International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT2063136

Survey on Data Science and Data Scientist

Ashwin Karanth*[1], Prakash Hongal[2]
[1]Department of Computer Science and Engineering, SKSVMACET, Laxmeshwar, Karnataka, India
[2]Assistant Professor Department of Computer Science and Engineering, SKSVMACET, Laxmeshwar, Karnataka, India

ABSTRACT

Data science is a study of extracting, collecting, gathering of the data, representation and protecting the data to be used for business purpose or for other usage like technical and other software related platform. Data science is a combination of database, software, different types of quantitative aptitudes, qualitative aptitudes and non-mathematical abilities. Data Science is study of flow of the information from extremely large amount of data present in an organization. This paper illustrates what is data science, its application and issues which has to be removed. Next session of this paper consists of different analysis regarding to Data Science. Next session of this paper illustrates whole process of Data Science and all the related data research issues for the data science. At the end of the paper author will attempt to investigate the issues, execution and difficulties in data science. We will also look over some possible suggestion which can be considered for further research.

Keywords : Data Scientist, Data Science, Management, Cloud Computing.

I. INTRODUCTION

Data science is the collection of large amounts of data that are merged and processed [1]. In other words, data science is a detailed study of flow of data from large amount of information present in an organization. It involves obtaining meaningful data from raw and unstructured information which is processed through various skills such as analytical, programming and business. Data science uses machine learning algorithm to solve the problem.

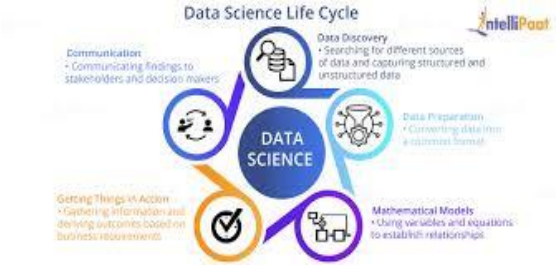

Fig 1: Data science life cycle

In the modern world where there is increase in digital technology, organization deals with huge amount of data daily which is structured and unstructured [1]. Develop in technology have enabled cost efficient and smarter storage space to store this huge data. Currently there is huge demand for certified data scientist who is highly skilled. Now they are one of the highest paid professionals in IT industry [1]. It is dream of many software engineers to be a data scientist, but it demands high skill and dedication to master it.

Components of data science are as follows:

1. **Data exploration**: It is main step a this consumes the most amount of time. When the Data Scientist it is in unstructured form, it has to be converted to structured form in order to do the further operation.

Here, unwanted data that is not required is removed and also it is checked whether any data is missing or not. Hence it is the most important step.

Modelling: By this step data is ready to be used for further process. In this step we actually use machine learning algorithms. The selection of the model and algorithm is depending on the data scientist whichever is relevant.

Testing of the model: In this step we check the performance of the model. The model is checked with the test data for the performance report. And make the required changes which is need to get the desired model.

Use of the models: Once we get the desired model after all the step, we can finally deploy the model to the production.

As any other type data science also have some application and issues in it. However, it is not considered as main issue but it shouldn't be neglected to be a data science expert. We will discuss about both in upcoming session.

II. LITERATURE REVIEW

Dr. S. Justus (2013), outlined that capacity, recovery and procedure for big data are advancing day by day rapidly. Testing is not facing any difficulties in this situation. J. Novling (2014), said that generating a lot of data for testing a data is a vital benchmark and quality manner in current machine learning. As maximum amount of data was stored electronically in same place has become a data science. Around 50 years ago John Tucky evolved this as Data analysis. Later after 30 years data analysis further evolved into study called Data Science by John Chambers and Leo Breiman(donoho,2015). When the term Data Scientist was coined by DJ Patil and Jeff Hammerbacker in 2009, there is a conflict among many people for actual meaning of the data scientist. Many statisticians consider them self as data science. But data science is not only Statistics but more than that.

III. APPLICATION OF DATA SCIENCE

E-commerce: E-commerce and other shopping industries had got huge benefits because of data science. By doing analysis we can predict what customers are interested in and related to that item we can give some suggestion to customer which can be helpful both to customers and provider. Data scientist will take the previous searches which are made by customer and based on that they predict the interests of the customer and can suggest the product which he may be interested to buy.

Healthcare: In the health-care industry, data science has made lot of reforms. The data science has made a huge influence on medical image such as X-Ray, CT-Scans, MRI's etc. Due to data science it is now possible to detect the flaws in the body automatically. We can also discover the new cures of diseases by analyzing several combinations of formulas and their effect on the particular gene to predict the outcome. Data science has made the doctors to analyze the disease and make relation between the variable of data and also it gives insights to doctor.

Transport: In transport sector, data science is making its mark in making the safer driving for the drivers. Through a lot of researches made data science has made self-driving cars the more efficient than ever. Also, the ola and uber are also using it for the fair price for the customer and driver based on some parameters like weather, traffic etc.

Manufacturing: Data science is used in industries to improve the cost and production to give more profits. And also, with the thorough analysis of data scientists and customers, the industries are taking better decisions and improve quality. Also, the data science

has taken away the slow man work job and introduced machines.

Banking: It is one of the biggest applications of data science. With the data science banks can manage the capital more efficiently and also improves the fraud detection and management of customer and employee's data. Banks have the ability to risk modeling through data science by which they detect their performance.

Finance: Just like bank the finance industries have the risk modeling.it also allows the companies to predict the customers necessity and stock market growth.it is also playing important role in making customers experience better. We can also boost their feedback and analyze customer review

IV. Characteristics of Data Scientist

Business Understanding: It is most important characteristics as data science mainly concentrate on business growth. To do the business growth one need to fully understand the business in order to take any decision [2]. A data scientist needs to understand business specification and requirements to develop analytics according to it. To understand the business more data scientists should be in touch with consumer in order to have knowledge about growing business.

Institution: As we know math is needed here but data scientists need to pick the proper algorithm to get the highest accuracy [2]. As all models will not give the same result, data scientist will decide for deployment of model.

Curiosity: Data science has been from many years before and there is progress being made day by day. Every day there is new method which are faster than the existing one. So, as a data scientist curiosity to learn new methods becomes very important [2].

Creative: To be a good data scientist one needs to be a creative in mind to solve the problem in a attractive way so that it is very effective for the production of the product. Thinking outside the box makes the product attractive and visually understandable [3].

Communicative: In order to solve any problem data scientist, need to have good communicative skills so that they can communicate with the customers about the problems they are facing. By understanding the problem correctly problem can be solved as per the need [3].

V. Data science process:

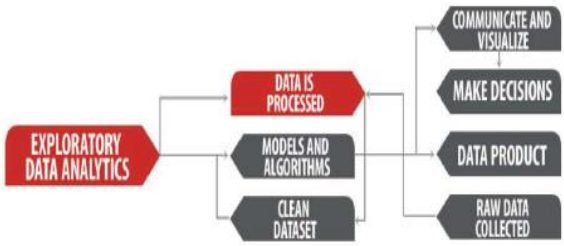

Fig 2. Steps involved in data science process

Frame the problem: First thing you have to do before solving the problem received is to define the problem [4]. When data scientist receives the data from the customer, it will be in ambiguous manner. He should convert into something actionable. When the problem is given to the data scientist it should be framed in such a way that it can easily understandable and solved.

Collect raw data needed for the problem: Once the problem is framed, data is needed to solve the problem. This process involves thinking what all data is needed to get the solution and find out the way to get that data. For that communication with the company is needed [4].

Process data for analysis: After collecting the raw data needed, it should be processed in order to do the analysis. Often data which are received will be

unprocessed [4]. Here the missing of data is checked and any corrupted values is checked such as invalid entry and many others. By this process it will be easy for the analysis team to do the analysis.

Perform in-depth analysis: In this step we have to apply mathematical, statistical, logical and technological skills to solve the problem. To solve the problem, we need to do the analysis of the problem with relevant algorithm that suits the problem to solve [4].

Communicate with the solution: When the product is ready with the solution, it is verified by the management and then showed to the customer for the feedback. Customer will be in communication for the further update as per current market business [4].

V. Advantages of data science

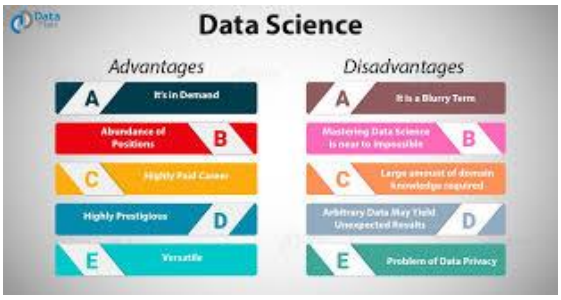

Fig 3: Advantages and Disadvantages of data science

High Demand: Data science is in great demand as there are very less who knows about that. It is one of the most demanded job in LinkedIn [5]. It is predicted that by 2026 it will create 11.5million jobs. So, data science makes a highly employable job in the industry.

Abundance of position: There are very less people who is good in data science. So, if we get good knowledge in it, we can get the high paying job [5]. To get a job as data scientist, one need to be experienced in engineering field.

A Highly paid carrier: As there are very less people who have knowledge about data science, it is in high demand. So, there is high paying jobs available in the market. Average salary of the data scientist is around 7 lakh which is very good compared to other jobs [5].it can range according to the skills and experience of individual in the field.

Data science is versatile: Data science has various application such as health care, banking and many others. We need the knowledge about the business in order to apply data science in it.

VI. Disadvantages

Mastering data science is difficult: As data science is a vast subject and has many algorithms and technique, it is nearly impossible to master the data science [5].

Large amount of domain knowledge required: In data science it is very important to have the domain knowledge as it plays main role[5]. For example, in banking sector if he wants to apply data science one should have good knowledge about banking domain in order to get the good result.

VII. Conclusion

In future data are made at unbelievable speed. To the end of this paper, we audit the introduction to data science, components to data science, Application of data science characteristics of data scientist, data science process and advantages and disadvantages of data science. Many methods utilized for the investigation purpose. We believe that in future analysts will give careful consideration to these methods to take care of the above mentioned points.

VIII. REFERENCES

[1]. “A deep desertion in data science: related issues and application” by IEEE 2019
[2]. https://intellipaat.com/blog/what-is-data-science/
[3]. https://www.educba.com/introduction-to-data-science/
[4]. https://medium.com/@storybydata/characteristics-of-a-data-scientist-ten-cs-4e3b185cc7cd
[5]. https://www.google.com/amp/s/data-flair.training/blogs/pros-and-cons-of-data-science/amp/

Cite this article as :

Ashwin Karanth, Prakash Hongal, "Survey on Data Science and Data Scientist", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.582-586, May-June-2020. Available at
doi : https://doi.org/10.32628/CSEIT2063136
Journal URL : http://ijsrcseit.com/CSEIT2063136