DATA SCIENCE - DOMAIN , TOOLS AND TECHNOLOGIES: A CASE STUDY

Juhi singh
Amity University Haryana

ABSTRACT: Data science is the investigation of the generalizable extraction of learning from information. It incorporates an assortment of parts and creates on techniques and ideas from numerous spaces, containing science, likelihood models, machine learning, factual learning, PC programming, information building, design acknowledgment and learning, perception and data warehousing planning to extricate an incentive from information. The motivation behind this paper is to give an outline Data science devices, proposing a grouping plan that can be utilized to contemplate open source information science programming. In parallel, this examination gives a diagram of existing center and other related areas of information science instruments.

Keywords: data science, data analysis tools

1. Introduction

Data science is an interdisciplinary field of logical techniques, procedures, calculations and frameworks to extricate learning or experiences from information in different structures, either organized or unstructured, like Data mining[1, 9] .Data science is an "idea to bring together insights, information examination, machine learning and their related strategies" to "comprehend and break down genuine wonders" with information. It utilizes systems and hypotheses drawn from numerous fields inside the wide territories of arithmetic, measurements, data science, and software engineering. Data science innovation ought not be mistaken for the data science.

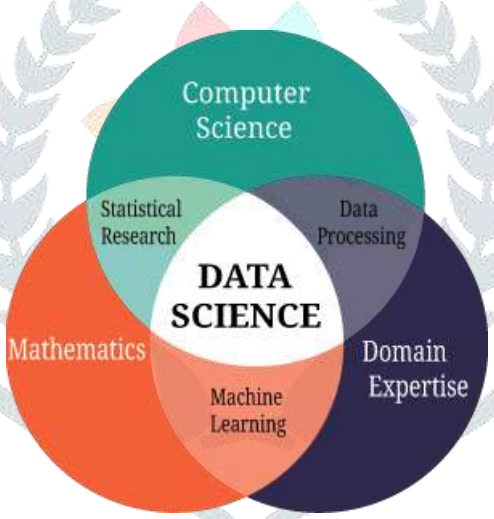

Figure 1: Data Science and its field [2]

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge[3].

Data science use some tools and technology to analyze the data for the extraction of knowledge or information for future use which helps in prediction and forecasting and decision making .As machine learning is the some of the core area of data science, which also helps to train or to learn from a previous data-set by using deep leaning . So indirectly it is to be said that data science can also helps in deep learning as machine learning [10]. Here are some tools and technology are mentioned in fig:2 with their percentage of usage, which is based on some survey

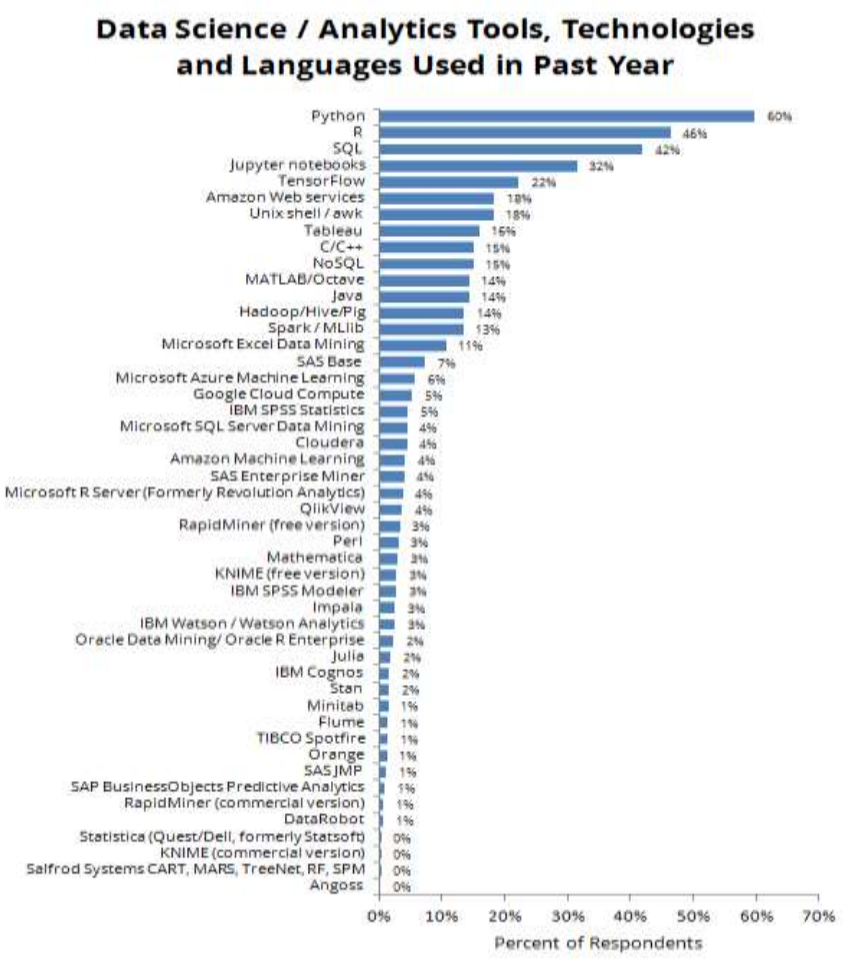

Figure 2: Data science analytical tools, techniques and languages [4]

2. DATA SCIENCE TECHNIQUES AND TOOLS

Data science is a multidisciplinary logical approach. A fitting arranged approach is to be take after for applying the systems of Data science while managing diverse kinds of information. The strategies that are utilized as a part of information science are gotten from different fields, for example, Artificial Intelligence, Big information, arithmetic, data and correspondence innovations and measurements [7].

Table 1: Tools of data science[5]

S.No.	Tools Name	Description
1.	Hadoop	Hadoop is an open source tool that is based on them Java-programming framework. Hadoop works on the huge amount of data to perform processing and storage tasks. Hadoop immediately developed as an establishment for big data processing tasks, for example business planning and scientific analytics.
2.	Map Reduce	MapReduce is a two step process in the first step is Map which collect the dataset and performs filtering and sorting operations on it. And in the second step, Reduce accept the output of Map as an input and then consolidate those data sets into smaller datasets. MapReduce libraries have been composed in many programming languages, with various levels of optimization.
3.	NoSQL	NoSQL databases are progressively utilized as a part of big data applications. NoSQL frameworks are likewise infrequently called "Not only SQL" to highlight that they may support SQL-like query languages. It provides a platform for storage

		and retrieval of data
4.	HBase	HBase is a data model which provide quick random access to large amounts of organized data. It is a part of the Hadoop environment that gives arbitrary ongoing read/compose access to information in the Hadoop File System
5.	R	R is a statistical computing and graphics environment based language. R provides the facilities for data manipulation, storage, performing calculation and graphical display. The R language is utilized among analysts and information mineworkers for creating statistical software and to perform the analysis on data.
6.	SQL	SQL remains for Structured Query Language which enables clients to get the information from database management systems. SQL is a language to perform storing, manipulating and retrieving operations on data in databases
7.	Weka	Weka stands for Waikato Environment for Knowledge Analysis [6]. It is based on the machine learning algorithms that is used to perform mining task from given datasets. It consists set of tools to perform tasks like data pre-processing, clustering, association, classification, regression and visualization.
8.	Rapid Miner	Rapid Miner is a data science software platform which provides an integrated environment used for business applications. It is an open source platform where analytics tasks can be performed. It also supports the concept of machine learning, deep learning and text mining
9.	Python	Python is an interactive, object-oriented, and high-level programming language. It provides interfaces to all real business databases. Python can be utilized as a scripting language and it can be ordered to byte-code for building huge applications.
10.	D3.js	D3 refers to 3 D's as Data, Driven and Documents. To create an interactive and dynamic representation of the web pages, it used JavaScript library. D3.js provides an environment to attach the data to DOM (Document Object Model) components.

3. Applications of Data Science

Data science is a subject that rose fundamentally from require, with respect to authentic applications instead of as an exploration area. Consistently, it has progressed from being used as a piece of as far as possible field of bits of knowledge and examination to being a general proximity in each part of science and industry. In this fragment, we look at a segment of the essential zones of employments and research where data science is starting at now used and is at the bleeding edge of advancement**8]**. Following are the some core application areas of Data science:

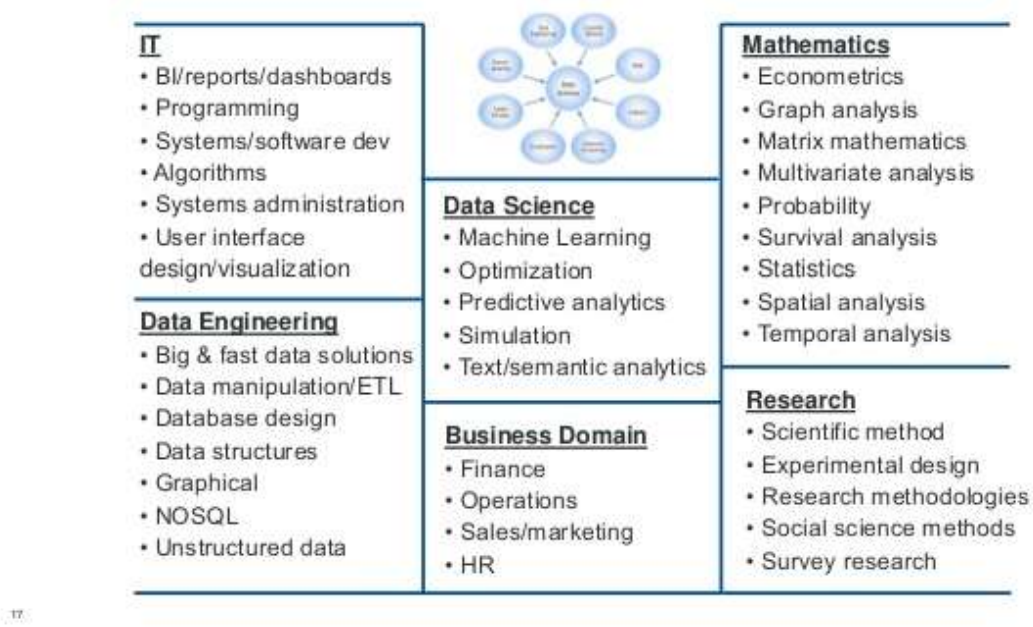

Figure 5: core application areas of data science[2]

Apart from the above mentioned areas of data science, it can be used in other domain also

- Business Analytics
- Prediction
- Security
- Computer Vision
- Natural Language Processing
- Bioinformatics
- Science and Research
- Revenue Management

4. Conclusion

Data science is an approach that incorporates gathering, readiness, administration, investigation and perception on colossal and complex datasets. to accomplish the objective, different ideas are combined like insights, software engineering, science, area information and perception. The computerization instruments are utilized by the information researcher to separate the new learning results, which increment the estimation of business. Everything in science is changing a result of the effect of Information Technology. Data science is a rising field, with awesome vulnerability, fast changes, and energizing openings. Information science is tied in with overseeing enormous nature of information with the ultimate objective of extricating important and reliable results. Data science isn't just manages the instruments and techniques to discover, oversee and dissect information, it is additionally about acquiring an incentive from information and making a translation of it from advantage for learning.

References:

1. https://en.wikipedia.org/wiki/Data_science

2. https://sites.google.com/site/yongyoondatascience/

3. https://www.artificial-intelligence.blog/terminology/data-science

4. https://www.kaggle.cm/surveys/2017

5. Neha Bhateja,"A business approach to Data Mining for software Development Process " , International Journal of Advanced Research in Computer Science, 8 (5), May-June 2017

6. Juhi Singh, "A comprehensive study on Open Source Tools and Techniques of Data Mining ", International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES) Volume 4, Issue 5, May-2018

7. Chalef, Daniel "Data Science Tools – Are Proprietary Vendors Still Relevant?". kdnuggets.com. Retrieved 2016-11- 07

8. Nishu Sethi," A comprehensive study on Data Science and Its Practical Applications", International Journal on Recent and Innovation Trends in Computing and Communication , Volume: 5 Issue: 5 ,pp 573 – 575

9. Nishu Sethi, Neha Bhateja, "?various Techniques of Data Mining-For Software Development Process", International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES) Volume 4, Issue 6,June-2018

10. Juhi Singh, "comprehensive study of deep learning", International Journal on Recent and Innovation Trends in Computing and Communication , Volume: 5 Issue: 5 ,pp 641-645