RESEARCH ARTICLE OPEN ACCESS

An introduction to Data Science

Vivek Kumar Sharma*, NishaVasudeva**
Computer Science Department, Arya College of Engineering and I.T, Jaipur

Abstract:

Data science is a technique to change the raw data into information. Data Science is a multidisciplinary domain that includes working with huge amounts of data, developing algorithms, working with machine learning and more to come up with business insights. Data science has attracted a lot of attention, promising to turn vast amounts of data into useful predictions and insights. Data extraction converts the unstructured data into pure and polish data that will be useful for further processing. The important characteristics that a data engineer should possess are good knowledge of machine learning, statistical skills, analytics, coding, and algorithmic experience.

Keywords **—Data Science, machine learning**

I. INTRODUCTION

Data science is a technique to change the raw data into information. It is the study of where the valuable data comes from, what it represents and how it can be turned into a valuable resource in the creation of business and IT strategies. Data Science is a multidisciplinary domain that includes working with huge amounts of data, developing algorithms, working with machine learning and more to come up with business insights. The term Data Science has emerged recently with the evolution of mathematical statistics and data analysis. It is the civil engineering of data. Its acolytes have a practical knowledge of materials and tools joined with a theoretical understanding of what is feasible .You will work with huge amounts of data. You will clean the data, prepare it and convert it into a format through which you can derive valuable insights from it. Data science can be explained as the science that deals with the identification, representation, and extraction of needful and meaningful information from a pool of data that are useful for the further growth of the business. It is actually a mixture of programming and analytics that works on unstructured raw data to create finely chopped useful pieces. The presence of a large amount of data with various structure and purpose, it is quite difficult to choose the most appropriate one. It is in this phase that the data Engineers set up databases and data storage to ease the data mining.

In a business firm, the amount of data creation increases rapidly and the data scientist helps such organizations to convert the raw data into valuable business data. Data extraction converts the unstructured data into pure and polish data that will be useful for further processing. The important characteristics that a data engineer should possess are good knowledge of machine learning, statistical skills, analytics, coding, and algorithmic experience. It includes:

- Understanding the problem
- Collecting enough data
- Processing the raw data
- Exploring the data
- Analysing the data
- Communicating the results

II. SUBSET OF DATA SCIENCE

The different subset of data science includes:

Data Analyst

It includes analysis of data using various tools and technologies. It can be done using various programming languages.

Data Architect

Data Architect performs the high-level strategies that include integrating, centralizing, streamlining and protecting the data. Data Architect should have high authority over various plans and should have good knowledge of various tools like Hive, Pig, and Spark etc.

Data Engineer

Data Engineer is supposed to work with a large amount of data where the logical statistics and programming languages club each other. The data engineer should have a software background.

III. COMPONENTS OF DATA SCIENCE

1. **Datasets**
   We need Collection of Data or a lot of data which can be analyzed, this data is fed to your algorithms or analytical tools.
2. **R Studio**
   R is an open source programming language and software environment for statistical computing and graphics that is supported by the R foundation. The R language is used in an IDE called R Studio. It is used for

- **Programming and Statistical Language**
  - Apart from being used as a statistical language, it can also be used a programming language for analytical purposes.

- **Data Analysis and Visualization**
  - Apart from being one of the most dominant analytics tools, R also is one of the most popular tools used for data visualization.

- **Simple and Easy to Learn**
  - R is a simple and easy to learn, read & write.

- **Free and Open Source**
  - R is an example of a FLOSS (Free/Libre and Open Source Software) which means one can freely distribute copies of this software, read it's source code, modify it, etc.

R Studio was sufficient for analysis, until our datasets became huge, also unstructured at the same time. This type of data was called Big Data.

3. Big Data

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

We had to come up with a tool, because no traditional software could handle this kind of data, and hence we came up with Hadoop.

4. Hadoop

Hadoop is a framework which helps us to **store** and**process** large datasets in parallel and in a distribution fashion.

Store

The storage part in Hadoop is handled by HDFS i.e Hadoop Distributed File System. It provides high availability across a distributed ecosystem. The way it function is like this, it breaks the incoming information into chunks, and distributes them to different nodes in a cluster, allowing distributed storage.

Process

MapReduce is the heart of Hadoop processing. The algorithms do two important tasks, map and reduce. The mappers break the task into smaller tasks which are processed parallely. Once, all the mappers do their share of work, they aggregate their results, and then these results are reduced to a simpler value by the Reduce process. If we use Hadoop as our storage in Data Science it becomes difficult to process the input with R Studio, due to its inability to perform well in distributed environment, hence we have Spark R.

5. Spark R

It is an R package that provides a lightweight way of using Apache Spark with R.it provides a distributed data frame implementation that supports operation like selection, filtering, aggregation etc but on large datasets.

Uses of Component of Data Science:

It contains three components i.e. (OPD) which are used for

a) Organizing Data
b) Packaging Data
c) Delivering Data

OPD Data Science Process

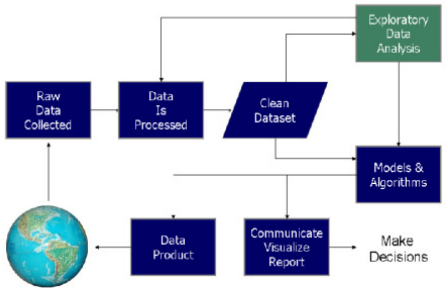
Step 1: Organize Data

It includes the physical storage and formatting of data and integrated finest practices in data management.

Step2:PackageData

In this the prototypes are created, the visualization is built and also statistics is performed. It includes logically joining and manipulating the raw data into a new representation and package.

Step 3 : Deliver Data In this process data is delivered to those who need that data.

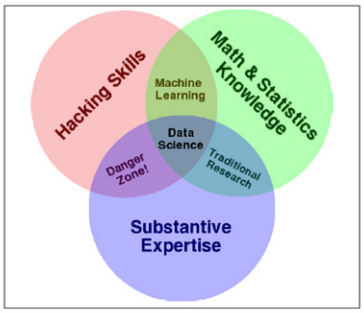


Data science process

IV. ANALYSIS OF DATA

- **Mathematical Expertise**

Before approaching the data, the data scientist should create a quantitative strategy through which exact dimensions and correlations of data can be expressed mathematically. The solutions to many business problems can be solved by building analytical models. It is a misconception that the lion's share mathematics includes the statistics. But, the fusion of both classical and Bayesian statistic is will be helpful.



- **Hacking Skills (technologies)**

Here we don't mean breaking a computer and taking out the confidential data. The hacking here refers to the clever technical skills that will make the solutions as faster as possible. Many technologies are very important in this area. Many complex algorithms are related to each task and hence the deep knowledge in core programming languages is a must. Data flow control is another sophisticated area. The man dealing with the problem should be tricky enough to find the loops and high dimensional cohesive solutions.

- **Business Acumen**

A data scientist should have a solid awareness of tactical business traps. He will be the one person in the organization that works closely with the data and hence he can create great strategies that will solve very minute problems.

V. DIFFERENTIATING DATA SCIENCE FROM BIG DATA

Big data consists of structured, unstructured and semi-structured data whereas data science deals with programming, statistical and problem-solving techniques. In big data, we will be using various methods to extract meaningful insights from large data. In data science, we will be using the above-

mentioned techniques to solve the problems. Irregular and unauthorized data will be dealing with data science.

The importance of data science is increasing day by day. There are many factors that enable its growth. Evolution of digital marketing is an important reason. The data science algorithms are used in every strategy in digital marketing to increase the CTR. Also, the data science will increase the performance. It will give way to real-time experimentation. One who can please the customers will win the business. Data science will create the best way for the same

VI. WHY DATA SCIENCE IS USED

Due to the incessant amount of data that we are creating, there is an urgent need to derive valuable insights from this data. Data is the oil of our generation. With the right tools, technologies, algorithms we can make sense of data and convert it into a distinctive business advantage.

VII. COMPARISON OF DATA SCIENCE WITH DATA ANALYTICS

Criteria	Data Science	Data Analytics
Various skills required	Data capturing, statistics, mathematics, problem-solving	Analytical, mathematical, statistical
Need to be experts in	Data mining	Data visualization
Type of data used	All types of data	Structured & mostly numeric data
Standard lifecycle	Explore, discover, investigate & visualize	Report, predict, prescribe & optimize

A lot of people confuse the role of a data scientist with the role of a data analyst. There are a lot of similarities but there are a lot of differences as well. So the above table gives you a high-level understanding of what are the major difference between a data scientist and a data analyst. One more key difference between the two domains is that data analysis is a necessary skill for data science. Thus data science can be thought of a big set while data analysis can be thought of as a subset of it. In this data science tutorial you will learn top tools, technologies, skills needed to be a successful data scientist. So this is your preliminary step to learn data science and become an accomplished data scientist.

VIII. APPLICATION OF DATA SCIENCE

Some of the major applications of data science are as below:

- Internet search
- Personalized recommender systems
- Image recognition
- Fraud detection
- Optimization techniques
- Stock market analysis
- Pathological diagnosis

IX. CONCLUSION

Data science is used for maintaining raw data change all the raw data into useful information. Here, we will work with huge amounts of data. We will clean the data, prepare it and convert it into a format through which you can derive valuable insights from it. There is different subset, component of data science. We analyse the data using different methods. Data extraction converts the unstructured data into pure and polish data that will be useful for further processing.

REFERENCES

[1] Deep Learning by lan Goodfellow,Yoshua Bengio and Aaron Courville.
[2] http://datascience.codata.org.
[3] Christopher Phethen ,Elena Simperl “The Role of Data Science in Web Science”, IEEE Intelligent Systems Vol-31,issue 3,2016.
[4] V.Dhar ,”Data Science and Prediction”, Comm. ACM,vol-56,,no 12,pp 64-73,2013.
[5] C.A. Mattmann,”Computing:A Vision for Data Science”,Nature,vol 493,no 7433,pp473-475,2013.