ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

CASE STUDY AND REPORT

Available Online at www.ijarcs.info

A Business Approach to Data Driven Science - A Study

Neha Bhateja
Department of Computer Science & Engineering
Amity University Haryana, India

Abstract: Data science provides a platform to extract useful information and derived new knowledge by performing the analysis on the huge data. It is a recently rising field that includes various activities, like data mining and data analysis. It utilizes the concept of mathematics, statistics, and information technology, pattern identification and data representation skills which provides high performance computing. The data scientists that are involved in extracting useful information can increase the value to the business by effectively used their skills.

Keywords: Data Science, Data Scientists, Data Analytics, Data Mining

INTRODUCTION

Data science includes two terms i.e. data and science. The data refers to the collection of unprocessed material (data) in terms of numbers and characters. Whereas, science is a study which is related to describing and discovering the things by experimenting and doing observations. So, data science is the discipline that deals with collecting, preparing, managing, analyzing, interpreting and visualizing large and complex datasets [1]. Data science is a developing interdisciplinary field that merge elements of mathematics, computer science, statistics and knowledge in a specific application space with a goal to find useful information from the increasingly large amount of data.

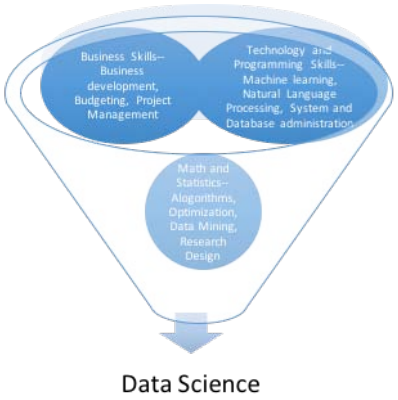

Fig: Skill Sets of data science [1]

DATA SCIENCE WORKFLOW

The means of information science are for the most part: accumulation and planning of the data, substituting between running the investigation and reflection to extract the output, and dispersal of results in terms of written documents [2].

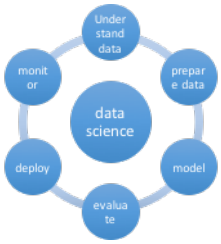

Fig: Data Science Process

firstly, the Data acquisition and cleanup process can be performed on large amount of available data by using Python libraries and other tools like open Refine and Wrangler. After that the data storage and management process can be performed by using NoSQL and MapReduce tools [4]. To extract the perception from the data visualization process is performed. And at last, the results are shared and distributed among the users.

DATA SCIENTIST- SKILLS

Data Scientist are those that are involved in the data driven approach of data science [3,6]. They can add value to overall business by their effective skills. Following are the skills that a data scientist must have.

Skills of a Data scientist

- Learning Abilities: the information researcher should have the ability to rapidly figure out how the information will be utilized in a specific manner.
- Communicating with the users: a data scientist must have abilities to learning the requirements and the inclination of clients. the data scientist has the ability to translate the process well in technical terms to the end users.
- Understanding of a complex system: after understanding the concept of application area, the data scientist must visualize how data will be used to different parts of the system and to the users.
- Knowing how data can be stored: data scientist must have a reasonable comprehension about how information can be put away and connected and having

knowledge about the metadata. Data scientist must have a reasonable comprehension about how information can be put away and connected and having knowledge about the metadata.

• Data analysis: at the point when information is accessible to the client for decision making, data researcher must know how to, extract the conclusions and doing review from the data.

• Visualization and presentation: a better data display can frequently be a more powerful method for imparting results to information clients. A better data display can frequently be a more powerful method for imparting results to information clients.

• Attention to quality: data scientist must know the impediments of the information they used and also having ability to know how to measure its precision, and have the capacity to make recommendations for enhancing the nature of the information in future.

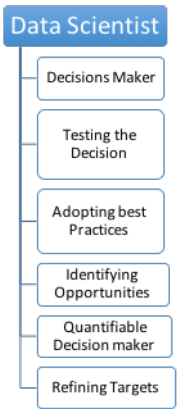

Fig: Ways Data Scientist Can Add value to Business

DATA SCIENCE TECHNIQUES AND TOOLS

Data science is a multidisciplinary scientific approach. An appropriate planned approach is to be follow for applying the techniques of data science while dealing with different types of data.

The techniques that are used in data science are derived from various fields such as Artificial Intelligence, Big data, mathematics, information and communication technologies and statistics [4].

Fig: Elements of Data Science

Tools used in data science:

• 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.

• 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.

• 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.

• 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.

• MapReduce: There are two steps of MapReduce. The first step is Map which collect the dataset and performs filtering and sorting operations on it. 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.

• 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.

• Weka: Weka stands for Waikato Environment for Knowledge Analysis [5]. 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.

• RapidMiner: 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

• 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.

• 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.

CONCLUSION

Data science is an approach that includes collection, preparation, management, analysis and visualization on huge and complex datasets. to achieve the target, various concepts are merged like statistics, computer science, mathematics, domain knowledge and visualization. The automation tools are used by the data scientist to extract the new learning outcomes, which increase the value of business. Everything in science is changing because of the impact of Information

Technology. Data science is an emerging field, with great uncertainty, rapid changes, and exciting opportunities.

REFERENCES

[1]. What is Data Science? http://www.datascientists.net/what-is-data-science

[2]. Dhar, V. (2013). "Data science and prediction". Communications of the ACM 56.

[3]. Richard Rivera Adam Haverson (2014) Data Scientist vs Data Analyst, Available at: https://www.captechconsulting.com/blogs/data- scientist-vs-data-analyst (Accessed: 14th May 2015).

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

[5]. G. Holmes; A. Donkin; I.H. Written (1994). "Weka: A machine learning workbench" (PDF). Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.

[6]. Nath," Challenges in Data Science: A Comprehensive Study on Application and Future Trends", IJARCSMS, ISSN: 232 7782, pg. 1-8, Volume 3, Issue 8, August 2015.