to determine meaningful differences between treatment and control groups. Multivariate generalization of this technique is often called “A/B/N” testing.

3.2 Network Analysis

This analysis algorithm is used to detect relationships between the nodes in a graph or in a network. It is useful for social networks where important information regarding user, his friends, relatives etc. can be obtained. Central users in the network can also be discovered by social network analysis.

3.3 Association Rule Learning

Set of techniques which find interesting relationship among variables in large datasets. These techniques include variety of algorithms to generate and test possible rules. Used in market analysis for their benefits where retailer can determine what products are sold.

3.4 Genetic Algorithm

It is an optimization technique that is based on the “survival of the fittest” for optimal solution. In this technique solutions are encoded as chromosomes that can be combined and mutated. The potential of chromosomes is determined by its fitness value. The best ones are copied into next generation. It is an evolutionary algorithm as it is well suited for solving non-linear problems. It can be used for the improvement of scheduling of jobs in manufacturing and further optimize the performance.

3.5 Sentiments Analysis

To determine subjectivity of reviewers and opinion is the aim of sentiment analysis which is a Natural Language Processing (NLP). With increase in popularity of various websites where users can provide their opinion for other users and items that are purchased, the internet is full of comments, reviews and ratings. The main aim of sentiment analysis is to classify user opinion. To determine reaction about the product and actions of user, companies use this analysis.

3.6 Machine Learning

A sub domain of the artificial intelligence and is related with the design and development of algorithms which senses the behaviors based on historical data. Machine learning basically focuses on automatic recognition of complex patterns and makes intelligent decision based on data.

3.7 Cluster Analysis

This is an unsupervised learning technique which classifies user in smaller subgroup with similar properties which are not known in advance. It is different from classification. In markets it can be used to find customer with similar taste.

3.8 Pattern Reorganisation

This technique generates some kind of output according to a given input value according to some specific algorithm.

3.9 Predictive Modelling

In this set of techniques a mathematical model is created or chosen according to the best probability of an outcome. This technique is helpful in determining customer's likelihood for the customer relation manager.

4. KEY APPROACHES TO ANALYZE BIG DATA

To analyze big data and generate insight, there are four key approaches to uncover hidden relationships [19].

4.1 Discovery Tool

These tools can be used for analysis along with traditional business intelligent source system. They are helpful throughout information lifecycle for rapid analysis of information from any data source. The user can draw new ideas and meaningful conclusions and can make intelligent decisions quickly.

4.2 Business Intelligence (BI) Tools

These tools are important for analysing, monitoring and performance management of data from different data sources. They help in taking the decisions for the business by providing comprehensive capabilities, ad-hoc analysis on the large scale platform.

4.3 Decision Management

These tools consist of business rules, predictive modeling and self -learning to perform action based on current context. They maximize customer interaction by suggestions. It can be integrated with Oracle Advanced Analytics to execute complex predictive models and make real time decision.

4.4 In-Database Analytics

They consist of different techniques which help in finding relationship and pattern in data. This technique is helpful as it removes the movement of data to different locations thus reducing total cost of ownership and information cycle time.

5. COMPARISON OF BIG DATA ANALYTICS TOOLS

The following Table I demonstrate the comparison of different tools in big data based on compatible data sources and its operating system. The main objective of this comparison is not to criticize which is the best tool in big data, but to demonstrate its usage and to create alertness in various fields.

Big Data Tools	Mode	Data Types	Data Sources	Database Support
JASPERSOFT	Commercial and Open Source	Structured and Unstructured data	JDBC, Delimited text, Positional text, LDIF, XML	Mongo DB, Cassandra, Redis, Riak, CouchDB Neo4j, HBase
SPLUNK	Commercial	Unstructured data Time-series textual	Files, the network scripted Outputs	Relational IBM Database 2, SAP, Sybase
TABLEAU	Commercial	Structured and Unstructured data	Database, Cubes, Hadoop Cloud	MySQL, Microsoft SQL Server, Oracle, EMC, GreenPlum
KARMASPHERE	Commercial and Open Source	Structured, Semi- Structured and Unstructured data	Web logs, Mobile devices, and Sensors	Base HDFS file data

Table 1 : Comparison of Big Data Analytics Tools

6. CONCLUSION

Big data provide very effective supporting processes for summing up of data sets which are too complex and large. This compulsory requirement gives the path for developing tools in big data research. Whereas these tools are generated both in real time and also in very large scale which comes from sensors, web, networks, audio/video, etc. Thus the aim of this survey is to enhance the knowledge in big data tools and their applications applied in various companies. It also provides obliging services for readers, researches, business users and analysts to make enhanced and quicker decisions using data. Data is everywhere, from social sciences to physical science, business and commercial world, for example, digitizing the past fifty year's newspapers will results the massive amount of data, in astronomy storing billions of astronomical objects, in biology storing genes, proteins and small molecules results in massive amounts of data. In business world such as handling millions of call data records in telecommunication, handling millions of transactions in banking and handling millions of transactions for multinational grocery store results in large data sets. Analyzing these large datasets and getting out meaningful information from it is a challenging in it. The big data is defined as datasets whose size is beyond the ability of typical database tools to store, capture, manage and analyze. The best way to define big data is via three vs. which are data volume, data velocity and data variety or variability. The data volume is regarding size (terabytes and petabytes), records, transactions and tables, files. The data velocity is about how frequently data is generated by an application in real time or in streaming. The data variety includes different types of data that is structured (simply RDBMS), semi-structured (XML, RSS feeds) and unstructured (text, human languages). Big data is remarkably diverse in terms data types, sources and entities represented. There are different technologies to deal with Big Data analysis, but most of them are complex and requires expertise to deal with them. Especially for non-computer scientists such as social scientists, they require good programming skills and knowledge of configuring and maintaining the infrastructure which almost makes it impossible for them to explore the large data sets or to perform ad hoc analytics on it. For example, how a social scientist can explore the data to find an event in 1975 by having the previous fifty years of newspaper data? Or how a social scientist can predict the human behavior by analyzing its previous 5 years of data gathered from different sources such as cell phone records with GPS tracking, search engine queries, internet transaction data, consumer behavior or its social network activity? There are plenty of Big Data analysis platforms or frameworks are available nowadays in the market, but the problem for non-computer scientists is to master them because of the complexity involves with them and where necessary to take training in order to use them for exploring Big Data in ad-hoc manners and doing analytics on large data sets. These systems inherit the problem of maintaining them as well, which might include at application or infrastructure level. In this research work I have represents a detailed study of Big Data, Big Data Analytics technologies and tools which provide a better idea about the big data world. There are many future important challenges in Big Data management and analytics that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal during the next years: Analytics Architecture, Statistical significance, Distributed mining , Hidden Big Data.

[1] References

[2] P. Russom, “Big Data Analytics,” TDWI Best Practices Report, 4th Quarter 2011, 2011

[3] Avita Katal, Mohammad Wazid, and R H Goudar, "Big data: Issues, challenges, tools and Good practices," In Contemporary Computing (IC3), 2013 Sixth International Conference, 2013, pp. 404-409.

[4] Simona Candrian and Richard Conrardy,” Big Data-Challenges and opportunities for Governments”.

[5] IBM.[Online].http://www.01.ibm.com/software/data/infosphere/biginsights/.[4]

[6] SAP Big Data Solutions. [Online]. http://www.sap.com/solution/big-data.html.5

[7] SAP HANA In-Memory Database. [Online]. http://www.saphana.com.6

[8] Yi-Man Ma, Che-Rung Lee, and Yeh-Ching Chung, "InfiniBand virtualization on KVM," in 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom),2012, pp. 777-781.7

[9] TERADATA Aster http://www.asterdata.com.

[10] https://cwiki.apache.org/confluence/display/Hive/HCatalog.

[11] http://www.cloudera.com/content/cloudera/en/home.html.

[12] The R Project for Staistical Computing. [Online]. http://www.r-project.org/

[13] Oracle Big Data. [Online].

[14] HP Vertica Analytics Platform. [Online]. http://www.vertica.com/

[15] J. Manyika, M. Chui, B. Brown, J. Bughhin, R. Dobbs, C. Roxburgh and A.H. Byers “Big data: The next frontier for innovation, competition and productivity,” The McKinsey Global Institute, Tech. Rep., May 2011.

[16] An Oracle White Paper, “Big Data Analytics, Advanced Analytics in Oracle Databases,” March 2013.