Volume 6, Issue 2, February 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Data Science Technologies & Big Data Analytics

T. Giri Babu
Research Scholar, Department of Computer Science
S.V.University, Tirupati, Andhra Pradesh, India

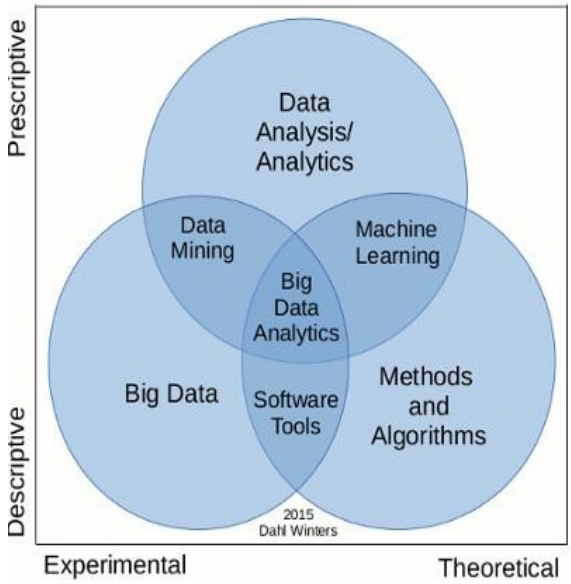
Dr. G. Anjan Babu
Associate Professor, Department of Computer Science
S.V.University, Tirupati, Andhra Pradesh, India

Abstract— Data science is about dealing with large quality of data for the purpose of extracting meaningful and logical results/conclusions/patterns. It's a newly emerging field that encompasses a number of activities, such as data mining and data analysis. It employs techniques ranging from mathematics, statistics, and information technology, computer programming, data engineering, pattern recognition and learning, visualization, and high performance computing. This paper gives a clear idea about the different data science technologies used in Big data Analytics.

Keywords— data science, analytics, data visualization, extraction, patterns

I. INTRODUCTION

. Data science solely deals with getting insights from the data whereas analytics also deals with about what one needs to do to 'bridge the gap to the business' and 'understand the business priories'. It is the study of the methods of analyzing data, ways of storing it, and ways of presenting it. Often it is used to describe cross field studies of managing, storing, and analyzing data combining computer science, statistics, data storage, and cognition. It is a new field so there is not a consensus of exactly what is contained within it.

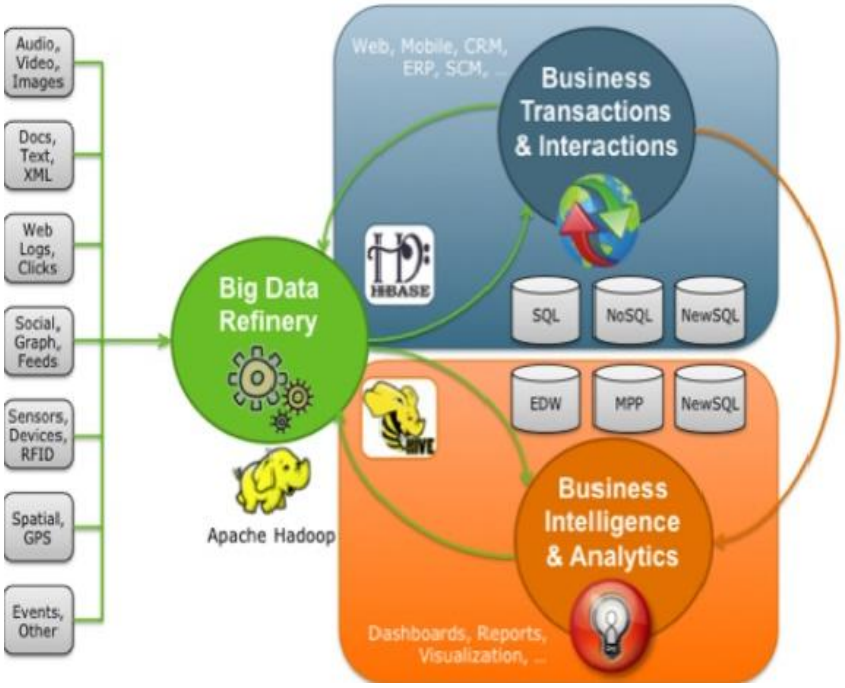


Fig(1) Fields of Data Science

Data Science is a combination of mathematics, statistics, programming, the context of the problem being solved, ingenious ways of capturing data that may not be being captured right now plus the ability to look at things 'differently' and of course the significant and necessary activity of cleansing, preparing and aligning the data. The actual process of Data Science is shown in fig (2).

II. BIG DATA

Big Data is the collection of massive amounts of information, whether unstructured or structured. Today, many organizations are collecting, storing, and analyzing massive amounts of data. This data is commonly referred to as "big data "because of its volume, the velocity with which it arrives, and the variety of forms it takes. Big data is creating a new generation of decision support data management. Businesses are recognizing the potential value of this data and are putting the technologies, people, and Processes in place to capitalize on the opportunities. A key to deriving value from big data is the use of analytics.



Fig(3) Next generation Big data Architecture

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 only changes the tools one can use for predictive analytics, it also changes our entire way of thinking about knowledge extraction and interpretation. Traditionally, data science has always been dominated by trial-and-error analysis, an approach that becomes impossible when datasets are large and heterogeneous. Ironically, availability of more data usually leads to fewer options in constructing predictive models, because very few tools allow for processing large datasets in a reasonable amount of time. In addition, traditional statistical solutions typically focus on static analytics that is limited to the analysis of samples that are frozen in time, which often results 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

III. TOOLS OF DATA SCIENCE TECHNOLOGIES

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.

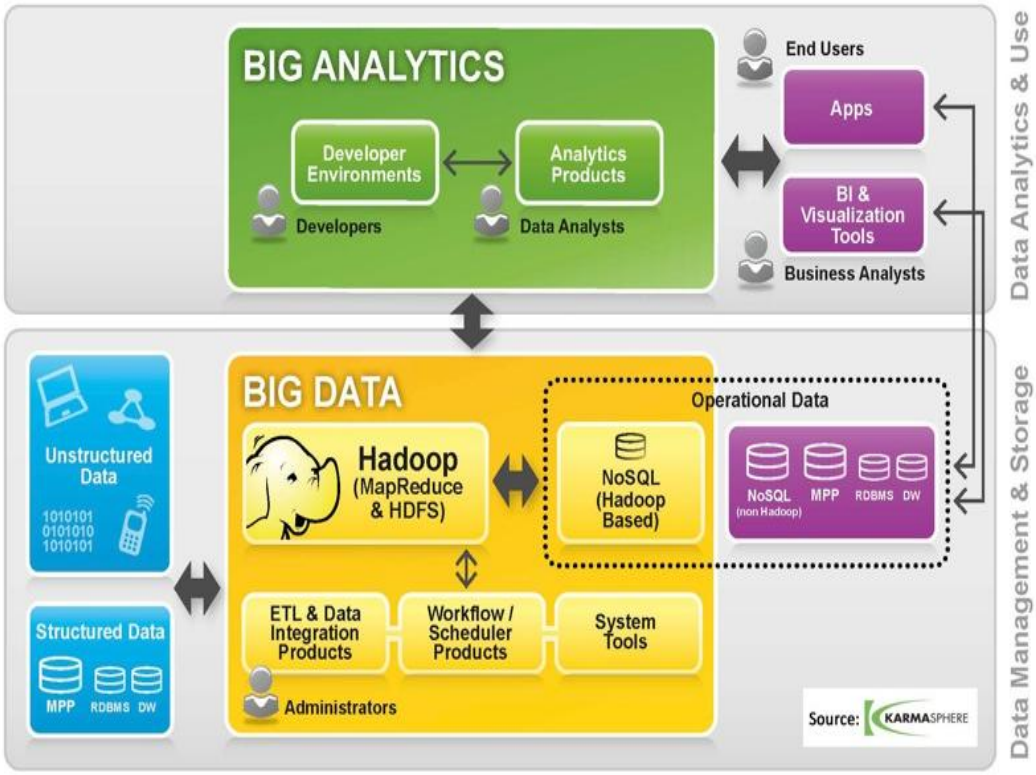
2) R

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.

4) 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

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.

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 **D**ata **D**riven **D**ocuments. We will explore D3.js for its graphing capabilities.

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.

IV. DATA SCIENCE TECHNOLOGIES WORK ON BIG DATA

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.

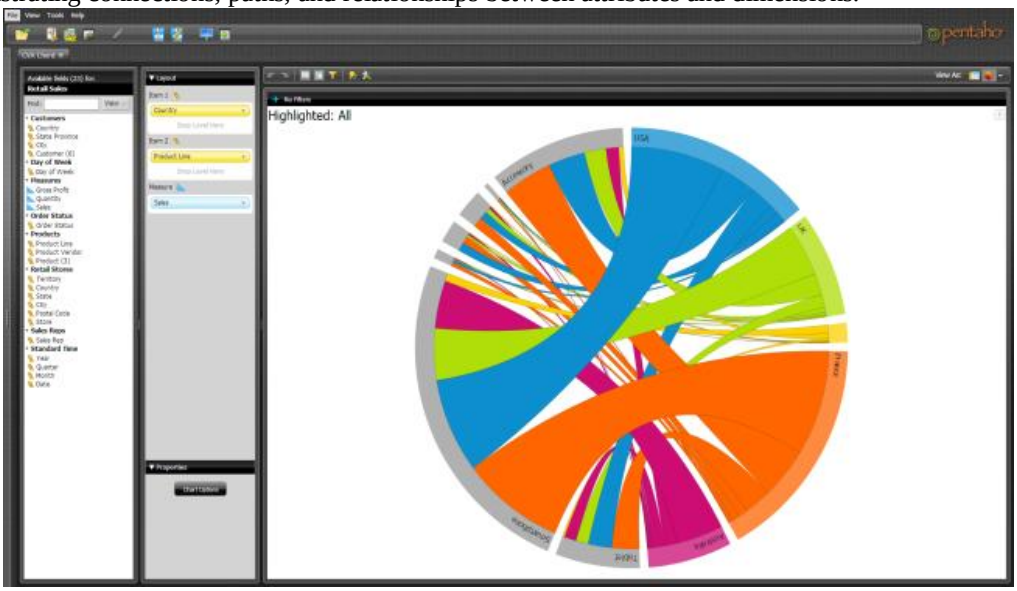


Fig (5): relationships between attributes and dimensions

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:

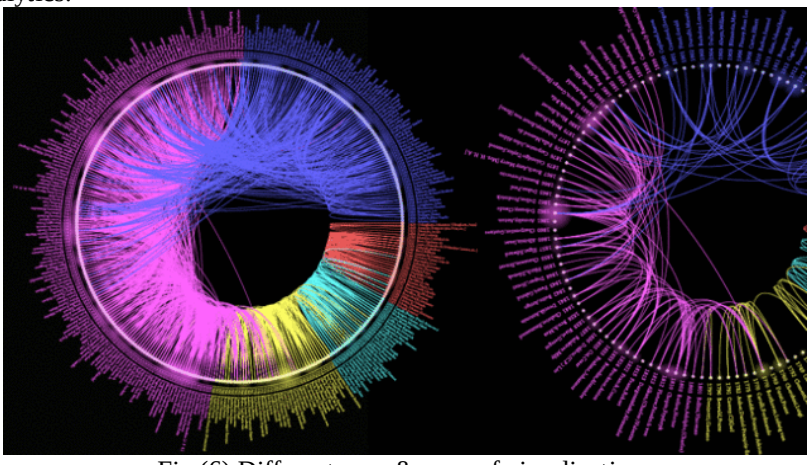


Fig (6) Different pros & cons of visualizations

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.

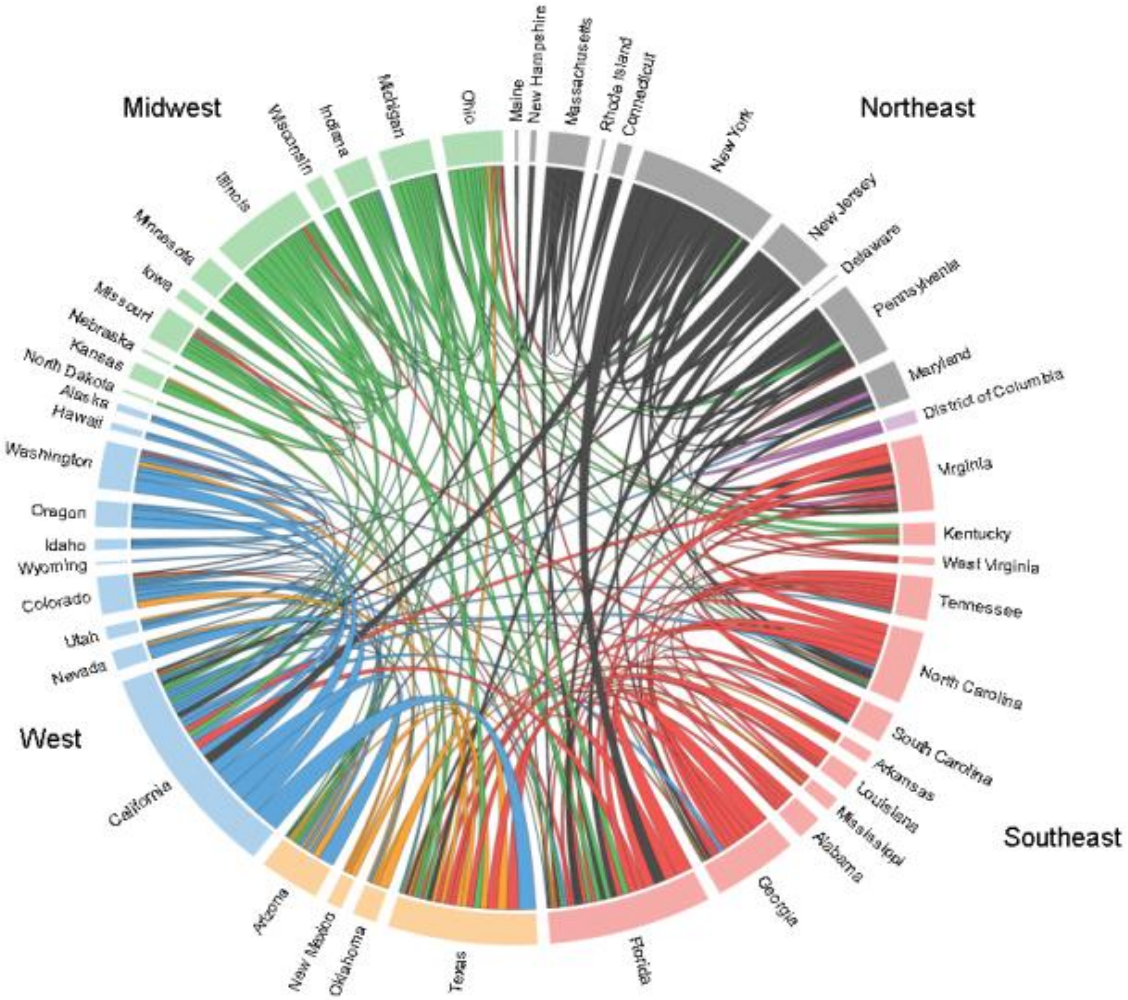


Fig (7) Showing Americans moving between states

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.

V. CONCLUSION

The analysis of big data requires traditional tools like SQL, analytical workbenches and data analysis and visualization languages like R. These tools can be used in various fields where data analytics is required. Many more tools have been introduced in the market and the existing products are also under constant improvement. The demand for better analytics tools is increasing constantly which is only going 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.

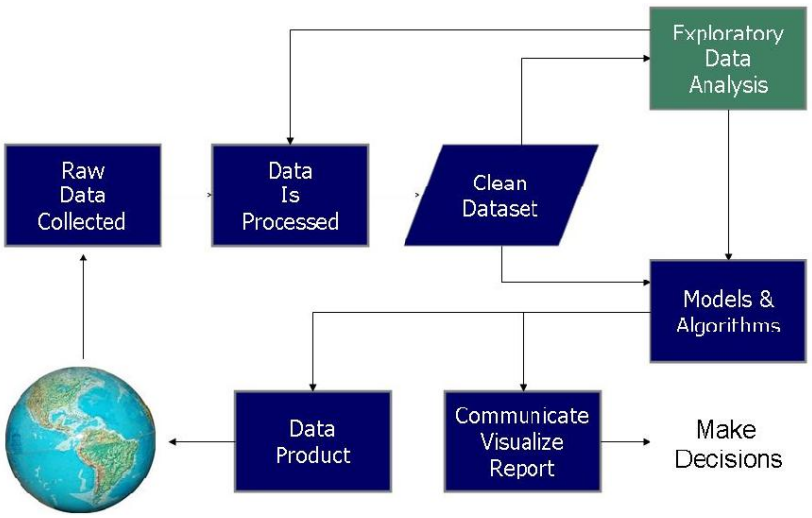


Fig (2) Data Science process

