IS BIG DATA THE CAUSE OF THE EMERGENCE OF DATA SCIENCE & HOW DATA SCIENTISTS BECAME THE MOST IMPORTANT AGENT FOR SOCIETY?

Dr. Kashif Qureshi[1] and Dr. Gajendra K. Saraswat[2*]

[1]*Dean Cum Professor in Faculty of Engineering & Technology, Chhatrapati Shivaji Maharaj University, Navi Mumbai, Maharastra, India,*

[2]*Professor in the Department of Mathematics, Modern Institute of Technology and Research Center, Alwar, Rajasthan, India,*

Abstract: *Data turns into big data because of volume, speed, velocity and so many V's. There are in total 42 V's, but we will discuss main V's. Handling big data is a very big task to handle big data Technology coined Data Science to handle big data and its associated problems. In this paper, I have discussed concepts of big data, big data analytics, Machine learning, data science and their related impacts. In this paper, We have discussed concepts of big data, big data analytics, Machine learning, data science and their related impacts on real-life and ease delivered by data science.*

Keywords: Big Data, Big Data Analysis, Big Data Analytics, Machine Learning, Data Science, Data Scientist

1. INTRODUCTION

The main source of big data is social media it's a big challenge to figure out meaningful information from the big data for an individual and institution and so on. Data science is a collection of artificial intelligence, machine learning and deep learning. Deep learning is the subset of machine learning and machine learning is a subset of artificial intelligence. By the use of various kinds of algorithms in machine learning with creating the model, train the model to solve the artificial intelligence-based problems.

Problems can be any type, social problem or technical problem. In general, there are three types of machine learning. Supervised, unsupervised, reinforcement. Supervised, unsupervised and reinforcement machine learning have their nature of solving the problems, There are 14 algorithms which solves a different kind of problems. Now the issue came which algorithm should be used for A.I problem. Machine Learning engineer understand this issue by the experience, finally, he/she knows very well how to solve a problem with the help of an algorithm.

2. ISSUES WITH DATA

Data accuracy, data cleaning, data visualization, data completeness, handling missing data and so on. The major issue with big data is how to get the required information from Mammoth of data or information. Big data is responsible for the data science creation, data science means the science of data, we are dealing with all aspects of big data to get the best and required information which is created by more than 4 billion people. This all we are trying to do by the help of machine learning, the meaning of the machine learning is to train the machines which can act/ perform like a human as much as possible, as an example Alexa, Siri, are the natural language processors, they understood easily what human wants, as an example if I say to Alexa, Alexa play song, then I can listen to a song very easily, Alexa turns off the lights, Alexa wake up me at 5:00 a.m.Creating and training a model we say it's machine learning. The machine understood how to perform or solve a problem to get its best solution. All A.I. based problems are handled & solved by machine learning engineers or data scientists.

SOURCES OF BIG DATA

Social media is the main source of big data. The uses of below mentioned social media websites as of January 2021 are as follows, Facebook (2.74 billion users), YouTube (2.29 billion users), WhatsApp (2 billion users), Facebook messenger (1.3 billion users), Instagram (1.22 billion users), TikTok (689 million users), QQ (617 million users), Douyin (600 million users) and Sino Weibo (511 million users).

5 TYPES OF SOCIAL MEDIA

Social networking, book marking, news forum, micro blogging and online forum sites.

DATA AND BIG DATA

Big data is fast, voluminous,& complex. It is very hard to deal with big data. Unlikely data is very easy eg: data of own written notes, data of my address, data of my University, data of my employment etc. It is very easy to know about someone's data, unlikely if more than 400 billion people data is available it becomes extremely complex and turns into big data. It has three virtues; it will be very fast, voluminous and complex.

GENERATING RATE OF BIG DATA

As per the statistics, 2.3 trillion Gigabytes generating by social media users, so far almost 50 trillion Gigabytes have been generated from social media. On an average medium companies storing 100 terabytes of data each day. As per the statistics Facebook processing 500 Terabytes of data each day, only YouTube has 1.325 billion users and YouTube uploads 300 hours of video every minute, 4.95 billion videos viewed by the users on YouTube every day, almost 1 billion new users surf YouTube every month, each month 3.25 billion hours video watched by the users on YouTube,

DATA SCIENCE

Data science is a collection of artificial intelligence, machine learning and deep learning and it is responsible to handle big data and all associated problems and solutions of big data. Big Data Analytics, machine learning, understanding and application on big data are the main responsibility of data scientists to get the best solution or information from Mammoth of big data by the usage of machine learning algorithms, Complex mathematical models, Complex statistical models, various software tools, various kinds of algorithms in order of creation model and training the model to generate the output like a human does. Data science is related to the future and prediction like weather forecasting, sales trend, stock market prediction etc.

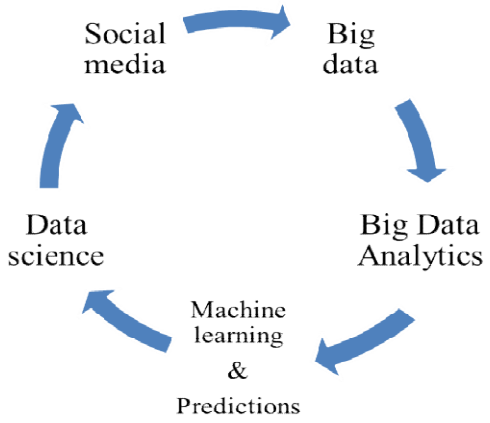
IMPORTANCE OF DATA SCIENCE

Prediction(s) is/are the main goal of the data science to get the best results,outputs or information for the betterment of any business, Business can have any domain or verticals, to get the maximum benefit from the customers or clients in terms of monetary & life -long association. Finally, data science gives us an edge over our competitors in the business world and all areas of life

APPLICATION OF DATA SCIENCE

Data Science is a part of the real world and real-time activities of humans. Data science playing an exceptional role in Health Care, fraud and risk detection, advanced image recognition, targeted advertising, product recommendations, speech recognition airline route planning, cars route planning, ships route planning, gaming, Augmented reality, self-assessment prediction, astrology, numerology, Finance, financial planning, predictions, predictions of sentiments, prediction of behavior and all areas of life

HIERARCHY OF DATA SCIENCE



DATA SCIENCE SOLVING GRITTY DEFIES

Data science giving solutions in the areas of Business intelligence, customer relationship management, education solution, Healthcare solution, security solutions, stock and investment analysis and prediction, human brain analysis and patterns, etc.

DATA SCIENTIST

He/she/ They solves the business problems, define the right requirements for the business growth, extract correct data and information, Implement right Technologies and tools to solve the business problems, implement Data Analytics, responsible for data cleaning, data visualization data mining evaluation, Automatic model-based decision making, They create the model and train the model, They keep surveillance and monitoring to scale up the performance of the model for the accurate prediction and business growth. They keep an eye on the latest trends of customer behavior, sentiment, customer hobbies, customer attitude, customer likes and dislikes and their criteria. The biggest challenge finding the experienced and wise data scientists, it is the hardest task because data scientist needs super skills and must cover the domain of machine learning engineer, data engineer, big data architect, big data engineer, business intelligence consultant, etc.

BIG DATATECHNOLOGIES

Hadoop is the most popular technology to process and store big data using HBase, hive, pig, NO SQL, cloud computing, data mining, data cleaning and using accurate and complete data. Python 3.6 or Python 3.10 is the best language to handle artificial intelligence problems uses machine learning algorithms applied to big data.

PYTHON

Python is a high-level language that has the maximum libraries so far, that is why it is the easiest language to understand and learn. Python has the greatest edge over other languages because of its libraries which are problem-oriented. **Main libraries of Python are** Scikit learn, Numpy, Pandas, Tableau, Tensor flow, Pytorch, Milk, Kailash, Pandas, Matplotlib, Seaborn, Nltk, etc.

3. CONCLUSIONS

In this paper, we have first investigated the background of big data, data science, data scientist, and the current state of data science. Then we have proposed several approaches to data science. The biggest bottleneck in the big data era is the production of capable data scientists. Tools and languages can be learned, but people who can manage real-world data science projects and who own the necessary big data analytics skills and knowledge are rare. Producing such capable people takes time. Hence, we need multiple approaches within data science.

We have proposed various approaches for data science based on a far-reaching assessment of current data science and domain knowledge in the field. In a nutshell, data science should cover (1) Data Analysis and Data Analytics, (2) eight-step data analytics lifecycle, (3) Big data technologies and model-building techniques, (4) Relation of Data analysis and data analytics, (5) Big data analytics as well as small data analytics, (6) real-world project experience Using automated tools (IBM Watson Analytics) or dashboards that use a black-box approach would be an important solution in training data scientists. However, the users of those tools should still be familiar with the methods implemented in the systems to choose the right method that fits the given data set and to interpret the outcomes properly. Those users should have the critical thinking and reasoning ability to explore the solution space provided in the tool and to determine whether the tool can indeed provide a satisfactory outcome. We believe that data science will get more and more attention, and it is time to do more research on how data science should be learned and implement? Data science educators should keep working on how to re-train people, processes, and technologies around big data to change for the better. As we get more experience, we will be able to improve data science programs and educate successful data scientists.

REFERENCES

[1] *Yeol Song, Yongjun Zhu, Big data and data science: what should we teach? (2015)*

[2] *V. Mayer-Schonberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt (2013)*

[3] *R. Thomson, C. Lebiere, S. Bennati, Human, model and machine: a complementary approach to big data, in Proceedings of theWorkshop on HumanCentred Big Data Research, HCBDR '14, 2014. (2014)*

[4] *A. Cuzzocrea Privacy and security of big data: current challenges and future research perspectives, in: Proceedings of the First International Workshop onPrivacy and Security of Big Data, PSBD '14, 2014. (2014)*

[5] *Big data, Nature 455(7209) (2018) 1–136.*

[6] *Dealing with data, Science, 331(6018) (2011),639–806.*

[7] *J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung, Bigdata: the next frontier for innovation, competition, and productivity, Tech. rep., McKinsey GlobalInstitute,availableat:http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. (2011).*

[8] *C. O'Neil, R. Schutt , Doing Data Science: Straight Talk from the Frontline, O'Reilly Media, Inc.,. (2013).*

[9] *Big data, http://en.wikipedia.org/wiki/Big data. (2014).*

[10] *G. Li, X. Cheng, Research status and scientific thinking of big data, Bull. Chin. Acad. Sci. 27(6) (2012), 647–657.*

[11] *Y. Wang, X. Jin Xueqi, Network big data: present and future, Chinese J. Comput. 36(6) 1125–1138.*

[12] *X.-Q. Cheng, X. Jin, Y. Wang, J. Guo, T. Zhang, G. Li, Survey on big data system and analytic technology, J. Softw. 25(9) (2014) 1889–1908.*

[13] *W.B. Arthur, The second economy, available at: http://www.images-et reseaux.com/sites/default/files/medias/blog/2011/12/the-2nd-economy.pdf. (2011)*

[14] *T. Kalil, Big data is a big deal, available at: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal. (2012)*

[15] *T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation. (2009)*

[16] *Data science, http://en.wikipedia.org/wiki/Data science, (2014*).

[17] *M. Loukides, What Is Data Science? O'Reilly Media, Inc., (2011)*

[19] *M.A. Stokes, China's nuclear warhead storage and handling system, Tech. rep., 2049 Project Institute. (2010).*

[20] *I.M. Easton, L.R. Hsiao, The Chinese people's liberation army's unmanned aerial vehicle project: organizational capacities and operational capabilities, Tech.rep., 2049 Project Institute. (2013).*

[21] *J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant , Detecting influenza epidemics using search engine query data, Nature, (2009) 72321012–1014.*

[22] *Big data for development: challenges & opportunities (2012), available at http://www.unglobalpulse.org/projects/Big Data for Development.*

[23] *Declaration to be the world's most advanced IT nation, available at:http://japan.kantei.go.jp/policy/it/2013/0614_declaration.pdf,*

[24] *Wikipedia*

[24] *Geekfoegeeks*