IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 3, Ser. II (May - June 2020), PP 01-06
www.iosrjournals.org

Towards Data Science: The Data Driven Era

Vishal Dwivedi[1], Dr.Sheenu Rizvi[2], Dr.Anuradha Misra[3]
[1]*(Department of Computer Science & Engineering, ASET /Amity University, India)*
[2]*(Department of Computer Science & Engineering, ASET / Amity University, India)*
[3]*(Department of Computer Science & Engineering, ASET / Amity University, India)*

Abstract: *Now's days, a vast number of data is swiftly produced in cyberspace. Data science is a composite number of pretexting disciplines. It refers to the study of extracting, collecting, gathering and representation of data to be used for business purpose or in technical issue. Data Science contributes a novel search way for natural and social science and goes by computer science in reasserting data. With a big amount of data now available in cyberspace, the organization in most of the field are series on data for their competitive advantages.*

This paper presents the challenges present in data and will discuss the differences in data science and other technologies like big data, the life cycle of data science, various data science technologies that are trending with its application. Also, the paper will describe the roles and responsibilities of a data scientist.

***Key Word**: Data Nature, cyberspace, data management*

Date of Submission: 06-05-2020 Date of Acceptance: 19-05-2020

I. Introduction

Data explosion is the great increment in the numbers of data in cyberspace that takes everyone in the Big Data time. The data now has no longer concise to qualitative values variables; In addition, data are everything that will be found in the cyberspace. Data Nature (the total number of data that are found inside the Cyberspace) [1]. The unique pattern exhibited by the fact existing in the natural world being surpassed by the process of gathering information in the nature of data. Data science is a collection of basics or primary theories which supports principle withdrawal of information and knowledge from the data. One of the closest technology or method that is similar and associated with data science is Data mining that is another trending technology that is the actual drawing out of data from several aspects and principals. There are approximately hundreds of algorithm and great deal of method for this field. From the day computers came to existence, we were continuously utilizing and dealing with the data.Data do no longer only intention to solve hassle based in reality but to extend analyzing information in order to learn about the phenomena and rule of information them. (For ex., discovering the increase pattern of data and predicting the scale of records in cyberspace ten years into the future). Supporting natural and social science with statistics applied sciences and strategies and exploring information nature can lead to transition toward the new science i.e., Data Science. If a person has been engaging in data science research, he or she is already become a Data Scientist.

In this given paper, I represent the challenges provided by the data and investigate why we need Data Science. Later on, I will discuss key issues like fundamental theories, new methods etc.

II. Fundamental Theories of Data Science

The Theory of Data Similarities: The key feature in mapping the relationships between data for data testing is data similarities. Research based topics includes the definition of similar measure, computational of similarities.

Data measurement and Data Algebra: -it is necessary to give complete and right theory of data science. The Relational Database Management System (RDBMS) was correct when data naturally fit into table, but it was to be known from the start that the relational model of data was incomplete. the model having imperfection become obvious primarily due to the problems while using the relational database management system with the fixed and particular structure.

Data Science research methods: It is a primary research method for data science that includes data search, data analysis and data approach. Data search explains the characteristics and structure of data set so that we can check the volume of data set and could select method for evaluate the data set.

Search of Data Nature: Fundamental rule of data search- many research records from the nature or man experiences are stored in large cyber space as in form of data. This data is the known as data nature. the search of data nature present in higher level than earlier, thus showing us to that many principles and nature laws are to be existing in data nature, for example prime numbers, Fibonacci number series etc.

III. Challenges while working with Data

As we all are aware of the fact that data is being continuously and rapidly increasing day on day basic .to keep the information and security within the data, data scientist work on several techniques and methodology. Various challenges are there arrives while working with the data set, some of them are listed below as:

Credibility of Data:The most difficult part while collecting data is too aware of while the data provided is correct or not. So, it is to know How the data which are given is providing the true or giving some irrelevant information? How do we come to know that the dataset contains unnecessary data? On the off chance that bogus information is blended with the right data, howwould we consider the certainty or validity of informational index? If some product review was provided by users and product is not used by user. and not even by competitors, these product reviews may not be valid or credible. Those results may be considered to be not truthful and data set is not credible.These are real difficulties in the data related research regions and will wind up significant part of information science inquire about. With the assistance of interpersonal organization like Facebook and other blogs growing, the difficulties are getting increasingly serious.

Existence in Cyberspace: Cyberspace is slowly turning into a section of people experiences. In straightforward words we are going to before long board each physical house and Net. However, how could we can live in cyberspace world? For example, as one is present and its presence in whole communication world and thus survival on one in many, in Net.By taking a straightforward example, The West Germanic, one in every of the net language teenagers want to communicate on-line, will be considered a communication technique in computer network. for many individuals, West Germanic is extremely tough to know as a result of it adopts words from varied totally different languages like English, French and lots of a lot of and mixes all of them along.

Research done with the Data: Data are the illustration of facts and they won't discover the laws inside universe. These developing strategies in information nature to explore the foundations within the wildlife could be a potential analysis area which will be useful for project. Data scientist acknowledges that increase within the quantity of information in several areas, several issues can't be resolved victimization ancient strategies, and that they notice the importance of information in research project. Additional on the info scientist are exploring new approaches to traumatize the info, like data processing technology and massive information analysis.

Knowledge Extraction from data set: In traditional computer era, the main concern was on how to enhance computer's performances with capacities. Now a day, the major challenges with the data is how to acquire or extract valuable information from the discrete data that are being continuously generated in cyberspace. Data science technology will help to extract the correct and only required information from the data set and that will help for the development of business strategies of any organization or bigger company along with the future planning and preparation for their research and development. Data science technique helps to extract only the valuable and required information from data set in cyberspace which have a bulk of data in it. Knowledge extraction itself is a major component in data science research.

IV. What is the Difference

Informa ionization is the process of collecting of objects and features in natural surroundings in the method of data in cyberspace. In this era, the present measure of data is increasing and being consubstantially processed in cyberspace. And the process walkdowns data explosion.

Data explosion forms data nature in Cyberspace. The data explosion is that the speedy increase within the quantity of printed information or knowledge and also the effects of this abundance. Because the quantity of obtainable knowledge grows, the matter of managing the knowledge becomes tougher, which might result in data overload.

Data Science with other Data Technologies: The traditional computer generation deals with data and their processes such as storage of data, sharing and accessing of date. Data Science uses various techniques and strategies, including: Data Acquisition, Data storage and Management. There is a vast field and applications of data science in future aspects in computer science and information technology. In the required situation, computer science builds the model for this real-world making use of computer language for the realities of the world (including human and its behavior) can be kept safely in the Computer. In these, the facts of computer system are kept in the processed data. The task of data modeling is the real method of dealing the stored data in proper manner. As we all know, the data technologies in CS were proposed to me employed in creating model for programs and facts for data computation by taking the help of computer science. So, above mentioned statements are the valuable explanation for the consequences of science of data science. These days, CS has established research on the required data well as data mining. So, the associated conferences and publications have originated from the IEEE.

Scholars as well as Data scientists are working continuously on how the model certainties with data, how to arrange and use the processed data, and the way to grow data technologies making use of computer belongs to an integral part of data science.

Data Science v/s BIG DATA: Big data in various industries drive towards the importance Data Science. The consequence of data science comprises of the questions of how data analysts and scientist could make use of use Big data for taking advantage of these things in scientist inquiries as well as researches. The continuous growth in the storage in cyberspace provides the prospect to obtain he set of big data in the several vital areas from Data Nature. It is quite simple to make use of the set of data; we can conduct further and superior data research. It is quite complicated to process big data by making use of existing technology because of their complexity and large scale. To reform these departing issues the new technology of data science has come to the rescue. Big Data technology has been empowering day by day as per the demand by various industries. Big data evolves is yet another issue in data science. Utilizing big data handles several issues in social and scientific areas and it is an integral part of data science too and the term is also known as Hadoop, this is yet another trending topic in data science technology.

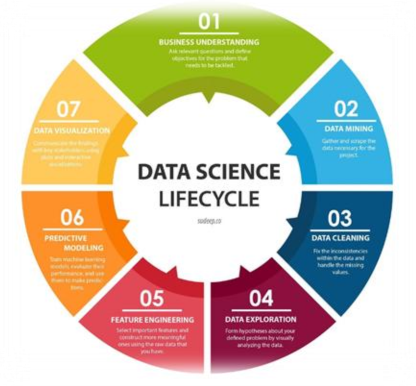
Data Science V/S Other computer Sciences: The term Data is used to represent the nature in CS with the information in term of natural phenomena of society, experiences and knowledge that comes with regular research and practices. Data can be known as representation and symbol of collected information and gathered knowledge. This shouldn't be equivalent to information, data and knowledge. The research object, methods, and goals of Data science are essentially differing from those CS, knowledge and information science. Data Science primarily supports social sciences and natural sciences, dealing with the driving forces that play major role behind the data science. Data science also termed as Data intensive science since it follows innovative research techniques with the data and information those are found inside the cyberspace. These days scientists and researches and earn achievements and honors from their biological outcomes can diminish these time-taking experiments and empower their efficiency.

On the contrary, research with the data target in data nature directly, rather than the facts in human behavior since a lot number of data exits from the references in human behavior and nature.

IV. Life Cycle of Data Science

Data Science is one of the growing aspects in the field of computer science and information technology and Emerging Trends. Some of the major components of Data Science are listed below: -

- Business Understanding
- Data mining
- Data cleaning
- Data Exploration
- Feature Engineering
- Predictive Modelling
- Data Visualization.



01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

Figure2: - Data Science Lifecycle chart

V. Trending Data Science Technologies

Here is the list of some of the technologies that are using data science

- Data management and Stewardship
- Fair data principals
- Data Integration
- Research data publications, quality and Indexing

- Infrastructure development
- Privacy aware analysis
- Machine learning /Deep learning
- Natural language processing & text-mining
- Semantics
- Data Curation
- Trend discovery & Analysis
- Graph mining & Knowledge extraction
- Social and weather sensors
- Scientific web services and work force.

VI. Technologies and Application

Method of Scientific research with Data:In the present era, computers are involved in scientific researches with the studies in vast amount and so do the data. Scientific researches and studies are confirmed with a thoughtful required for reforming in term of research, development& analysis approaches.

Domain driven data techniques:These days, researches and analysis require the integration of plentiful methods; like the blend of biological experiments as well as computation yield biometrics. Another major problem is how that is thinkable to assimilate the methods of data to a specific and customized research area.Some other applications of data science include Natural Language Processing, Security of Data, Computer vision and Business Analytics.[11]

VII. Big Data Applications

- In Healthcare
- In manufacturing
- In Media and Entrainment
- In IoT
- In Government agencies

VIII. Future Scopes and Plan

The famous American engineer and Professor Edward Deming said:

"In God we Trust, all other must bring data".

The data can be collected from anywhere through human behavior or social belongings. The major concern is to manage and store those data for further use, evaluate and to predicts some important aspects of organization
Data science is long term planned it will be task that will last to the decades and may be even more, with respect to this new and emerging technology, the data scientist should follow the pointed methods as to:

- Evolve in expanding data science as a new science and the world know the capacities and potential of the same except only developing some individuals.
- Clarify and improve the meaning of data science thoroughly
- Explore the methodology for data science
- Build up the theories of data science
- To deploy domain knowledge consisting data science (for example bio informatics, social networks).

And a lot more can be done since data science is one of the most growing and trending technologies that all aspects of data can be sorted, stored and will be kept for further use and predictions.
The recent study conducted by Indeed.com, a job provider portal that there is a huge increase in demands of data scientist in recent years. The graphical plot for the same will be discussed as:

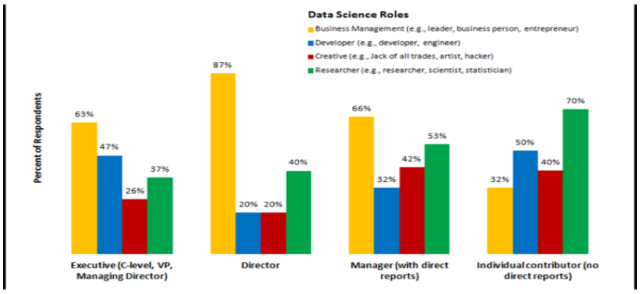


Figure 3: - The Data Science Job across Organization level.

IX. Data Scientist Skill Set

To be extremely fruitful as an information researcher, the programming abilities need to contain both computational angles — managing enormous volumes of information, working with ongoing information, distributed computing, unstructured information, just as factual perspectives

- Statics Machine learning optimization
- Programming CS fundamentals
- Visualization and Domain Knowledge
- Big data/Cloud computing
- Business data
- Communication Storytelling

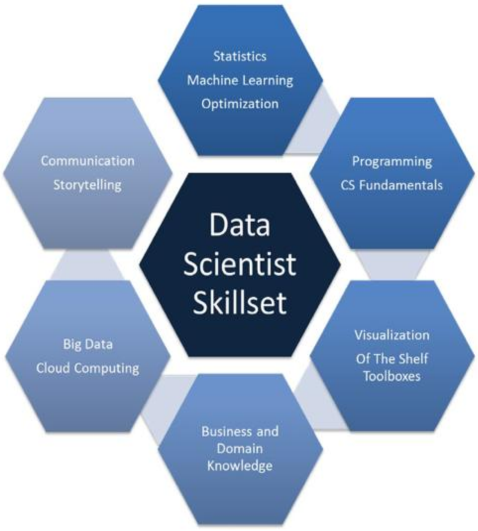


Figure 4: - Several Skills, A Data Scientist must have.

X. Conclusion

Data oriented researchers should be move towards this new science that is Data science apart from developing individual or separate data analysis and other technologies of their own.

Achievements in the present information arranged business condition requires having the option to consider how these basic ideas apply to specific business issues. This is helped by applied system that themselves are a piece of information science. There is solid proof that business execution can be improved generously by mean of information science procedure. Educational Institutions must promote this new technology for the research purposes and their future views and path ways to the new era of data science.

There is a quote which suites best with regards to Data science:

"The New Era will be Data-Driven era"

References

[1]. Dhar, V (2013) Data Science and Prediction. Google scholar.
[2]. Zhu, Zhong and Xiong(2009)., Zhu &Xiong(2009), Data Science and Methodology,
[3]. Data Science in action,F.Provost , T.Fovcett, Big data(2013), Google scholar.
[4]. Hayashi, C. (2010) What is Data Science? Fundamental concepts and Heuristic examples.
[5]. Loikides, M.(2010) What is Data science ? An O'Reilly Radar Report.
[6]. Data Science blogs: https://blogs.cisco.com/security/cyberspace-what-is-it

[7]. Data Science Journal paper : https://towardsdatascience.com/guide-to-reading- academic-research-papers-c69c21619de6
[8]. Handbook on R for Data Science by Hadley Wickham & Garrett.
[9]. Understanding Machine learning: from theories to Algorithm Shai Shalev.
[10]. Think Stats by Allen B.Downey., a book on Data Science grasp on probability.
[11]. Sanyukta Shrestha, Archana singh, Sanya Sachdev “A Deep Dissertion of Data Science: Related issue and its Applications”

Vishal Dwivedi, et. al. “Towards Data Science:The Data Driven Era.” *IOSR Journal of Computer Engineering (IOSR-JCE),* 22(3), 2020, pp. 01-06.