Volume 9, Special Issue No. 1, February 2018

ISBN: 978-93-5300-455-2

International Journal of Advanced Research in Computer Science
(ISSN: 0976-5697)

REVIEW ARTICLE

Available Online at www.ijarcs.info

SHRI SHANKARLAL SUNDARBAI SHASUN JAIN COLLEGE FOR WOMEN
ICRACC 2018
(23rd–24th February 2018)

Big Data Analytics and Data Science – A Review on Tools and Techniques

Abirami.K, RajaMeenakshi.S and Supriya.R
Shri Shankarlal Sundarbai Shasun Jain College for Women, T.Nagar, Chennai, India

ABSTRACT

Big Data is a humongous amount of data in any form. Big data Analytics is the process to perform statistical analysis of data. Data science is go-ahead approaches that analyze past and current data in exploratory way to predict the occurrence of particular event in future. Tools used in big data and data science are to extract knowledge from the data in addition tools yield intelligence data solutions to business. This paper objective is to provide the essential details about Big Data, Data Science and tools in which through innovative analytics are achieved. Tools in Big data platform can be classified to accomplish few key tasks akin to Storage, Querying and Analysing. This paper also deals with Apache Hadoop, Hive, EXCEL, R-Programming, and Tableau these tools obviate the programming aspect and provide a GUI to build predictive models.

***Keywords* -** Image Processing, Classification, Grading, Neural Network

I. INTRODUCTION

Big data deals with large amount of structured, semi structured and unstructured data. Data that is very large or unstructured must be converted to structured data which cannot be achieved by relational database engines. This type of data requires the concept of big data which provides structured data. Its process includes capturing of data, storing, search, updating and analyzing. Five dimensions of big data are volume (quantity), variety (difference), velocity (Speed), Veracity (accuracy) and Value (scope). Big data information arrives from different domains like healthcare, Banking Sectors and equity market. These data can be processed through many tools and techniques.

Data science is an interdisciplinary field of mathematics, statistics, computer science to extract knowledge and patterns from complex data. There is overwhelming flow of data from various sources like internet, e-commerce sites, cell phones and social media. This data can be structured or unstructured like videos, audio etc. Data science as a whole is related to compiling, purifying and analyzing the data. Data science is a composite of algorithms and tools from various disciplines to gather data, obtain insights, extract meaningful information and explicate it for decision making.

Data science is highly applied in various fields such as Marketing, Social media, healthcare, education, security, biological science etc.

Data Analytics also known as Analysis of data. Data analysis refines the data using diverse techniques. Analysis process evolves collection of data, applying statistical methods to derive data for decision making. Data analysis methods can be a quality or quantity based. How data analysis techniques imply to big data is a challenge in this era

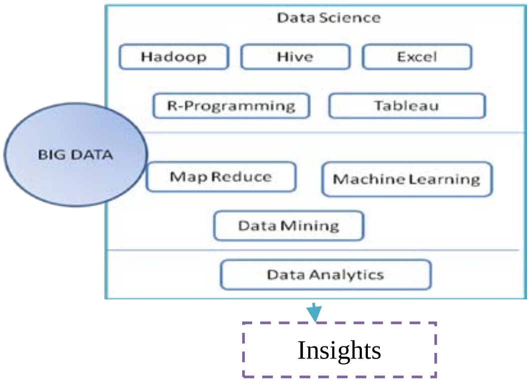


Fig: 1 stack of Bigdata, Data science, Tools and techniques

Above diagram depicts the flow of data from the various sources, also tools and techniques involve in data science and big data. When the enormous amount of data generated by the social media like Facebook, twitter, also data from healthcare domains turned into Big Data. Traditional tools and techniques fall short to handle big data. Thus, we couldn't derive knowledge from the data scattered around the system.

Both big data analytics and data science deals with unstructured data but big data analytics commonly used in financial services, retail marketing and healthcare whereas data science used to predict models in the new-fangled fields.

This paper explores the tools like Hadoop (storage), Hive (Querying), Excel (Analysing), R-Programming (Statistical Computing in Data Science), Tableau (Visualization) and the data mining, machine learning, Map Reduce techniques.

1. Hadoop

Hadoop is an open source framework from Apache.

Hadoop can handle data in huge volume. It is used to store the colossal amount of data in Big Data.

Hadoop Distributed File System and Map Reduce Engine are the main components in Hadoop. With the Hadoop Distributed File system the data is stored once on the server and consequently read and re-used many times thereafter.

It uses client/server structural design, with each cluster consisting of a single Name Node that manages file system operations and along with Data Nodes that manage data storage on individual compute nodes.MapReduce is to schedule and process across the cluster.hadoop is a data warehouse whereas MapReduce process the data by dividing into smaller units.

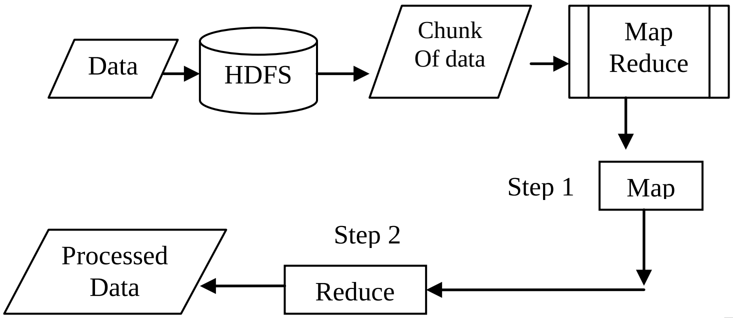


Fig 1.1 Data Process in Hadoop

Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data.

2. Hive

Apache provided another platform named as Hive to handle (querying) big data. HiveQL is a query language from Hive. Result set from this HiveQL is transferred to SQL Query set.

It does generate the MapReduce Scripts.

Like RDBMS hive facilitates storing of object in a binary stream and Meta data information through Hive Metastore.

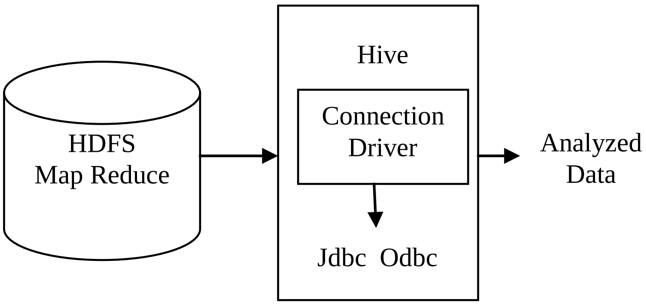


Fig: 2.1 Hive architecture

Hive is uniquely placed to come up with querying of data, powerful analysis, and data précising while working with huge volumes of data. The vital part of Hive is the HiveQL query which is an SQL-like interface that is used widely to query that is stored in databases.

Also Hive enables access data from Apache HBase storage system.

Connect Excel to Hadoop:through HiveODBC.it is possible to have a Hive add in to the Excel.

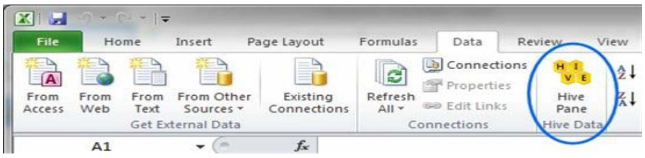
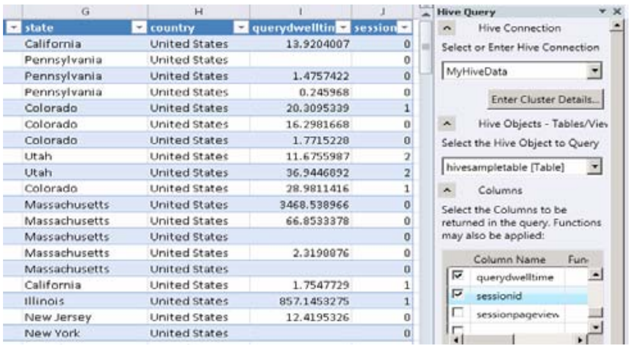


Fig 2.2 Hive Pane in Excel

Using the Hive Pane odbc connection is established and Hive data is imported to excel.



state	country	querydwelltin	session
California	United States	13.9204007	0
Pennsylvania	United States		0
Pennsylvania	United States	1.4757422	0
Pennsylvania	United States	0.245968	0
Colorado	United States	20.3095339	1
Colorado	United States	16.2981668	0
Colorado	United States	1.7715228	0
Utah	United States	11.6755987	2
Utah	United States	36.9446892	2
Colorado	United States	28.9811416	1
Massachusetts	United States	3468.538966	0
Massachusetts	United States	66.8533378	0
Massachusetts	United States		0
Massachusetts	United States	2.3190876	0
Massachusetts	United States		0
California	United States	1.7547729	1
Illinois	United States	857.1453275	1
New Jersey	United States	12.4195326	0
New York	United States		0

Hive Query

Hive Connection

Select or Enter Hive Connection

MyHiveData

Enter Cluster Details...

Hive Objects - Tables/Viev

Select the Hive Object to Query

hivesampletable [Table]

Columns

Select the Columns to be returned in the query. Functions may also be applied:

	Column Name	Fun
☑	querydwelltime	
☑	sessionid	
□	sessionpageview	

Fig 2.3 Hive data into Excel

3. Excel

Data stored in hadoop can be accessed by excel using ad hoc techniques. we can import data from hive to excel using HDInsight Services. Microsoft power query in excel used to extract the needed information from the data source. To accomplish this HDInsight cluster data is used.

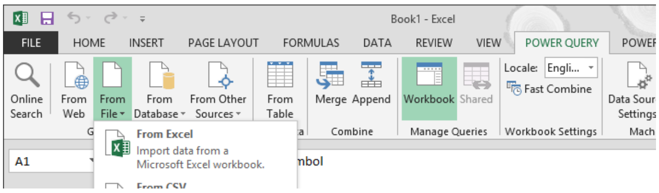


Fig 3.1 Power Query in Excel

Data cleaning and filtering are the crucial task of excel in big data.

4. R-Programming

It becomes the de facto programming language for data science. It performs statistical analysis of data.

RStudio framework is used to compile and execute R programming.

Data are manipulated using R programming. With the use of R Engine and IDE R-programming process the data and yields the data as statistical report.

Predominantly this language used to provide analyze patterns before the happening of the events.

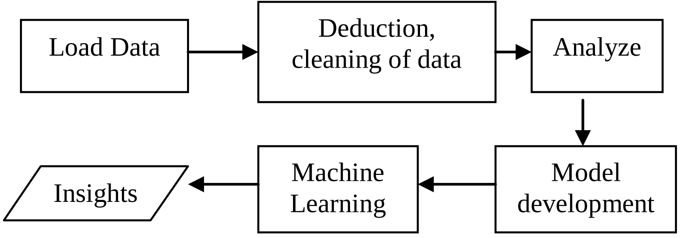


Fig 4.1 Data Process in R-Programming

5. Tableau

Tableau is primarily used in business intelligence platform which offer data visualization and exploration capabilities. When combined, Tableau and R offer one of the essential and complete data analytics solutions in the industry today,

providing businesses with incomparable abilities to see and understand their data.
A scalable object to various sources from tableau is the additional benefit.

Tableau has an interactive dashboard which renders the image as a result from analysis very quickly. The dashboard will also give you rich visualizations. The data visualization dashboard offer an in depth knowledge into the data.

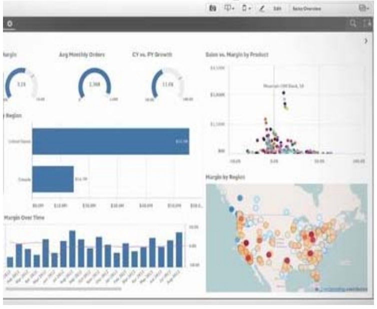


Fig 5.1 Tableau data output

It works with huge datasets and it includes more visualization tools.

6. Techniques

A. Machine learning (ML), a associate-field of artificial intelligence (AI), focuses on the task of enabling computational systems to learn from data about how to perform a desired task automatically. Machine learning applications including decision making, forecasting or predicting and it is a key enabling technology in the deployment of text mining and big data techniques in the diverse fields of engineering, healthcare, science, , business and finance.
Two types of Machine learning are:

- **Supervised ML:** The program is “practiced” on a pre-built set of “training examples”, which then facilitate its ability to reach an exact conclusion when given new data.
- **Unsupervised ML:** finite amount of data should be given as the program input and must find patterns, relationships therein.

Working sample of Machine Learning algorithm:

ML described as learning a target function (f) that best maps input variables (M) to an output variable (N).

$$N = f(M)$$

In the learning process we develop predictions for the future (N) and provided (M) new inputs.

7. Data Mining

Mining is the analyse process from various view and give a summary to useful information. This task can be automatic or semi automatic on heavy data sets.
Data mining is a knowledge Discovery process which applies to the enormous set of data. Like Big Data.

8. Map Reduce

It’s coupled with hadoop hdfs to handle big data. It incorporates stages.map stage, shuffle stage, reduce stage. Semantically the first two stages distribute the data and the reduce phase does the computation.:

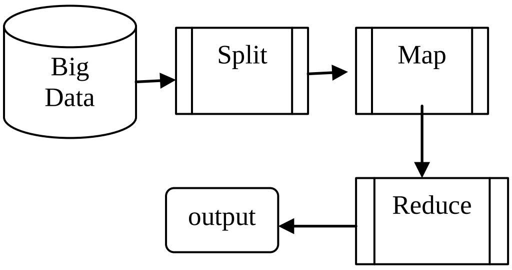


Fig 8.1 Steps in Map Reduce

Limitation of this technique is user has to follow a logic of his own.

9. Conclusion

This paper presents the overview on big data, data science and related tools, techniques. Tools discussed here provide the efficient way of handling big data. Understanding of this tools and its usage enables the researchers, business people, and academician to work with big data in an effective manner.

Future work:
Processing of ontology based data set using big data tools and to derive the insight for healthcare domain.

10. References

[1] M. R. Wigan and R. Clarke, ―Big Data’s Big Unintended Consequences‖, IEEE Computer Society, , vol. 46, no. 6, (2013), pp. 46-53.

[2] Abdul Raheem Syed, Kumar Gillela, Dr. C. Venugopal, “The Future Revolution On Big Data”, In International Journal of Advanced Research in Computer and Communication Engineering, 2013.Volume: 2 (Issue:6) , Page No. 2446- 2451.

[4] Daniel, B. K., Big Data and analytics in higher education: opportunities and challenges. British Journal of Educational Technology, 2015,46, 904–920.

[5] AmirGandomi, MurtazaHaider - Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, April 2015, Volume 35, Issue 2, Pages 137-144.

[6] Salisu Musa Borodo, Siti Mariyam Shamsuddin, Shafaatunnur Hasan - Big Data Platforms and Techniques, Indonesian Journal of Electrical Engineering and Computer Science Indonesian Vol. 1, No. 1, January 2016, pp. 191 -200.

[7] Proyag Pal,Triparna Mukherjee , Dr. Asoke Nath- Challenges in Data Science: A Comprehensive Study on

Application and Future Trends, August 2015,Volume 3, Issue 8.

[8] https://importoioweb.staging.wpengine.com/post/best-big-data-tools-use/.

[9] https://towardsdatascience.com/the-10-statistical-techniques-data-scientists-need-to-master-1ef6dbd531f7

[10] http://www.datascientists.net/what-is-data-science