International Journal of Research Publication and Reviews

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

A Review on Data Science Technologies

Mr. Varun M R[1], *Prof. Hetal Rana*[2]

[1]MTech (AI and DS) Student, MVJ College Of Engineering, Bangalore Email: vamr20ee@cmrit.ac.in
[2]Assistant Professor, MVJ College Of Engineering, Near ITPL Bangalore Email: hetalrana@mvjce.edu.in

ABSTRACT

Data science plays a crucial role in handling vast amounts of data to derive meaningful and coherent insights. This emerging field encompasses various activities such as data mining and analysis, utilizing techniques from mathematics, statistics, information technology, computer programming, data engineering, pattern recognition and learning, visualization, and high-performance computing. This paper provides a clear overview of different data science technologies.

Keyword: Data Science, Analytics, Data Visualization, Extraction, Patterns,

1. INTRODUCTION:

Data science is all about extracting insights from data through analysis. It involves understanding the needs of businesses and finding solutions to their problems by studying and modelling data. This field encompasses various techniques for analysing, storing, and presenting data effectively.

Data science indeed involves the integration of computer science, statistics, data storage, and perception to manage, store, and analyse data. It's a relatively new field, so there isn't a consensus on its exact boundaries. Data science combines mathematics, statistics, programming, and problem-solving contexts, enabling the capture and analysis of data in new ways.

The ability to see things from a different perspective and importantly, the significant and foundational task of cleaning, preparing, and organizing data. The actual process of Data Science is illustrated in Figure 1.

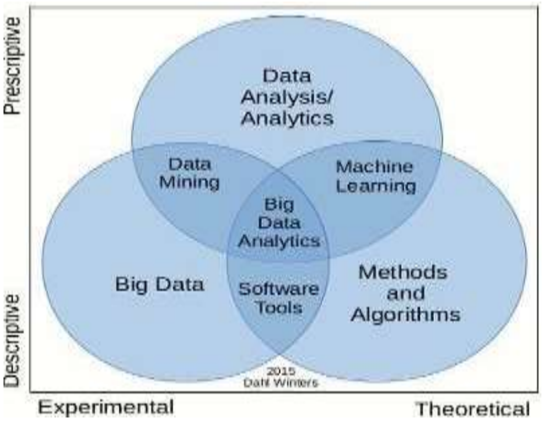

Figure1. Fields of Data Science

Data science has garnered significant attention in academic and industrial circles, leading to the establishment of new research institution and organization,

For example, the Columbia University Institute for Data Sciences and Engineering and New York University Centre for Data Science have emerged as prominent players in this field. Several universities, including the University of California at Berkeley, Columbia University, and Fudan University, among others, have launched data science courses and degree programs.

Cleveland and Smith proposed that data science should be considered a distinct discipline. Major companies such as Facebook, Google, EMC, and IBM have created job roles for data scientists. Harvard Business Review has described the data scientist as "the sexiest job of the 21st century."

Despite the growing interest, there is no universally accepted definition of data science. It is seen as a new and evolving field with research objectives and scientific challenges that differ from more established branches of science. The problems addressed by data science are often distinct from those tackled by traditional or social sciences.

2. TOOLS OF DATA SCIENCE TECHNOLOGIES

A. R Programming

R is a programming language designed for statistical computing and data analysis. It is widely used by data scientists and analysts for high-dimensional data visualization and analysis. R's popularity has grown significantly in recent years, as evidenced by various studies and surveys.

Originally developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, R is named after the first initials of its creators. It is freely available under the GNU General Public License and its source code is primarily written in FORTRAN, C, and R. R is a GNU project, and pre-compiled binary versions are available for various operating systems.

R provides a command-line interface (CLI), but there are also several graphical user interfaces (GUIs) available. It offers a wide range of statistical techniques, from modelling to clustering, classification, and summarization. The packages developed by the R community play a crucial role in extending its functionality.

While R is primarily used for statistical computing, developers proficient in languages like C, C++, Java, .NET, or Python can write their own code to interact with R objects. Advanced users can also leverage their preferred computational tools for computationally intensive tasks.

In addition to statistical analysis, R also provides graphical packages for creating flexible, informative, and publication-quality visualizations. These capabilities make R a powerful tool for data miners and analysts alike.

B. Python

Python is a versatile, open-source programming language known for its simplicity and versatility. It features a straightforward syntax that is easy to learn and use, making it accessible to beginners while also being powerful enough for advanced users. Python is equipped with robust libraries for data manipulation and analysis, making it a popular choice for scientific computing and quantitative fields such as finance, oil and gas exploration, physics, and signal processing.

Python's versatility is evident in its use across various industries and projects. For example, it has been utilized in optimizing Space Shuttle mission designs, processing images captured by the Hubble Space Telescope, and playing a crucial role in the physics experiments that led to the discovery of the Higgs Boson, often referred to as the "God particle."

Overall, Python's combination of ease of use, powerful libraries, and wide-ranging applications has made it a preferred language for developers across different domains.

Python's popularity, as shown by the TIOBE index, surpasses that of Perl, Ruby, and JavaScript by a significant margin. Its agility and productivity in developing solutions are highly regarded among modern languages. Python's future hinges on the adoption of Python SDKs by service providers and the continued expansion of Python modules to enrich the ecosystem of Python applications.

C. Hadoop

Hadoop has become synonymous with handling large datasets. It is an open-source software framework for distributed storage and processing of big data across computer clusters. This means you can easily scale your data storage and processing capabilities without worrying about hardware failures. Hadoop offers massive storage capacity for any type of data, significant processing power, and the ability to handle virtually unlimited simultaneous tasks or jobs. However, Hadoop is not beginner-friendly. To fully harness its power, a good understanding of Java is essential. Despite the learning curve, Hadoop is highly valuable because many other companies and technologies rely on it or integrate with it. Nevertheless, Hadoop's MapReduce framework is designed for batch processing and may not be suitable for interactive applications, real-time operations like stream processing, or more complex algorithms.

D. Visualization Tools

Data visualization is an essential aspect of data analysis, helping to transform complex data into understandable and visually appealing representations. Here are some tools commonly used for data visualization:

1. **Tableau**: Tableau is a powerful tool that offers a user-friendly interface for creating interactive and visually appealing data visualizations. It allows users to analyse and present data without requiring advanced programming skills.

2. **D3.js**: D3.js is a JavaScript library that provides powerful capabilities for creating custom data visualizations on the web. It allows developers to bind data to DOM elements and apply data-driven transformations to create dynamic and interactive visualizations.

3. Data wrapper: is a web-based tool that simplifies the process of creating charts and maps. It offers a range of customizable templates and allows users to quickly upload data, choose a visualization type, and publish the final visualization.

These tools offer different approaches to data visualization, catering to a wide range of user needs and preferences. Whether you prefer a user-friendly interface or the flexibility to create custom visualizations, there is a tool that can meet your requirements.

3. CONCLUSION

Data science relies on fundamental tools such as SQL, analytical workbenches, and data analysis and visualization languages like R. These tools are essential across various industries where data analysis is crucial. The market is continuously evolving, with new and improved tools regularly emerging. There is a growing demand for advanced analytics tools, which is expected to increase even further in the future.

4. REFERENCES

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

[2]. D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, 2012, "Interactions with big data analytics," interactions, Vol. 19, No. 3, PP. 50–59, May 2012

[3]. Eckerson, W. 2011. "Bigdata Analytics: Profiling the Use of Analytical Platforms in User Organizations," TDWI.

[4]. Lekha R. Nair and Sujala D. Shetty, "Research in Big Data and Analytics: An Overview," International Journal of Computer Applications, ISSN NO:0975-8887Vol.