Indian J.Sci.Res. 14 (1): 176-180, 2017 ISSN: 0976-2876 (Print)
ISSN: 2250-0138 (Online)

A SURVEY ON DATA ANALYTIC TOOLS

D. MAGHESH KUMAR[a1], D. CHRISTYSUJATHA[b] AND M. CHANDARAKUMAR PETER[c]

[abc]Department of Software Engineering. Periyar Maniammai Univeristy, Thanjavur, India

ABSTRACT

In the current circumstances the volume of data are created and put away by different enterprises are quickly developing on the web accordingly information researchers are confronting a ton of difficulties for keeping up an immense measure of information as the quickly developing ventures require the huge data for upgrading the business and for prescient examination of the data. As information proceeds down its way of development, a noteworthy test has gotten to be the manner by which to manage the blast of information and investigation of this information. The term 'Big Data', refers to data sets whose size (volume), complexity (variability) and rate of growth (velocity) make them hard to capture, manage, process or analyzed. The main objective of this paper is to give a top to bottom examination of various data analytic tools accessible for performing Big data analysis. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, Apache Hive, No SQL and HPCC. This paper surveys different tools available for big data analytics and assesses the advantages and drawbacks of each of these tools based on various metrics.

KEYWORDS: Big Data, Hadoop, Map Reduce, Apache Hive, No SQL, HPCC.

Under the touchy increment of worldwide information, the term of Big data is for the most part used to portray gigantic datasets. Contrasted and customary datasets, Big data ordinarily incorporates masses of unstructured data that need all the more ongoing investigation. What's more, Big data additionally achieves new open doors for finding new values, helps us to pick up an inside and out comprehension of the concealed qualities, and furthermore brings about new challenges, e.g., how to successfully arrange and oversee such datasets.

Big Data is as an accumulation of vast dataset that can't be prepared utilizing customary figuring systems. Big Data is not merely a data rather it has become a complete subject which involve various tools, techniques and framework. The need of Big data created from the substantial organizations like facebook, yahoo, Google, YouTube and so forth with the end goal for the analysis of enormous amount of data also Google contains the large amount of information. Recently, industries become interested in the high potential of big data, and many government agencies announced major plans to accelerate big data research and applications [Fact, 2012]. The era of big data has come beyond all doubt [Manyika et.al., 2011].

Big Data is a term that alludes to dataset whose volume (measure), complexity and rate of growth (velocity) make them excessively troublesome, making it impossible to captured, managed, processed or analyzed by traditional technologies and tools, such as relational databases. There are different advanced technologies in the market from various vendors including Amazon, IBM, Microsoft, and so forth., to deal with Big Data.

The data in it will be of three types.

Structured data: Relational data.

Semi Structured data: XML data.

Unstructured data: Word, PDF, Text, Media Logs.

FIVE V`S OF BIGDATA

There are many properties associated with big data. The prominent aspects are Volume, Variety, Velocity, Variability and Value. and they are shown in figure 1.

Volume

Volume alludes to size of data. Volume speaks to the size of the data how the data is large. The size of the data is spoken to in terabytes and petabytes**.**

Variety

Variety makes the data too large. The files come from different sources and of any sort, it might be structured or unstructured, for example, text, audio, videos, log files and more are only the tip of the iceberg.

Velocity

Velocity alludes to the speed of data preparing. The data comes at very fast. Sometimes 1 minute is too late so big data is time sensitive.

[1]**Corresponding author**

Veracity

Veracity alludes to clam or, predispositions and variation from the norm when we managing high volume, speed and assortment of data, the all of data are not going 100% right, there will be filthy data.

Value

The potential estimation of Big data is colossal. Value is fundamental hot spot for enormous data since it is critical for organizations, IT framework to store large amount of values in database. All enterprises and e-commerce systems are keen in improving the customer relationship by providing value added services. For that, contemplate on client states of mind and trends in the market are to be analyzed. In addition, clients can likewise inquiry the data store to discover business trends and appropriately they can change their techniques. By making Big data open to all, it makes transparency on functional analysis. . Supporting continuous choices and exploratory investigation in various areas datasets can do superb things for enterprises.

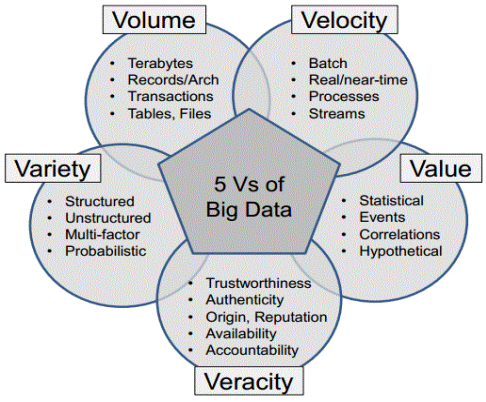

Figure 1: Properties of big data

Complexity: Data management gives extremely complex process, especially when immense volumes of data originate from several sources. This information should be connected, associated and related keeping in mind the end goal to make to gather the information that is thought to be conveyed by these data [Sonka, 2014].

5C`S OF DATA VISUALIZATION

Capture

Create data source connections to most any data source in short span of time.

Clean

The increased volume of data introduces an overall degradation in the quality of the data coming in. So the data is to cleaned.

Combine

Captured data in a common repository comes from varied sources. Making a single new source by combining the data which come from varied sources

Calculate

Calculations seem to gain a mind of their own in every reporting system.

Control

Control is what you gain by putting together the C's into your data before visualization and agile in your decision making process

BIGDATA EXAMPLES

Financial Services

To anticipating the client bank related exercises, for example, Insurance and debit cards by utilizing Bigdata analysis. The financial services sector has experienced uncommon changes in the most recent couple of years. Clients are need a more individual/protection information benefit from their banks. Bigdata can change the way work together to bridles the ideal incentive in client data, re-modeling the association with the market. Bigdata can include both inward and outside data utilizing a several sources [www.acquia.com].

Airlines and Tracking Companies

Fuel usage and Airlines examples are checked to improve skill and give effectiveness by utilizing Bigdata. Today Bigdata is comprehended and acquired as high volume, high speed and high assortment of data. It can create economic value and help with operations, decision making and risk management and customer services. Carriers are gathering the data as great way, yet they haven't generally well at utilizing it. In the Airlines business immense measure of data are putting away and handling then dropped even as carriers accumulation more of it. Terabytes of client's data is coasting around at any given time inside its system. These works are done by the Bigdata Analytics.

Healthcare Providers

Healthcare providers is utilized to track the drug effectiveness. The Healthcare providers suppliers separate themselves by specific patient results and experience from the patient, reference from patients and therapeutic work force. It is additionally enhancing and overseeing costs, more adaptable patient checking. To enhance nature of care and reduction cost of care [www.acquia.com].

Telecommunications

To analyze client practices and demand patterns. Today telecom organizations are extremely famous. They have a lot of data and they require legitimate burrowing and investigation both structured and unstructured data. Client area and travel patterns are comprehended by global mobile communications. To distinguish impacts from extortion by utilizing Bigdata at the same time diminishing the volume of support. Telecom division's utilization of data to develop at annual rate of data.

Finally Bigdata has the ability to

- Identify, combine and manage multiple sources of data.
- Make advanced analytics models for determine and processing outcomes.
- Process the organizational data for yielding better decisions.

Agriculture

In monetary development and sustenance security of agro based nation, Agriculture assumes a vital part. Crops determination is the fundamental issue from agriculture planning. Different parameters are utilized as a part of Agriculture industry that implies it relies on the productivity rate, market rate, market farm, government controls and policies.[Kumar et.al., 2015]

DATA ANALYTIC TOOLS

Column Oriented Database

In usual, online exchanges are for the most part utilizing row oriented database with high handling speed, yet it yields falls in the performance on query execution. Rather than rows, columns can be utilized as a part of column oriented database. It permits colossal data compression. A column-oriented database stores each column continuously. i.e. on disk or in-memory each column on the left will be stored in sequential blocks.

For analytical queries that perform aggregate operations over a small number of columns retrieving data in this format is extremely fast. As PC storage is optimized for block access, by storing the data beside each other we exploit locality of reference. On hard disk drives this is particularly important which due to their performance characteristics provide optimal performance for sequential access. The major column-oriented databases include:

- MonetDB (open source)
- C-Store (open source)
- Teradata
- Vectorwise/Paraccel
- Sybase IQ

Nosql Database

It is for the most part concentrate on the storage and recovering extensive volumes of structured, semi structured or Unstructured. Nosql [Cattell, 2011] gives read-write consistency. The most important result of the rise of NoSQL is Polyglot Persistence. NoSQL does not have a prescriptive definition but we can make a set of common observations, such as:

- Not using the relational model
- Running well on clusters
- Mostly open-source
- Built for the 21st century web estates
- Schema-less

Application developers have been frustrated with the impedance mismatch between the relational data structures and the in-memory data structures of the application. Using NoSQL databases allows developers to develop without having to convert in-memory structures to relational structures.

In a distributed system, managing consistency (C), availability (A) and partition toleration (P) is important, Eric Brewer put forth the CAP theorem which states that in any distributed system we can choose only two of consistency, availability or partition tolerance. Many NoSQL databases try to provide options where the developer has choices where they can tune the database as per their needs.

Hadoop

Hadoop is a name that represents two items, one a child's toy and the other an open source framework for distributed storage and processing of big data. In both contexts, interaction with Hadoop is foundational in personal growth and development. This learning path covers content that is critical to your success in this realm. It takes you on a journey that explains the Hadoop conceptual design, then it looks how to use the application and then manipulate data without the use of complex coding(Figure 2). This open source framework which allows to store and performing huge amount of data in a Bigdata environment among clusters of commodity hardware. It made to scale up from single servers to thousands machines. Hadoop is mainly written in java that provides distributed storage. Hadoop Architecture simply have four modules. Hadoop Common, Hadoop YARN [Vavilapalli, et.al., 2013], HDFS [Borthakur, 2008], Hadoop MapReduce [Dean and Ghemawat, 2008].

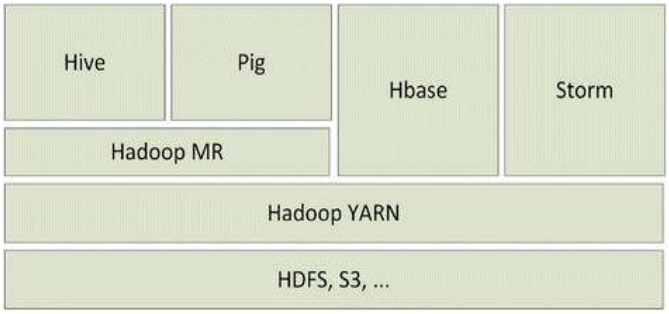

Figure 2: Hadoop Stack showing different components.

Map Reduce

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce [Dean and Ghemawat, 2008] model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. It is the data retrieving programming paradigm which allows for parallel massive job execution.

Hive

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

To process structured data by using Hive. It is a Data warehouse structure. Firstly, Hive was developed by Facebook, later the Apache foundation captures it up and developed it further works. Hive is used by different companies for example. Amazon. It is not a relational database, Hive [Olston, et.al., 2008] is mainly used for developing SQL Scripts.

Pig

It is a "Perl-like" language which allows mainly for query execution over data across the cluster, instead of "SQL-like "language. Yahoo was developed the PIG [Thusoo et.al.]. It is also an open source framework.

Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

Platfora

One of the main issue of Hadoop is very low level Map Reduce implementation. PLATFORA is a platform that simplifies Hadoop jobs automatically. To create an abstraction layer in which anyone can exploit to collect and organize datasets related to Hadoop.

CONCLUSION

In this paper, we discussed about Bigdata terminologies, numerous technologies and tools concerned in Bigdata. In actual-world applications managing and mining Big Data is Challenging mission, As the data concern large in a extent, distributed and decentralized control and complicated. There are numerous challenges at data, model and system level. We want computing platform to handle this Big Data. With Big Data technologies we able to offer maximum

applicable and accurate social sensing comments to better understand to society at real-time.

REFERENCES

Fact sheet: Big data across the federal government, 2012. http://www.whitehouse.gov /sites/default/files/microsites/ostp/big data fact sheet 3 29 2012.pdf

Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C. and Byers A.H., 2011. Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Sonka S., 2014. Big Data and the Ag Sector: More than Lots of Numbers", International Food and Agribusiness Management Review, **17**(1).

http://www.acquia.com/examples-big-data-projects.

Kumar R.G. et. al., 2015. Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique", 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM).

Cattell R., 2011. Scalable sql and nosql data stores. ACM SIGMOD Record, **39**(4):12–27.

Hadoop., [http://hadoop.apache.org/]

Vavilapalli V.K., Murthy A.C., Douglas C., Agarwal S., Konar M., Evans R., Graves T., Lowe J., Shah H. and Seth S. 2013. Apache hadoop yarn: Yet another resource negotiator. Proceedings of the 4th annual Symposium on Cloud Computing, 5.

Borthakur D., 2008. HDFS architecture guide. Hadoop Apache Project.

Dean J. and Ghemawat S., 2008. MapReduce: simplified data processing on large clusters. Commun ACM, **51**(1):107–113. 10.1145/1327452. 1327492

Olston C., Reed B., Srivastava U., Kumar R. and Tomkins A., 2008. Pig latin: a not-so-foreign language for data processing. In Proceedings of the ACM SIGMOD international conference on Management of Data. ACM, 1099–1110.

Thusoo A., Sarma J.S., Jain N., Shao Z., Chakka P., Anthony S., Liu H., Wyckoff P. and Murthy R., : Hive: a warehousing solution over a map-reduce framework. Proceedings of the VLDB.