e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

DATA SCIENCE AND ITS RELATION TO BIG DATA AND MACHINE LEARNING

Astitva Arora *1, Dr. Parveen Kumar Gupta*2

*1Research student, Department of information technology, Jagan Institute of management studies, Rohini sector -5, Delhi, India.

*2Professor, Department of information technology, Jagan Institute of management studies, Rohini sector -5, Delhi, India.

ABSTRACT

With the multiple devices owned by us generating a massive amount of data, 2.5 quintillions of data are producing by us every day. Many big companies like Facebook, Instagram, Google, etc., are dealing massive data every second. This led to the creation of terms data science and big data. Data scientists are responsible for collecting, analyzing, and interpreting big data.

Massive data generated by companies help them to gain useful insights from the data and getting ahead of their existing competitors. Data analysts and data scientists need to analyze driven data as accurately as possible.

In this paper, we are going to find out the significance of data and how we can manipulate it by our needs predict our future, and informative insights.

Keywords: Big Data, Data Science, Machine Learning

I. INTRODUCTION

Over the past few years, Big Data is revolutionizing the way companies store their data, it allows them to manipulate large volumes of data. Every company in the world rather small or multinational companies generates data. It might be customer information or sales data. This data improves their quality of services and products they offer.

To manipulate a large volume of data we need a data scientist or data analyst to manipulate and clean raw data generated and find useful insights from them. They play a very vital role in the growth of an organization. Big Data has tremendous potential for industrial significance in a very diversity of areas, like healthcare, transportation, e-business, power supervision, and economic services [3].

But, when faced with this vast quantity of knowledge, the standard approach suffers to perform information analysis. Real systems will generate a huge quantity of knowledge from various resources, creating it advanced and tough to perform information management, processing, and analysis. it's a tough drawback for many industries and organizations to include today's "healthcare corporations," "IT departments," "government agencies" and "research institutions". To solve such quite drawback, a separate space was created for Big Data science and new trends square measure required for analysis and education efforts for fast and triple-crown development [2].

Machine learning is a powerful tool used in the big data processing. The relation between human intelligence and its potential to learn through resources is similar to the relation between machine learning and big data [4]. As considering data as a resource machines can learn through it and solve complex problems.

With considerations from prior researches, we will find out the significance of Big Data and its relation with machine learning and data analytics.

II. LITERATURE REVIEW

Below are the studied papers for this issue addressed in the title. Table 1 shows the name of the author and the issue addressed with the key points. Most of the papers have the same issue related to big data or machine learning.

Table 1.

S.no	Author	Issue address	Key points	conclusion
1.	Foster Provost and Tom Fawcett	Paper on data science, big data technologies	Decision making, analysis,	Explaining relations between data science and its applications working with big data.
2.	Vishnu Vandana	Paper on the need for	Big data, Machine	New methods to assist

	Kolisetty and Dharmendra Singh Rajput	research aimed at proposing new techniques that can be used to analyze BD	learning, Data analysis, Big data implications, Big data challenges.	machine learning algorithms and improvement in big data analysis.
3.	Wullianallur Raghupathi and Viju Raghupathi	Paper on big data analytics for healthcare researchers and practitioners.	Big data, Analytics, Hadoop, Healthcare, Framework, Methodology	Explaining how big data analytics helping in the healthcare sector.
4.	Han Liu , Alexander Gegov and Mihaela Cocea	Paper on Unified Framework for Control of Machine Learning Tasks towards Effective and Efficient Processing of Big Data	Big Data, Computational Intelligence, Data Mining, Machine Learning, Data Processing, Predictive Modelling	Explaining cross-validation and a way to measure the extent to which an algorithm is suitable to build a predictive model on the basis of the existing data provided.
5.	Anand Gupta, Hardeo Kumar Thakur, Ritvik Shrivastava , Pulkit Kumar , Sreyashi Nag	Paper on A Big Data Analysis Framework Using Apache Spark and Deep Learning	Spark, Mlps, Deep Learning,Data manipulation, Machine learning models,	Explaining how models are trained when input data and how its proficiency is increasing while every single cycle it completes.
6.	H.V. Jagadish	Paper on Big Data and Science: Myths and Reality	Big data, data mining, computing architecture	Explaining reality and myths about big data and approaches

III. METHODOLOGY

Machine learning – Starting Phase

Machine learning is defined as a computer gaining experience automatically and processing efficiently and effectively. We provide computer algorithms that will compute complex problems more accurately while running. For example, a toddler learning to walk on his/her practices day to day learning to control his motor muscles, once it gains control, he will be able to walk efficiently same is with the computer once it gains experience with resources provided it will for much faster than initial.

There are many real-time applications we take for granted which use machine learning and making our life east like computer vision, speech recognition, bio-surveillance, robot control, bots, accelerating in empirical science, and many others [5]. The framework of machine learning on big data is interacting with components, examine and generating output.

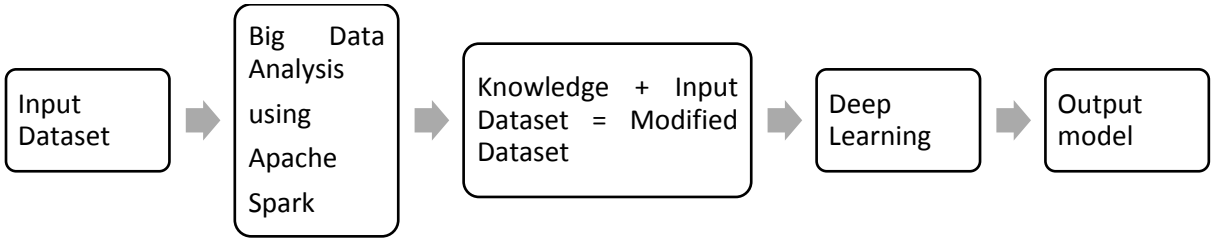

Fig 1. The framework of machine learning

Fig1 shows the working of a problem and computing its logic to get useful insights. Recent research conducting real-time experiments while taking dataset of H-1b Visa Application Record and Cardiac Arrhythmia Dataset, we find out the accuracy level is increasing while running this model again and again. This shows the machine is adapting and learning by itself from the resources and inputs we are providing.[6].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

Big Data – Knowledge For Machine

Massive data has been organized into 5 categories: volume which deals with the quality of data, velocity which deals with the speed of data generation, variety deals with which type of data and in what format we are getting it, veracity is quality of data captured and value is impacted it gives.[7]

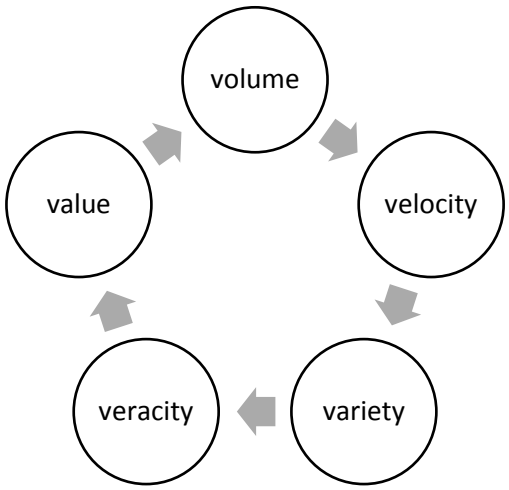

Fig 2. 5 v's of big data

For our convenience, we distribute it in a stack where there are 3 layers big, data, and value.

The big layer is the most important layer as it deals with the volume and velocity of data where the data layer depends more on heavily om technology advancement and the highest layer deals with the strategic power of big data.

Data collected is in raw unstructured form, we have to clean, manipulate it so that our model can train itself from the inputs. Machine learning models and algorithms are used to train and test also known as learning and prediction. The activity called predictive modeling helps it to make predictions from previous predicted data.[2]

The process of transforming and analysis of big data is machine learning models often required cleaned and processed data to train its particular build model. Data is scraped from different sources; the format may be different. Data is transformed into one particular formatting for input, then data is integrated from various data set scrapped earlier. Data is processed and cleaned for further discovery and then we visualize it and make a meaningful decision.

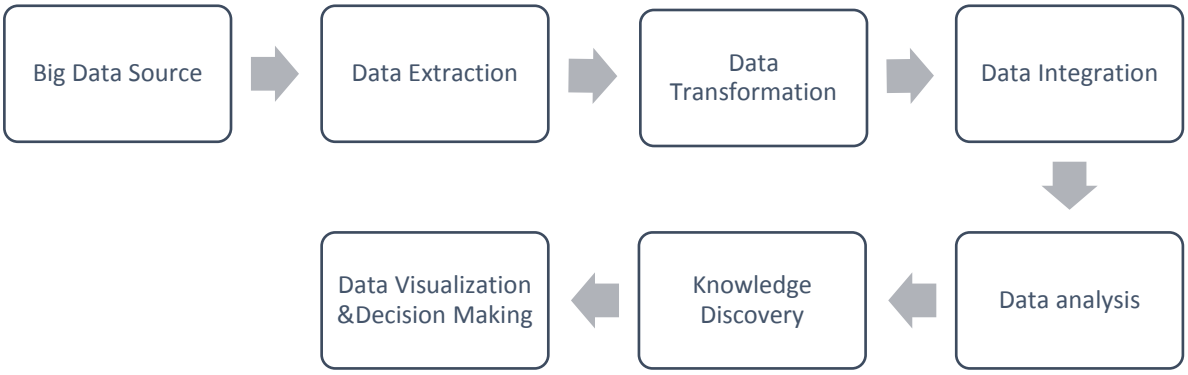

Fig3. Process of big data

Data Science – Prediction and Analysis

Principles, processing, and techniques used in understanding the problems via automated analysis of data in data science. Business intellectuals are more interested in improving decision-making through data science.

Big data and data science are not interchangeable terms whereas big data is handling data characteristics whereas data science is using data to find meaningful insights. Yes, both fields are working together but have different approaches [8].

The architecture used by a data scientist is to identify and evaluate problems then acquire or scrape data from different sources and explore it with context to the problem addressed and build a model through a defined machine learning model to train a model from driven data and monitor for optimized results. Once the accuracy level is considerable it will deploy a solution for the given model.[9]

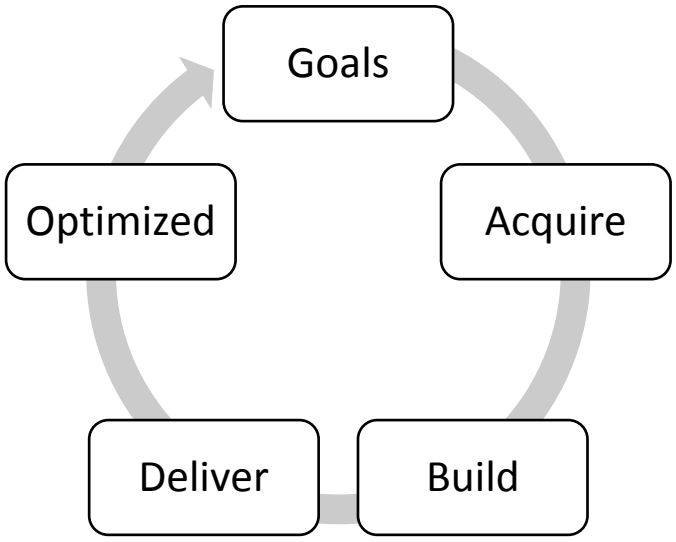

Fig4. The cycle of data science

This work is done by some existing tools and programming languages like python, R, Sol, Java, Julia, and Scala. Mainly python and R is most popular among computer scientist and statistician. Visualizing data, modeling, mathematical algorithms are implemented in seaborn, matplotlib, Scikit-learn, TensorFlow, PyTorch, Pandas, Numpy.[9][10].

IV. RESULTS AND DISCUSSION

We find out some of the positive and negative aspects of perusing and putting time into data science. As a growing career, we have surveyed and provided points below.

Advantages

- It is a highly paid career
- Future prediction is possible
- While using machine learning products are becoming smarter
- Wide numerous applications are there for implementing data science

Disadvantages

- It does not have a definite definition so it depends on the field that the company is specializing in.
- Mastering data science is nearly not possible because it is a mixture of various other fields.
- A large number of skills and information is required.
- The data used in solutions may breach data privacy.

V. CONCLUSION

As we have discussed above the significance of data science and its processing. We find out the machine learning models and data play a very major role in finding the insights from the given knowledge. Models working continuously gaining knowledge by inputs of data and making it more efficient. Through the logistic regression approach, we can see the score and model efficiency. So, while conducting experiments on models and reading other research papers we can find out the significance of data science in every sector and how It can change the way we see things.

Almost every sector in the industry uses its previous data to find their mistakes and predict future data.

As a result, we can say that data science is working parallel with the bid data and machine learning.

VI. REFERENCES

[1] Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51–59.

[2] Vishnu Vandana Kolisetty, Dharmendra Singh Rajput (2019). A Review on the Significance of Machine Learning for Data Analysis in Big Data.Jordanian Journal of Computers and Information Technology 06(1):1

[3] W. Raghupathi and V. Raghupathi, "Big Data Analytics in Healthcare: Promise and Potential," Health Information Science Systems, vol. 2, no. 1, pp. 1-10, 2014.

[4] H. Liu, A. Gegov and M. Cocea, "Unified Framework for Control of Machine Learning Tasks Towards Effective and Efficient Processing of Big Data," Springer Data Science and Big Data: An Environment of Computational Intelligence, pp. 123–140, 2017.

[5] Mitchell, T. (2006). The discipline of machine learning. (Vol. 9) Carnegie Mellon University, School of Computer Science, Machine Learning~….

[6] Anand Gupta,Hardeo Kumar Thakur,Ritvik Shrivastava,Pulkit Kumar, Sreyashi Nag, A Big Data Analysis Framework Using Apache Spark and Deep Learning

[7] Zhou, L., Pan, S., Wang, J., & Vasilakos, A. (2017). Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, 350–361.

[8] Jagadish, H. (2015). Big data and science: Myths and reality. Big Data Research, 2(2), 49–52.

[9] Alex Castrounis, What Is Data Science, and What Does a Data Scientist Do?(2020)

[10] Jake VanderPlas Python Data Science Handbook first edition(2017), preface XV

[11] Parveen Kumar Gupta, Rahul Rishi, Ranjit Biswas(2013).a comparative analysis of temporal data models, published in International Journal of Advanced Computational Engineering and Networking, (full text available online at journal's website).

[12] Gupta, P., Rishi, R., & Biswas, R. (2013). A comparative analysis of temporal data models. International Journal of Advanced Computational Engineering and Networking,

[13] Kaliraman, J., Gupta, Praveen kr. and Bansal S., "Vehicle Security System and Accident Information System by using VLSI", contributed and presented in international conference ICICCT 2016, sponsored by CSI and organized by JIMS, Rohini, Delhi, ISBN 978-93-85777-66-0, Excel Publishers India, pp 122-130.

[14] Gupta, Praveen Kumar, Rishi, R. and Biswas, R, "An investigative analysis of timestamping approaches of temporal models", published in International conference (i.e. ICACCT 2013), publisher – Inderscience, UK.

[15] Parveen Kumar Gupta, Master Data Management, international journal of engineering, technology, science and research volume- 4 issue 6 page 268 – 272.

[16] Deep learning architecture, technowhiz 2019.

[17] Big data challenges, technowhiz 2016.