WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# A Survey on Big Data Analytics Data Science Technologies

***Dr.S.Prakasam***

*Associate Professor, Department of Computer Science & Applications, SCSVMV University, Enathur,Kanchipuram Tamil Nadu, India*

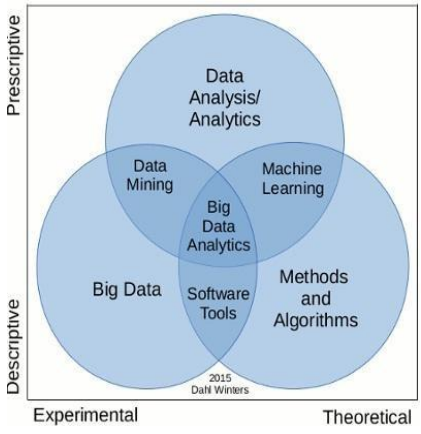
A B S T R A C T

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 Big data Analytics used in data science technologies

Keywords: analytics, data science data visualization, extraction, patterns evaluations

## 1. INDRODUCTION

. 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.

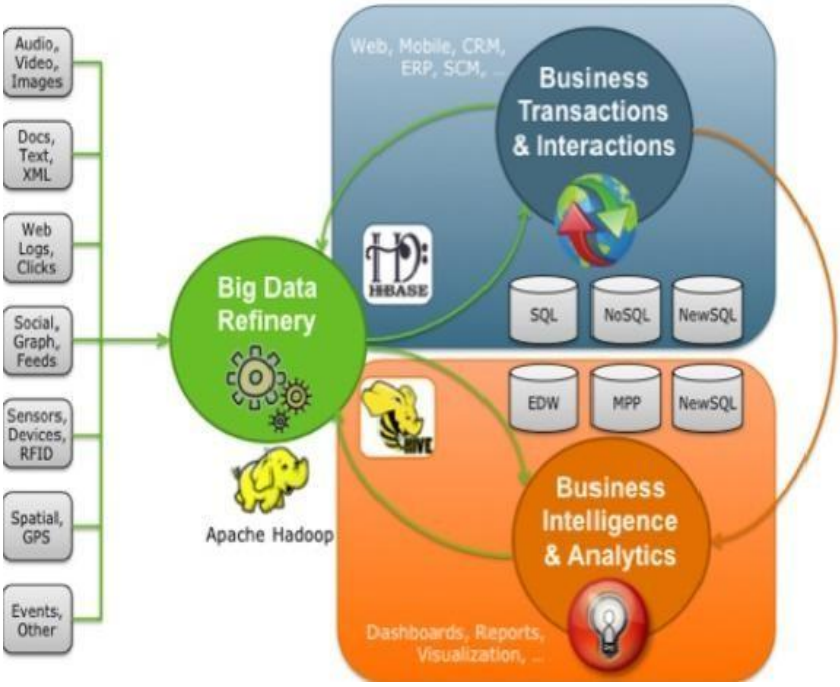



Fig(1) Fields of Data Science

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).

## 2. 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(2) Next generation Big data Architecture

.

**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

**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

## 3.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.

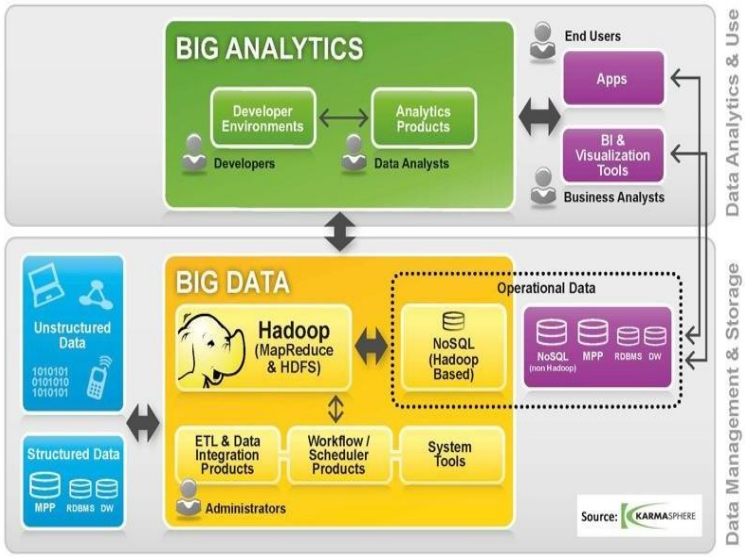
### 2) R

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(3) 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) 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.

**5) 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.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.

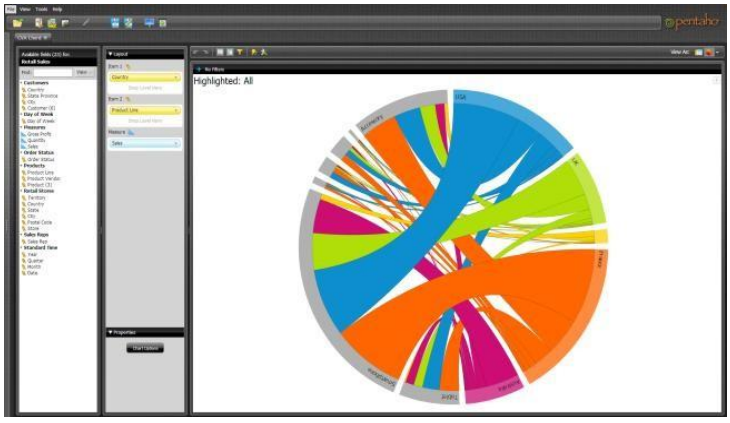



Fig (4): relationships between attributes and dimensions

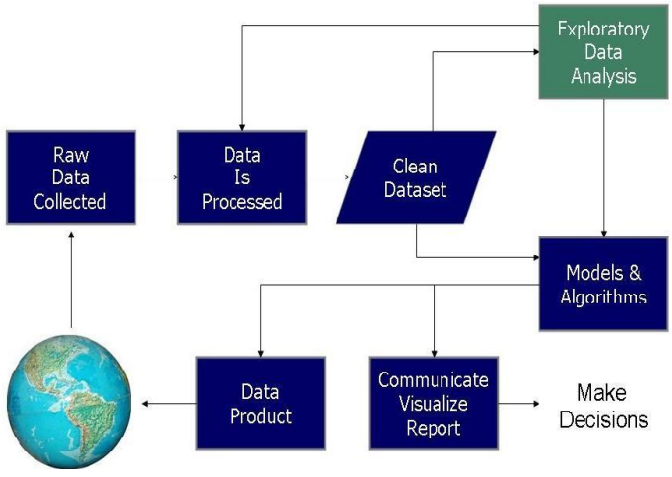



Fig (5) Data Science process

**Advantages of Big Data**

The importance of it leads to intense competition and increased demand for big data professionals. here we will discuss the advantages of it.

**Cost Savings** The implementation of Real-Time Analytics tools may be expensive, it will eventually save a lot of money. Some tools of it like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are to be stored and these tools also help in identifying more efficient ways of doing business.

**Time Reductions** The high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analysis immediately and make quick decisions based on the learnings.

**Better sales insights,** which could lead to additional revenue. Real-time analytics tell exactly how your sales are doing and in case an internet retailer sees that a product is doing extremely well, it can take action to prevent missing out or losing revenue.

**Control online reputation** tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, tools can help in all this.

**Understand the market conditions** By analysis you can get a better understanding of current market conditions. For example, by analyzing customers' purchasing behaviors, a company can find out the products that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.

**Increased productivity** Modern tools are allowing analysts to analyze more data, more quickly, which increases their personal productivity. In addition, the insights gained from that analytics often allow organizations to increase productivity more broadly throughout the company.

**Fraud detection** One of the big advantages of analytics systems that rely on machine learning is that they are excellent at detecting patterns and

anomalies. These abilities can give banks and credit card companies the ability to spot stolen credit cards or fraudulent purchases, often before the cardholder even knows that something is wrong.

**Applications of Big Data**

**1. Tracking Customer Spending Habit, Shopping Behavior:** In big retails store (like Amazon, Walmart, Big Bazar etc.) management team has to keep data of customer's spending habit (in which product customer spent, in which band they wish to spent, how frequently they spent), shopping behavior, customer's most liked product (so that they can keep those products in the store). Which product is being searched/sold most, based on that data, production/collection rate of that product get fixed. Banking sector uses their customer's spending behavior-related data so that they can provide the offer to a particular customer to buy his particular liked product by using bank's credit or debit card with discount or cash back. By this way, they can send the right offer to the right person at the right time.

**2. Recommendation:** By tracking customer spending habit, shopping behavior, Big retails store provide a recommendation to the customer. E-commerce site like Amazon, Walmart, Flipkart does product recommendation. They track what product a customer is searching, based on that data they recommend that type of product to that customer.

As an example, suppose any customer searched bed cover on Amazon. So, Amazon got data that customer may be interested to buy bed cover. Next time when that customer will go to any google page, advertisement of various bed covers will be seen. Thus, advertisement of the right product to the right customer can be sent.

YouTube also shows recommend video based on user's previous liked, watched video type. Based on the content of a video, the user is watching, relevant advertisement is shown during video running. As an example suppose someone watching a tutorial video of Big data, then advertisement of some other big data course will be shown during that video.

**3. Smart Traffic System:** Data about the condition of the traffic of different road, collected through camera kept beside the road, at entry and exit point of the city, GPS device placed in the vehicle (Ola, Uber cab, etc.). All such data are analyzed and jam-free or less jam way, less time taking ways are recommended. Such a way smart traffic system can be built in the city by Big data analysis. One more profit is fuel consumption can be reduced.

**4. Secure Air Traffic System:** At various places of flight (like propeller etc) sensors present. These sensors capture data like the speed of flight, moisture, temperature, other environmental condition. Based on such data analysis, an environmental parameter within flight are set up and varied.By analyzing flight's machine-generated data, it can be estimated how long the machine can operate flawlessly when it to be replaced/repaired.

**5. Auto Driving Car:** Big data analysis helps drive a car without human interpretation. In the various spot of car camera, a sensor placed, that gather data like the size of the surrounding car, obstacle, distance from those, etc. These data are being analyzed, then various calculation like how many angles to rotate, what should be speed, when to stop, etc carried out. These calculations help to take action automatically.

**6. Virtual Personal Assistant Tool:** Big data analysis helps virtual personal assistant tool (like Siri in Apple Device, Cortana in Windows, Google Assistant in Android) to provide the answer of the various question asked by users. This tool tracks the location of the user, their local time, season, other data related to question asked, etc. Analyzing all such data, it provides an answer.As an example, suppose one user asks "Do I need to take Umbrella?", the tool collects data like location of the user, season and weather condition at that location, then analyze these data to conclude if there is a chance of raining, then provide the answer.

**7. IoT:**

- Manufacturing company install IOT sensor into machines to collect operational data. Analyzing such data, it can be predicted how long machine will work without any problem when it requires repairing so that company can take action before the situation when machine facing a lot of issues or gets totally down. Thus, the cost to replace the whole machine can be saved.
- In the Healthcare field, Big data is providing a significant contribution. Using big data tool, data regarding patient experience is collected and is used by doctors to give better treatment. IoT device can sense a symptom of probable coming disease in the human body and prevent it from giving advance treatment.

  IoT Sensor placed near-patient; new-born baby constantly keeps track of various health conditions like heart bit rate, blood presser, etc. Whenever any parameter crosses the safe limit, an alarm sent to a doctor, so that they can take step remotely very soon.

**8. Education Sector:** Online educational course conducting organization utilizes big data to search candidate, interested in that course. If someone searches for YouTube tutorial video on a subject, then online or offline course provider organization on that subject send ad online to that person about their course.

**9. Energy Sector:** Smart electric meter read consumed power every 15 minutes and sends this read data to the server, where data analyzed and it can be estimated what is the time in a day when the power load is less throughout the city. By this system manufacturing unit or housekeeper are suggested the time when they should drive their heavy machine in the night time when power load less to enjoy less electricity bill.

**10. Media and Entertainment Sector:** Media and entertainment service providing company like Netflix, Amazon Prime, Spotify do analysis on data collected from their users. Data like what type of video, music users are watching, listening most, how long users are spending on site, etc are collected and analyzed to set the next business strategy.

## 5.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] H. Kitano, Artificial intelligence to win the Nobel prize and beyond: Creating the engine for scientific discovery, AI Magazine 37(1) (2016), 39–49. https://www.aaai.org/ojs/index.php/aimagazine/article/view/2642

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

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

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

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

[8] For Big Data Analytics There's No Such Thing as Too Big The Compelling Economics and Technology of Big Data Computing, White Paper, March 2012, By: 4syth.com, Emerging big data thought leaders

[9] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011 [11] Oracle Information Architecture: An Architect's Guide to Big Data, An Oracle White Paper in Enterprise Architecture August 2012.

[10] H. Moed, The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature, 2012, ResearchTrends, http://www.researchtrends.com.

[11] P. Zikipoulos, T. Deutsch, D. Deroos, Harness the Power of Big Data, 2012, http://www.ibmbigdatahub.com/blog/harnes s-power-big-data-book-excerpt.

[12] E. Dumhill, "What is big data?", 2012 ,http://strata.oreilly.com/2012/01/what-isbig-data.html.