Use of Data Science Technologies and Big Data Analytics

K.B.V. Rama Narasimham, K. Nethaji babu

Abstract— Information science is about managing huge nature of information with the end goal of separating important and coherent results/conclusions/designs. It's a recently developing field that envelops various exercises, for example, information mining and information examination. It utilizes strategies extending from arithmetic, measurements, and data innovation, PC programming, information designing, design acknowledgment and learning, representation, and superior processing. This paper gives a reasonable thought regarding the diverse information science innovations utilized as a part of Big information Analytics.

Index Terms— Big Data Analytics, Data science technologies, data wrapper, Hadoop, Big Data

1. Introduction

Information science exclusively manages getting experiences from the information while investigation additionally manages about what one needs to do to 'overcome any issues to the business' and 'comprehend the business cloisters'. It is the investigation of the techniques for dissecting information, methods for putting away it, and methods for displaying it. Regularly it is utilized to portray cross field investigations of overseeing, putting away, what's more, examining information joining software engineering, insights, information stockpiling, and comprehension. It is another field so there is not a agreement of precisely what is contained inside it.

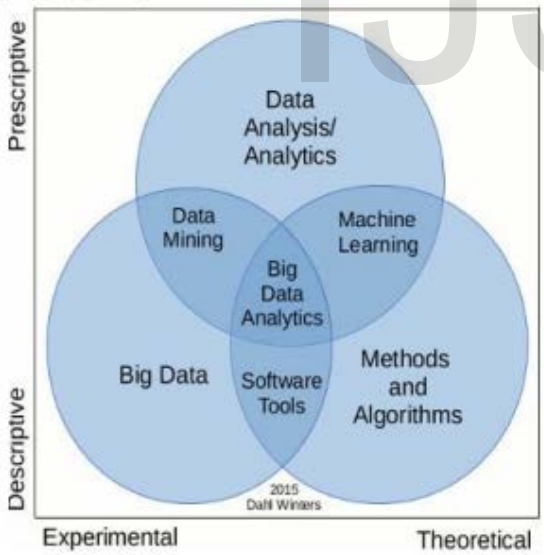

Fig 1. Data Science Fields

- *K.B.V Rama Narasimham completed his M.Tech in Computer Science and Engineering and he is currently working as a assistant professor in Guntur Engineering College.*
- *K. Nethaji babu completed his M.Tech in Computer Science Engineering from Guntur Engineering College, his areas of interest are Cloud Computing and Big Data.*

Information Science is a mix of arithmetic, measurements, programming, the setting of the issue being explained, sharp methods for catching information that may not be caught at this moment in addition to the capacity to take a gander at things" in an unexpected way" and obviously the huge and fundamental movement of purifying, get ready and adjusting the information. The real procedure of Data Science is shown in FIG 2

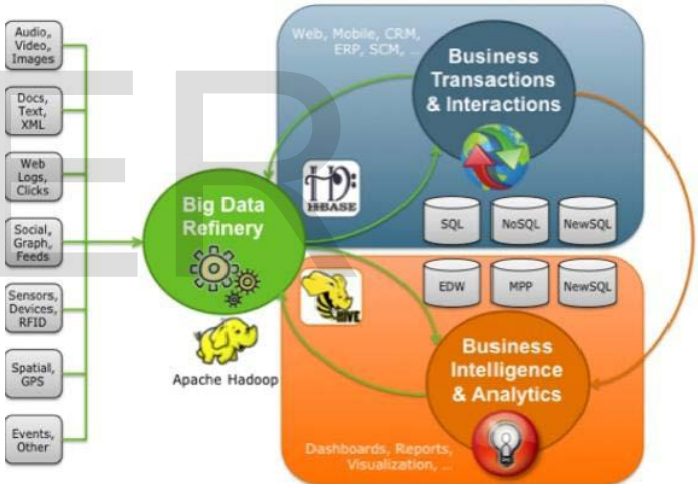

Fig 2 Big Data Architecture

2. Big Data

Big data is the collection of huge amounts of data, whether unstructured or structured. Today, many organizations are grouping, storing, and analyzing massive amounts of data. This data is normally said as "big data "because of its volume, the velocity with that it arrives, and the type of forms it takes. Big data is making a new generation of decision support data management. Businesses are identifying the potential worth of this data and are putting the technologies, people, and Processes in place to maximize the opportunities. A key to deriving worth from big data is the use of analytics.

Machine Learning is a branch of Computer Science that, instead of applying high-level algorithms to solve problems

in explicit, imperative logic, applies low-level algorithms to discover patterns implicit in the data. (Think about this like how the human brain learns from life experiences vs. from explicit instructions.) The more data, the more effective the learning, which is why machine learning and big data are intricately tied together.

Big Data Analytics

Big data not solely changes the tools one can use for predictive analytics, it also changes our entire manner of thinking regarding knowledge extraction and interpretation. Traditionally, data science has continually been dominated by trial-and-error analysis, an approach that becomes not possible once datasets are massive and heterogeneous. Ironically, availability of additional data sometimes leads to fewer choices in constructing predictive models, because very few tools allow processing massive datasets in a reasonable quantity of time. In addition, traditional statistical solutions usually focus on static analytics that is limited to the study of samples that are frozen in time, which typically ends up in surpassed and unreliable conclusions.

Let's begin with a real world example, looking at a farm that is growing strawberries

What would a farmer need to consider if they are growing strawberries? The farmer will be selecting the types of plants, fertilizers, pesticides. Also looking at machinery, transportation, storage and labor. Weather, water supply and pestilence are also likely concerns. Ultimately the farmer is also investigating the market price so supply and demand and timing of the harvest (which will determine the dates to prepare the soil, to plant, to thin out the crop, to nurture and to harvest) are also concerns.

Let's think about the data available to the farmer, here's a simplified breakdown:

1. Historic weather patterns
2. Plant breeding data and productivity for each Strain
3. Fertilizer specifications
4. Pesticide specifications
5. Soil productivity data
6. Pest cycle data
7. Machinery cost, reliability, fault
8. Water supply data
9. Historic supply and demand data
10. Market spot price and futures data

3 DATA SCIENCE TECHNIQUES TOOLS

3.1. 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.

3.2. 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:

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

IJSER

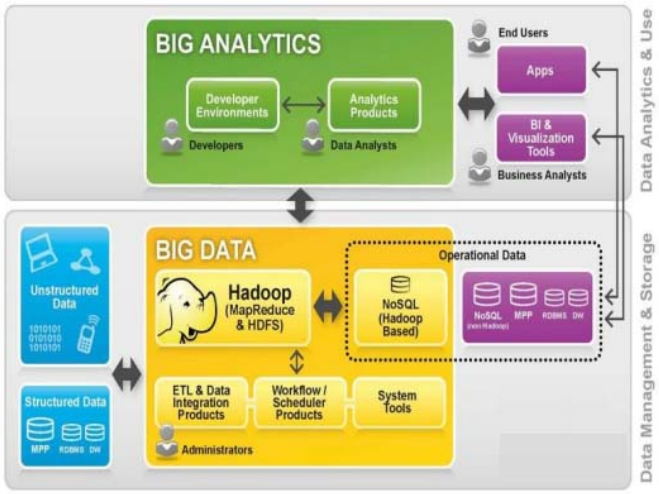

Fig 3 Data analysis and Data Management Relation

3.3. HADOOP:

The name Hadoop has become synonymous with huge data. It's an open-source software system framework for distributed storage of very massive datasets on pc clusters. All that means you'll be able to scale your data up and down while not having to stress regarding hardware failures. Hadoop provides massive amounts of storage for any kind of data, enormous process power and the ability to handle just about limitless concurrent tasks or jobs. Hadoop is not for the info beginner. To truly harness its power, you really have to be compelled to understand Java. It might be a commitment, but Hadoop is actually value the effort – since loads of different firms and technologies run off of it or integrate with it. But Hadoop Map scale back is a batch-oriented system, and doesn't lend itself well towards interactive applications; real-time operations like stream processing; and alternative, more refined computations.

3.4 Visualization Tools

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

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.

4. Work on Big data by data science technologies

Algorithms used for mining and analytics are being applied to Big Data sets, which implies a different approach to data management and processing. But it also means that ideas such as data exploration & data discovery are beginning to permeate modern every-day BI solutions. Below is an example from Pentaho where you can see that a chord does a good job of demonstrating connections, paths, and relationships between attributes and dimensions.

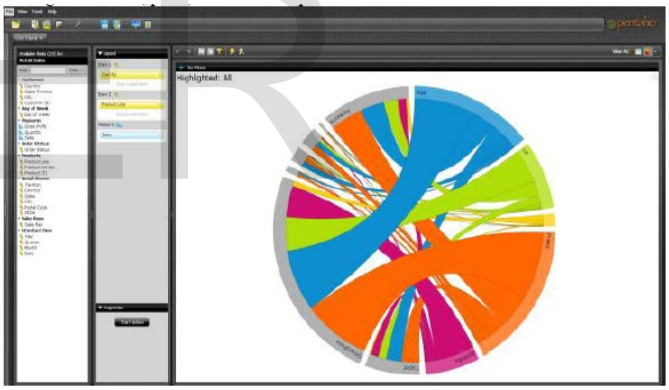

Fig. 4 relationships between attributes and dimensions

That comes from bigdatagov.org. We also use Chords often for our "data scientists" in Web analytics who are looking for paths to maximize conversions. Taking the chord idea to the next extreme comes from a project by Colin Owens at http://www.owensdesign.co.uk/hitch.html where he is exploring different pros & cons of visualizations that demonstrate relationships. Here you can see some of the chord's shortcomings in terms of showing influencers, a key aspect to marketing analytics:

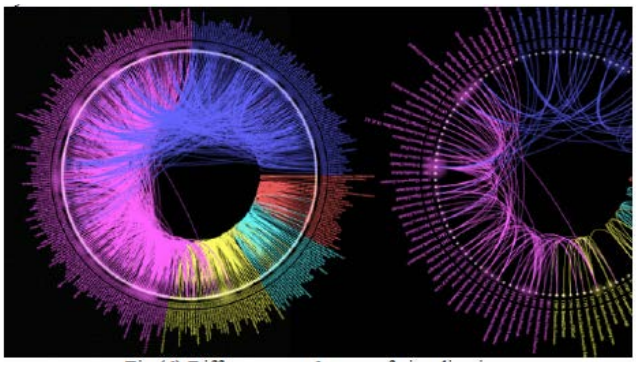

Fig. 5 pros and cons of visualization

But here is a great example of where the chord shines by using a data set that makes sense to most of us, not just statisticians. This should give you a good idea of the utility of a chord graph. In this case, Chris Walker used 2012 U.S. census data to show Americans moving between states in the U.S.

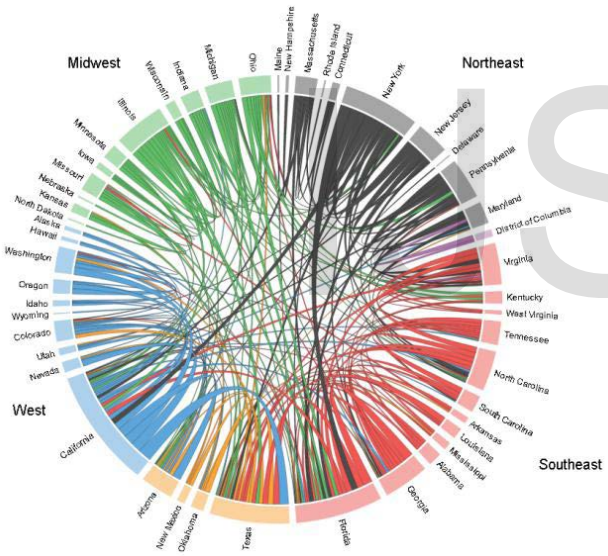

Fig 6 American Transportation between states

When you hover and select areas of the radial chord, you can easily see paths (very important in Web analytics and marketing) with size of links related to migrations.

5 Conclusion

The analysis of big data needs traditional tools like SQL, analytical workbenches and data analysis and visualisation languages like R. These tools can be utilized in numerous fields wherever data analytics is needed. Many additional tools are introduced within the market and also the existing products are underneath constant improvement. The demand for better analytics tools is increasing constantly that is solely planning to increase further in future.

REFERENCES

[1] Eckerson, W. (2011) "BigDataAnalytics: Profiling the Use of Analytical Platforms in User Organizations," TDWI, September. Available at http://tdwi.org/login/default-login.aspx?src=%7bC26074AC-998F-431B- BC994C39EA400F4F%7d &qstring=tc%3dassetpg

[2] "Research in Big Data and Analytics: An Overview" International Journal of Computer Applications (0975 –8887) Volume 108 –No 14, December 2014

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

[4] Douglas, Laney. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

[5] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, "Interactions with big data analytics," interactions, vol. 19, no. 3, pp. 50–59, May 2012

[6] Ari Banerjee senior analyst, heavy reading, "Big data and advanced analytics in Telecom: A Multi-Billion-Dollar Revenue Opportunity," December 2013.

IJSER