International Journal of Research in Engineering, Science and Management
Volume-1, Issue-11, November-2018
www.ijresm.com | ISSN (Online): 2581-5792

Data Science and Analytics

Aman Rathore
Student, Department of Computer Science Engineering, AITR, Indore, India

***Abstract*: with the collection, preparation, analysis, visualization, management, and preservation of large collections of information. Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining Although the name Data Science seems to connect most strongly with areas such as databases and computer science, many different kinds of skills including nonmathematical skills are also needed here. Data Science is much more than simply analysing data. There are many people who enjoy analysing data who could happily spend all day looking at histograms and averages, but for those who prefer other activities, data science offers a range of roles and requires a range of skills. Data science includes data analysis as an important component of the skill set required for many jobs in the area, but is not the only skill. Data scientists play active roles in the design and implementation work of four related areas such as data architecture, data acquisition, data analysis and data archiving. In the present paper the authors will try to explore the different issues, implementation and challenges in area called Data science.**

***Keywords*: data science**

1. Introduction

Imagine a world without data storage; a place where every detail about a person or organization, every transaction performed, or every aspect which can be documented is lost directly after use. Organizations would thus lose the ability to extract valuable information and knowledge, perform detailed analyses, as well as provide new opportunities and advantages. Anything ranging from customer names and addresses, to products available, to purchases made, to employees hired, etc. has become essential for day-to-day continuity. Data is the building block upon which any organization thrives. "Unstructured data" can incorporate messages, features, photographs, online networking, and other client produced substance. Information science frequently obliges dealing with an awesome measure of data and composing calculations to concentrate bits of knowledge from this information". The field of data science uses information planning, insights, and machine learning to research issues in different spaces, and for example, advertising improvement, extortion discovery, setting open strategy, and so forth. Data science researchers utilize the capacity to discover and translate rich information sources; oversee a lot of information notwithstanding equipment, programming, and transfer speed imperatives; consolidate information sources; guarantee consistency of datasets; make representations to help in comprehension of information; construct scientific models utilizing the information; and display and impart the information experiences/discoveries. The basic flow control in a data science process can be summed up in the Fig. 1.

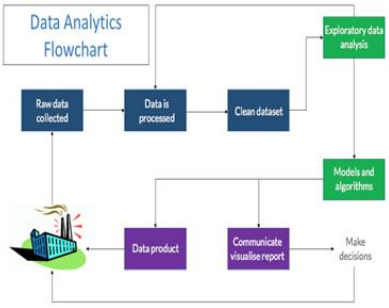

Fig. 1. Data science process

2. Steps of data science

The three segments included in data science are arranging, bundling and conveying information (the ABC of information). However bundling is an integral part of data wrangling, which includes collection and sorting of data. However what isolates data science from other existing disciplines is that they additionally need to have a nonstop consciousness of What, How, Who and Why. A data science researcher needs to realize what will be the yield of the data science transform and have an unmistakable vision of this yield. A data science researcher needs to have a plainly characterized arrangement on in what manner this yield will be accomplished inside of the limitations of accessible assets and time. A data scientist needs to profoundly comprehend who the individuals are that will be included in making the yield. The steps of data science are mainly: collection and preparation of the data, alternating between running the analysis and reflection to interpret the outputs, and finally dissemination of results in the form of written reports and/or executable code. The following are the basic steps involved in data science.

A. Data wrangling and munging

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. A data wrangler is a person who performs these transformation operations. This may include further munging, data visualization, data aggregation, training a

statistical model, as well as many other potential uses. Data munging as a process typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data using algorithms (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.

B. Data analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

C. Convey data

Conveying data includes methods to transform the mathematical or statistical conclusions drawn from the data into a form that can be easily understood and interpreted by those in need of it. Conveying data is empowering the development starting with one perspective then onto the next, empowering a beginner to turn into an expert, current technology to appear to be new and allowing the modeled information to be seen by apprentices and making new technology to appear like it was an integral part of the system.

3. Evolution of data science

In 1962, John Tukey wrote about a shift in the world of statistics, saying, "… as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt…I have come to feel that my central interest is in data analysis" Tukey is referring to the merging of statistics and computers, at a time when statistical results were presented in hours, rather than the days or weeks it would take if done by hand.

In 1974, Peter Naur authored the Concise Survey of Computer Methods, using the term "Data Science," repeatedly. Naur presented his own convoluted definition of the new concept:

"The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

In 1977, The IASC, also known as the International Association for Statistical Computing was formed. The first phrase of their mission statement reads, "It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."

In 1977, Tukey wrote a second paper, titled Exploratory Data Analysis, arguing the importance of using data in selecting "which" hypotheses to test, and that confirmatory data analysis and exploratory data analysis should work hand-in-hand.

In 1989, the Knowledge Discovery in Databases, which would mature into the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, organized its first workshop.

In 1994, Business Week ran the cover story, Database Marketing, revealing the ominous news companies had started gathering large amounts of personal information, with plans to start strange new marketing campaigns. The flood of data was, at best, confusing to company managers, who were trying to decide what to do with so much disconnected information.

In 1999, Jacob Zahavi pointed out the need for new tools to handle the massive amounts of information available to businesses, in Mining Data for Nuggets of Knowledge. He wrote:

"Scalability is a huge issue in data mining… Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data… Another technical challenge is developing models that can do a better job analyzing data, detecting non-linear relationships and interaction between elements… Special data mining tools may have to be developed to address web-site decisions."

In 2001, Software-as-a-Service (SaaS) was created. This was the pre-cursor to using Cloud-based applications.

In 2001, William S. Cleveland laid out plans for training Data Scientists to meet the needs of the future. He presented an action plan titled, Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics. It described how to increase the technical experience and range of data analysts and specified six areas of study for university departments. It promoted developing specific resources for research in each of the six areas. His plan also applies to government and corporate research.

In 2002, the International Council for Science: Committee on Data for Science and Technology began publishing the Data Science Journal, a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues.

In 2006, Hadoop 0.1.0, an open-source, non-relational database, was released. Hadoop was based on Nutch, another open-source database.

In 2008, the title, "Data Scientist" became a buzzword, and eventually a part of the language. DJ Patil and Jeff Hammerbacher, of LinkedIn and Facebook, are given credit for initiating its use as a buzzword.

In 2009, the term NoSQL was reintroduced (a variation had been used since 1998) by Johan Oskarsson, when he organized a discussion on "open-source, non-relational databases".

In 2011, job listings for Data Scientists increased by 15,000%. There was also an increase in seminars and conferences devoted specifically to Data Science and Big Data. Data Science had proven itself to be a source of profits and had become a part of corporate culture.

In 2011, James Dixon, CTO of Pentaho promoted the concept of Data Lakes, rather than Data Warehouses. Dixon stated the difference between a Data Warehouse and a Data Lake is that

the Data Warehouse pre-categorizes the data at the point of entry, wasting time and energy, while a Data Lake accepts the information using a non-relational database (NoSQL) and does not categorize the data, but simply stores it.

In 2013, IBM shared statistics showing 90% of the data in the world had been created within the last two years. In 2015, using Deep Learning techniques, Google's speech recognition, Google Voice, experienced a dramatic performance jump of 49 percent.

In 2015, Bloomberg's Jack Clark, wrote that it had been a landmark year for Artificial Intelligence (AI). Within Google, the total of software projects using AI increased from "sporadic usage" to more than 2,700 projects over the year.

In the past ten years, Data Science has quietly grown to include businesses and organizations world-wide. It is now being used by governments, geneticists, engineers, and even astronomers. During its evolution, Data Science's use of Big Data was not simply a "scaling up" of the data, but included shifting to new systems for processing data and the ways data gets studied and analyzed.

4. Conclusion and future scope

Data analytics is a process through which data is cleaned, analyzed and modelled using tools. This data is then used to derive insights. The insights are then used for business related decision-making purposes. There are many techniques that data analysts use in different fields of work. In the world of business, Data analytics is used for making strategies to get the desired business results. Today, data analytics has become a big career option in India. As a result, big data analytics courses are in huge demand.

Businesses have realized the importance of utilizing big data analytics to maximize their profits. They know that it is vital for their growth and for the future health of their business. Today, major business decisions are taken by utilizing the insights derived from data related to the organization or industry related data. As competition increases and customers are flooded with choices, it has become important to move faster in the market and that too with accuracy.

Data analytics provides both speed and accuracy to business decisions. It provides accuracy as it is based on statistical models and hi-tech tools that help fine tuning and analysing the data. This field also provides answers to present business problems as well as give a view of future trends. It is preparing the companies to make products for the future and aspire to connect with the customers of tomorrow.

As data analytics also allows to improve business process and maximise conversion rates, it helps the organizations in cutting unnecessary costs and reduce the cost of running the company. With all its obvious benefits, it is quite natural to say that data analytics is going to become important in a big economy like India.

India is a popular destination for a lot of companies who outsource their work to other countries. This is due to the lower cost of operations and manpower in India. This is further aided by the skilled and English-speaking youth of India. Data analytics is one such field where outsourced opportunities are available in India. As a country teeming with young people and tremendous outsourced work coming in, the scope for this sector is big in India.

Today, as advancements in the field of data analytics are being made, the process is getting automated. Machines are analyzing big chunks of data in an automated process. With more and more smart machines entering our daily lives, more and more data is getting created every hour. All this data can be used and analyzed for understanding customer behaviour or predicting future trends. With the help of machines, data analysts are finding it possible to make sense of the data in a quicker and easier way.

References

[1] Dhar, V. (2013). "Data science and prediction". Communications of the ACM 56.
[2] Jeff Leek (2013-12-12). "The key word in 'Data Science' is not Data, it is Science". Simply Statistics.
[3] Hal Varian on how the Web challenges managers. http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers
[4] Parsons, MA, MJ Brodzik, and NJ Rutter. 2004. Data management for the cold land processes experiment: improving hydrological science. Hydrol Process. 18:3637-653.
[5] Tukey, John W. The Future of Data Analysis. Ann. Math. Statist. 33 (1962), no. 1, 1-67.
[6] Tukey, John W. (1977). Exploratory Data Analysis. Pearson. ISBN 978-0201076165.
[7] Peter Naur: Concise Survey of Computer Methods, 397 p. Student literature, Lund, Sweden, 1974.
[8] KDD-89: IJCAI-89 Workshop on Knowledge Discovery in Databases. August 20, 1989, Detroit MI, USA
[9] Database Marketing Business Week. September 04, 1994
[10] From Data Mining to Knowledge Discovery in Databases. Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. AI Magazine Volume 17 Number 3 (1996)
[11] "Statistics=Data Science?" C.F.Jeff Wu. University of Michigan, Ann Arbor.
[12] Data Mining and Knowledge Discovery.
[13] Mining Data for Nuggets of Knowledge Dec 10, 1999 Mining Data for Nuggets of Knowledge. http://knowledge.wharton.upenn.edu/article/mining-datafor-nuggets-of-knowledge/
[14] Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics, William S. Cleveland, 2001.
[15] The International Council for Science: Committee on Data for Science and Technology in 2002.
[16] CTO of Pentaho promoted the concept of Data Lakes, rather than Data Warehouses James Dixon in 2011.