Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey on Importance and Tools Used: Big Data

Rama Devi Gunnam*, Pratyusha Gudavalli, Reshma Pothuri
Asst. Professor, Department of CSE, MICT,
India

Abstract— Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. The primary purpose of this paper is to provide an in-depth analysis of different platforms available for performing big data analytics. The challenges include analysis, capture, search, sharing, storage, transfer, visualization, and privacy violations. The tools like Hadoop, R, Python and Visualization Tools like Tableau, D3, and Data Wrapper.

Keywords— Big data, Visualization Tools, Techniques, Hadoop, Tableau.

I. INTRODUCTION

Big Data will be stored at different places and also the data volumes may increase with the increase in data and hence to collect data from various places becomes expensive. Big Data is creating new generation of decision support data management. An example of big data might be petabytes (1,024 terabytes) or Exabyte's (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact centre, social media, mobile data and so on).

II. BIG DATA

❖ Importance of Big Data

When big data is effectively and efficiently captured, processed, and analysed, companies are able to gain a more complete understanding of their business, customers, products, competitors, etc. which can lead to efficiency improvements, increased sales, lower costs, better customer service, and/or improved products and services.
The effective use of big data exist in the following areas

- Using information technology (IT) logs to improve IT troubleshooting and security breach detection, speed, effectiveness, and future occurrence prevention.
- Use of social media content in order to better and more quickly understand customer sentiment about you/your customers, and improve products, services, and customer interaction.
- Fraud detection and prevention in any industry that processes financial transactions online, such as shopping, banking, investing, insurance and health care claims.
- Use of financial market transaction information to more quickly assess risk and take corrective action.

❖ Sources of Big Data

1) *Public Data*: Public data includes data that is publicly available like data generated by government sectors, weather data, Wikipedia, research data, open source data and other data which is freely available to the public. This type of data accessible to all is referred to as Public Data.
2) *Transactional Data*: Every enterprise will have some kind of applications which performs different kinds of transactions like Mobile Applications, Web Applications and many more. In order to support the transactions of these type, there are one or more relational databases which works at backend. This type of data is structured and it is referred to as Transactional Data.
3) *Social Media*: Huge amount of data is being generated on social networks like Twitter, LinkedIn, Face book, etc. Thus social media has to capture and manage unstructured.
4) *Enterprise Data*: Huge amount of data comes from enterprises in different formats. Formats may be in the form of flat files, Word documents, emails, spreadsheets, PowerPoint presentations, HTML pages, pdf files, XMLs, legacy formats, etc. This data which is spread across the organization in different formats is referred to as Enterprise Data.
5) *Activity Generated data*: Data that has been generated by machines that surpasses the data volume generated by humans. These include data from various machines like images from medical devices, data from sensors, surveillance videos, satellites data and data from mobile towers. These types of data are referred to as Activity Generated data.
6) *Archives*: Archives are the data which is very rarely required or which is not required anymore for any organization. Now a day's cost of the hardware is so cheap that none of the organization would like to discard any data, they would like to capture and store as much data as possible. Archived data includes records of ex-employees, old bank transactions, scanned documents, agreements copies, completed projects, this type of data which is less frequently accessed is referred to as Archive Data.

❖ **Barriers**

1) *Unstructured data*: There are two types of data in storage, structured and unstructured data. Structured data has a high degree of organization, and is typically stored in a relational database that can be easily searched. Unstructured data is, obviously, not structured in any meaningful way, including such things as photographs, videos, MP3 files, etc. Unstructured data is difficult to search and analyze.

2) *I/O barriers*: If you're dealing with something like mapping genomes, gathering information from the Mars Rover or running sophisticated weather simulations, the transaction volumes of these data sets challenge traditional storage systems, which don't have enough processing power to keep up with the huge number of I/O requests.

3) *Management*: There are a million and one storage management tools out there. The most basic one – and one still in wide use even in business, believe it or not, is a simple Excel spreadsheet – but vendors from EMC to Hitachi Data Systems to NetApp offer solid storage management solutions. The trouble is, though, that data-sharing standards are still lacking and escaping vendor-lock is a never-ending challenge.

4) *The WAN*: As cloud computing becomes mainstream, the simplest way to break down data silos is to leverage the cloud to help with everything from search to backups to raw processing. However, as more storage moves into the cloud, the more the WAN will impede on Big Data progress. The WAN, unfortunately, isn't keeping up with Moore's Law, nor with the storage-specific analog Kryder's Law. Any Big Data storage solution must include some combination of redundant MPLS links, WAN optimization and CDN services.

5) *Security:* As you break down data barriers, certain people may get access to data (say HR records) that they should never, ever see. Thus, authentication, access and security in general are a major Achilles heel of Big Data storage.

III. TOOLS OF BIG DATA

1) Python

Python is a powerful, flexible, open-source language that is easy to learn, easy to use, and has powerful libraries for data manipulation and analysis. Its simple syntax is very accessible to programming novices, and will look familiar to anyone with experience in Mat lab, C/C++, Java, or Visual Basic. For over a decade, Python has been used in scientific computing and highly quantitative domains such as finance, oil and gas, physics, and signal processing. It has been used to improve Space Shuttle mission design, process images from the Hubble Space Telescope, and was instrumental in orchestrating the physics experiments which led to the discovery of the Higgs Boson (the so-called "God particle").

According to the TIOBE index, Python is one of the most popular programming languages in the world, ranking higher than Perl, Ruby, and JavaScript by a wide margin. Among modern languages, its agility and the productivity of Python-based solutions are legendary. The future of python depends on how many service providers allow for SDKs in python and also the extent to which python modules expand the portfolio of python apps.

2) R

R is an open source programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians for developing statistical software and data analysis. According to Rexer's Annual Data Miner Survey in 2010, R has become the data mining tool used by more data miners (43%) than any other. The S language is often the vehicle of choice for research in statistical methodology, and R provides an open source route to participation in that activity.

R is emerging as a defacto standard for computational statistics and predictive analytics. R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:

- An effective data handling and storage facility.
- A suite of operators for calculations on arrays, in particular matrices.
- A large, coherent, integrated collection of intermediate tools for data analysis.
- Graphical facilities for data analysis and display either on-screen or on hardcopy.
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

3) Hadoop

The name Hadoop has become synonymous with big data. It's an open-source software framework for distributed storage of very large datasets on computer clusters. Fig(4) Relation between Data Management and Data Analysis All that means you can scale your data up and down without having to worry about hardware failures.

Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop is not for the data beginner. To truly harness its power, you really need to know Java.

It might be a commitment, but Hadoop is certainly worth the effort – since tons of other companies and technologies run off of it or integrate with it. But Hadoop Map Reduce is a batch-oriented system, and doesn't lend itself well towards interactive applications; real-time operations like stream processing; and other, more sophisticated computations.

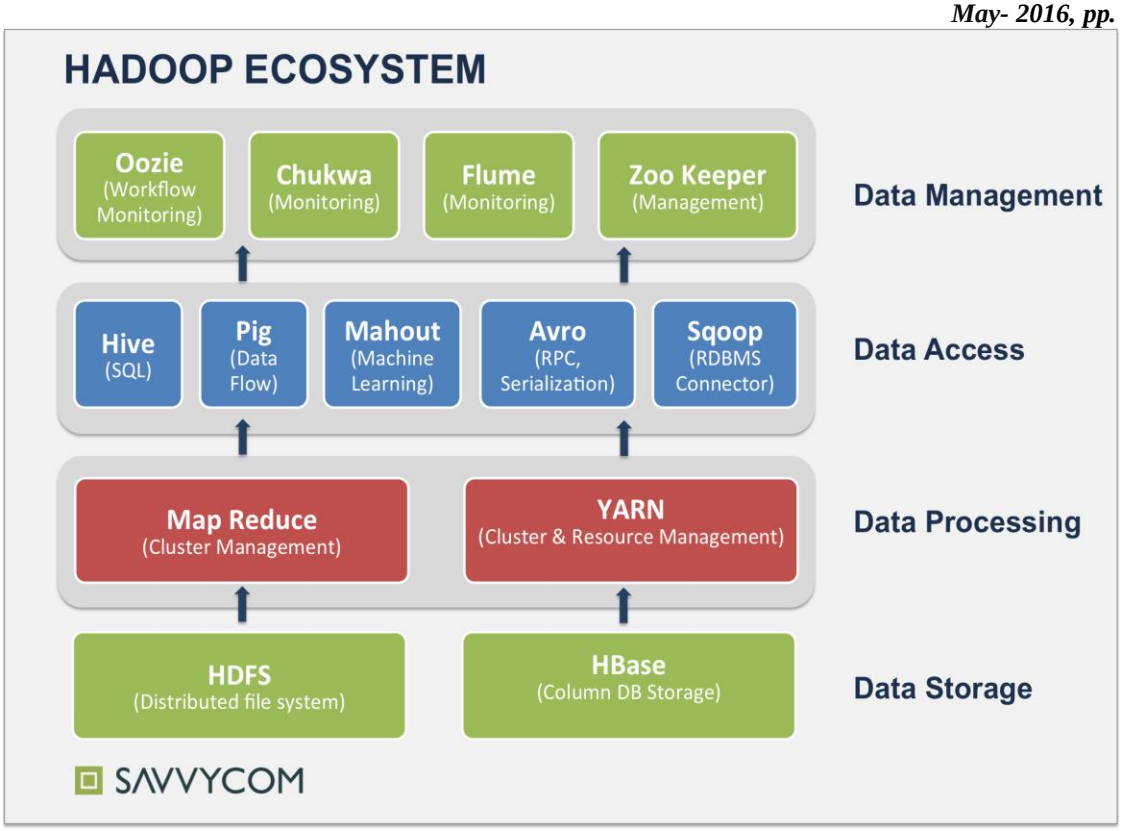

Fig 1: Apache Hadoop Ecosystem

4) Hive

Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users.

5) PIG

PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

6) WibiData

WibiData is a combination of web analytics with Hadoop, being built on top of HBase, which is itself a database layer on top of Hadoop. It allows web sites to better explore and work with their user data, enabling real-time responses to user behavior, such as serving personalized content, recommendations and decisions.

7) PLATFORA

Perhaps the greatest limitation of Hadoop is that it is a very low-level implementation of MapReduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional databases.

PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

8) SkyTree

SkyTree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data. Machine learning, in turn, is an essential part of Big Data, since the massive data volumes make manual exploration, or even conventional automated exploration methods unfeasible or too expensive.

9) Big Data in the cloud

As we can see, from Dr. Kaur's roundup above, most, if not all, of these technologies are closely associated with the cloud. Most cloud vendors are already offering hosted Hadoop clusters that can be scaled on demand according to their user's needs. Also, many of the products and platforms mentioned are either entirely cloud-based or have cloud versions themselves.Big Data and cloud computing go hand-in-hand. Cloud computing enables companies of all sizes to get more value from their data than ever before, by enabling blazing-fast analytics at a fraction of previous costs. This, in turn

drives companies to acquire and store even more data, creating more need for processing power and driving a virtuous circle.

IV. VISUALIZATION TOOLS

Data visualization is a modern branch of descriptive statistics. It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information". Some of the tools are

1) D3

You should use D3.js because it lets you build the data visualization framework that you want. Graphic / Data Visualization frameworks make a great deal of decisions to make the framework easy to use. D3.js focuses on binding data to DOM elements. 3 stand for Data Driven Documents. We will explore D3.js for its graphing capabilities.

2) Data wrapper:

Data wrapper allows you to create charts and maps in four steps. The tool reduces the time you need to create your visualizations from hours to minutes. It's easy to use – all you need to do is to upload your data, choose a chart or a map and publish it. Data wrapper is built for customization to your needs; Layouts and visualizations can adapt based on your style guide.

3) Tableau

This software adopts a very different mental model as compared to using programming to produce data analysis. Think about the first GUI that made computers public-friendly, suddenly the product has been repositioned. "Pretty Graphs" are useless if they just look pretty and tell you nothing. But sometimes making data look pretty and digestible also makes it understood to the average person. Tableau occupies a niche to allow non-programmers and business types to do guaranteed hiccup-free ingestion of datasets, fast exploration and very quickly generate powerful plots, with interactivity, animation etc.

V. CONCLUSION

Big Data is going to continue growing during the next years,and each data scientist will have to manage much more amount of data every year.This This data is going to more diverse,larger,faster and is becoming the new scientific data research and for business applications.Big datamining is new era which is help to discover knowledge. Big data analysis helps business people to make better decisions and researchers to identify new opportunities. This paper presents fundamental concepts of Bigdata like characteristics, sources, statistics, frameworks and technologies to handle big data

REFERENCES

[1] Suman Arora, Dr.Madhu Goel, "*Survey Paper on Scheduling in Hadoop*" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[2] Puneet Singh Duggal ,Sanchita Paul ," *Big Data Analysis:Challenges and Solutions*", *International Conference on Cloud, Big Data and Trust 2013, Nov 13-15.*

[3] Singh and Reddy, "A Survey on platforms for big data Analytics" Journal of Big Data 2014.

[4] Eckerson, W. (2011) "Big Data Analytics: Profiling the Use of Analytical Platforms in User Organizations," TDWI, September. Available at http://tdwi.org/login/default-login.aspx?src=%7bC26074AC-998F-431B-C994C39EA400F4F%7d&qstring=tc%3dassetpg.

[5] Blog post: Thoran Rodrigues in Big Data Analytics, titled "10 emerging technologies for Big Data", December 4, 2012

