International Journal of Research (IJR)
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Vol. 9 Issue 07
July 2022

The Concept of Data Science in the Era of Big Data

[1]Osuolale Peter POPOOLA, PhD., [2]Olukunle Dejuwon OLANIYAN, [3]Gedion Gbenga OYETORO
[1,2&3] Mathematics and Statistics Department, Adeseun Ogundoyin Polytechnic, Eruwa
Oyo State, Nigeria
Corresponding Author Email: Osuolalepeter@yahoo.com

Abstract

The omnipresence of data in our contemporal world is helping to reshape our world. With the unprecedented rate of data creation, and the increasing role data plays in most of our lives, it is easy to assume that the digital revolution is the most important life-changing event of this era. The high volume, high velocity and wide variety of data generated by the digital revolution are commonly referred to as Big Data. Before, researchers were facing the problem of how to store big data, but now, the main focus is not how to building a framework and solutions to store big data, companies like Hadoop, Hbase, CouchDB, (storage platform that stores structured and unstructured Data) and others have successfully solved the problem of storage, the focus has shifted to the processing of these big data. However, the volume and variety of data have far outstripped the capacity of manual analysis. As data continue to grow in size and complexity, new algorithms need to be developed so as to learn from eclectic data sources At the same time, computers have become far more powerful, networking is ubiquitous, and algorithms that can connect datasets to enable broader and deeper analyses than previously possible. The limitation of conventional statistics to manage and analyze big data has inspired data analysts to venture into data science.. Hence, the concept of data science. Data Science is a combination of multiple disciplines that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it. This paper gives general overview of big data revolution, the concept of data science, it uses, it structures, it relationship with other disciplines, application areas and, how it works.

Keywords: *Big Data, Data Mining, Data Algorithms, Data Science*

1.0.Introduction

Computer and internet revolution have enhanced data revolution by providing means to deal with high dimensional large data, and its easy storage and fast transmission and remote access. We started from very slow manual calculating machine, and then came the scientific hand calculator. That was of no comparison to the mainframe computer of 1970's. The introduction of personal computer (PC) in 1980's, laptop in the later part of twentieth century and later tablet and super computers have revolutionized the computing forever. Couple with the access to the high-power internet superhighway the computing technology has changed the way we think and work. As a

more recent development, the introduction of smart phone and other communication technologies have created much interconnected world. No doubt, the advents of internet and smart phones have brought about data revolution into our world. With the unprecedented rate of data creation, and the increasing role data plays in most of our lives**,** it is easy to assume that the digital revolution could be the most important life-changing event of this era. The high volume, high velocity and wide variety of these data generated by the digital revolution are commonly referred to as Big Data. Big Data is often defined as a data set that is huge, multidimensional, and complex so that traditional data processing methods, techniques and applications are inadequate to deal with them. There are challenges to managing such a huge volume of data such as capture, store, data analysis, data transfer, data sharing, etc. The common features of Big Data are high "Volume", "Velocity" and "Variety" which are popularly known as 3V (Popoola and Nuhamah, 2018) The High-volume refers to increasing Exabyte data generated by machines, networks, and human interaction; high-velocity refers to the speed at which data are created, processed, and stored; and high-variety relates to the range and complexity of data types and sources. Big Data are data that are so large and complex that traditional data-processing applications become insufficient to capture, store, and analyze the data. Instead, a network of human skills, advanced technologies, and data access infrastructure are essential to handle big data and his of course, is a key challenge for statisticians and policymaking organizations seeking to incorporate big data in their toolkits. However, over the past few years, the Internet has democratized the creation, access, and analysis of large datasets.

1.1. Big Data Revolution:

In a recent study, it was reported that 90% of the entirety of the world's data has been created within the previous two years. In just two years, the world have collected and processed 9x the amount of information than the previous 92,000 years of humankind combined. And it isn't slowing down. It's projected the world had already created 2.7 Zettabytes of data, and by 2025, that number will balloon to an astounding 85 Zettabytes. What do we do with all of this data? How do we make it useful to us? What are it's real-world applications? These questions are the domain of data science. As data continue to grow in size and complexity, new algorithms need to be developed so as to learn from eclectic data sources. The actual size of the big data and how fast its volume is increasing is a real surprise to many people. Even with the mainframe computer the volume of data was measured in kilobytes to megabytes. It is only after the

introduction of PC and laptop the volume of was counted in gigabytes. But now it is much bigger, and the phenomenal growth of data is so much and so fast that total data volume in the world doubled only in two years. Data are now measure in Terabytes, Exabyte, Petabytes, Zettabytes, and Yottabytes. Big Data can't be stored and analyzed in any ordinary computer because of its size. People are now using internet and cloud technology to store big data (cloud computing e,g Microsoft Azure). Big data always requires lots of cleaning and formatting before it could be used and analyzed. Often interest in big data is to find correlation, classification, clustering, and structure. Artificial Intelligence, Data Mining and Machine Learning algorithms are often used to analyze big data using appropriate statistical method. (Sahinoglu and Cueva-Parra, 2010). The limitation of conventional statistics to manage and analyze big data has inspired data analysts to venture into data science.

1.1. Sources of Big Data:

Unlike traditional data (sample survey, censuses and administrative) that are compiled for specific purposes, big data is a byproduct "found" in business and administrative systems, social networks, and the internet of things. Social networks are online platforms that help people build social relations with others having similar interests e.g Facebook, Twitter, LinkedIn etc. Users create blogs and profiles, share pictures, and exchange messages and thereby provide human-sourced information that is digitalized and stored. Data in social networks are often ungoverned and unstructured. In its big data classification, the United Nations Economic Commission for Europe (UNECE, 2013) also includes in social networks internet searches and mobile data that can be more widely understood as human-sourced information. Traditional business systems are processes and procedures defined by businesses to provide value to their customers and generate process-mediated data, including administrative records. Business systems record well-governed, structured information on transactions, positions, and metadata related to business events (commercial transactions such as registering a customer or receiving an order) stored in relational database systems. The internet of things is a system of data-producing interrelated computing devices with embedded sensors and internet connectivity that measure and record events and situations in the physical world. Their output is structured machine-generated data (sensor records, computer logs, webcam, and mobile phone location/GPS (UNGP. 2012).

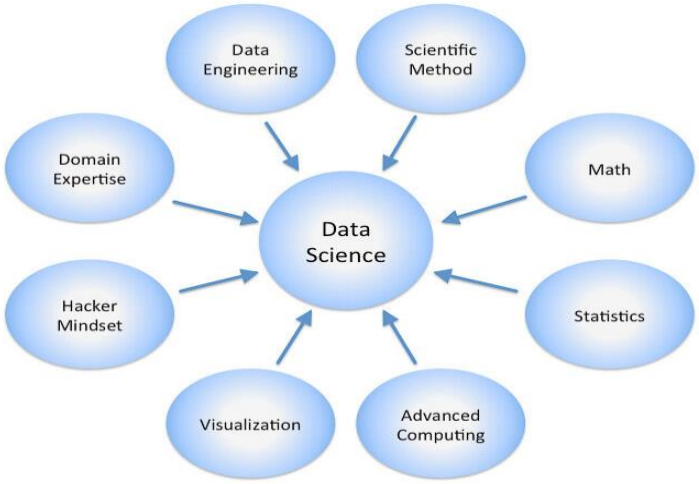
2.0.**Data science**:

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, (*Dhar, 2013*), (Jeff, 2013)and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data. Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyze actual phenomena" with data (*Hayashi,1998)* It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge (*Cao, 2017)* However, data science is different from computer science and information science. According to Tony and Bell (2009), defined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge. Data science not only requires conventional statistical methods, it also needs skills such as statistical signal processing, pattern recognition, data mining, machine learning, bioinformatics, meta-analysis etc. (Khan, 2020). Khan (2020) discussed various statistical models to Meta-analyse data from heterogeneous primary studies including the inverse variance heterogeneity (IVhet) model.

2.1. How Does Data Science Work?

Data science involves a plethora of disciplines and expertise areas to produce a holistic, thorough and refined look into raw data. Data scientists must be skilled in everything from data engineering, math, statistics, advanced computing and visualizations to be able to effectively sift through muddled masses of information and communicate only the most vital bits that will help drive innovation and efficiency. Data scientists also rely heavily on artificial intelligence, especially its subfields of machine learning and deep learning, to create models and make predictions using algorithms and other techniques.

Here is a diagram showing some of the common disciplines that a data scientist may draw upon. A data scientist’s level of experience and knowledge in each often varies along a scale ranging from beginner, to proficient, and to expert, in the ideal case.



2.2. Data Science Applications

2.2.1. In Marketing: for tasks such as targeted marketing, online advertising, and recommendations for cross-selling. Data science also is applied for general customer relationship management to analyze customer behavior in order to manage attrition and maximize expected customer value.

2.2.2. The finance industry: uses data science for credit scoring and trading and in operations via fraud detection and workforce management. Major retailers from Wal-Mart to Amazon apply data science throughout their businesses, from marketing to supply-chain management. Many firms have differentiated themselves strategically with data science, sometimes to the point of evolving into data-mining companies. But data science involves much more than just data-mining algorithms. Successful data scientists must be able to view business problems from a data perspective. There is a fundamental structure to data-analytic thinking and basic principles that should be understood. Data science draws from many ‘‘traditional’’ fields of study. Fundamental principles of causal analysis must be understood.

2.2.3. Business world: How about if you could understand the precise requirements of your customers from the existing data like the customer’s past browsing history, purchase history, age and income. No doubt you had all this data earlier too, but now with the vast amount and variety

of data, you can train models more effectively and recommend the product to your customers with more precision. Wouldn't it be amazing as it will bring more business to your organization?

2.2.4. Data Science in decision making: How about if your car had the intelligence to drive you home? The self-driving cars collect live data from sensors, including radars, cameras, and lasers to create a map of its surroundings. Based on this data, it takes decisions like when to speed up, when to speed down, when to overtake, where to take a turn – making use of advanced machine learning algorithms.

2.3.5. In predictive analytics (Weather Forecast): Data from ships, aircraft, radars, satellites can be collected and analyzed to build models. These models will not only forecast the weather but also help in predicting the occurrence of any natural calamities. It will help you to take appropriate measures beforehand and save many precious lives.

2.2.6. Healthcare: Data science has led to a number of breakthroughs in the healthcare industry. With a vast network of data now available via everything from EMRs to clinical databases to personal fitness trackers, medical professionals are finding new ways to understand disease, practice preventive medicine, diagnose diseases faster and explore new treatment options.

2.2.7. Cars Manufacturing Company: Car Manufacturing like Tesla, Ford and Volkswagen are all implementing predictive analytics in their new wave of autonomous vehicles. These cars use thousands of tiny cameras and sensors to relay information in real-time. Using machine learning, predictive analytics and data science, self-driving cars can adjust to speed limits, avoid dangerous lane changes and even take passengers on the quickest route.

2.2.8. Courier services: Courier Company like UPS has turns to data science to maximize efficiency, both internally and along its delivery routes. The company's On-road Integrated Optimization and Navigation (ORION) tool uses data science-backed statistical modeling and algorithms that create optimal routes for delivery drivers based on weather, traffic, construction, etc. It's estimated that data science is saving the logistics company up to 39 million gallons of fuel and more than 100 million delivery miles each year.

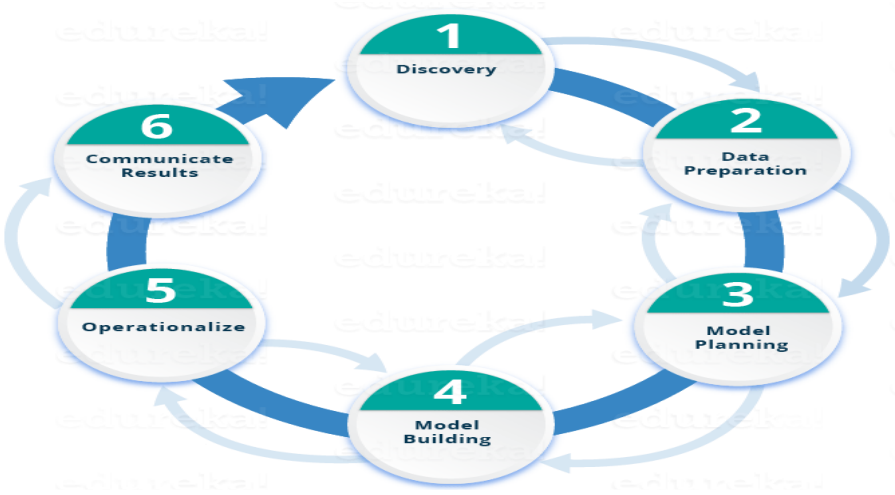
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Vol. 9 Issue 07
July 2022

2.2.10. Entertainment Industry: Do you ever wonder how Spotify just seems to recommend that perfect song you're in the mood for? Or how Netflix knows just what shows you'll love to binge? Using data science, the music streaming giant can carefully curate lists of songs based off the music genre or band you're currently into. Really into cooking lately? Netflix's data aggregator will recognize your need for culinary inspiration and recommend pertinent shows from its vast collection.

2.2.11. Cyber security: Data science is useful in every industry, but it may be the most important in cyber security. International cyber security firm Kaspersky is using data science and machine learning to detect over 360,000 new samples of malware on a daily basis. Being able to instantaneously detect and learn new methods of cybercrime, through data science, is essential to our safety and security in the future.

3.0. The Lifecycle of Data Science

Here is a brief overview of the main phases of the Data Science Lifecycle:



3.1. Phase 1—Discovery: Before you begin the project, it is important to understand the various specifications, requirements, priorities and required budget. You must possess the ability to ask the right questions. Here, you assess if you have the required resources present in terms of people, technology, time and data to support the project. In this phase, you also need to frame the business problem and formulate initial hypotheses (IH) to test.

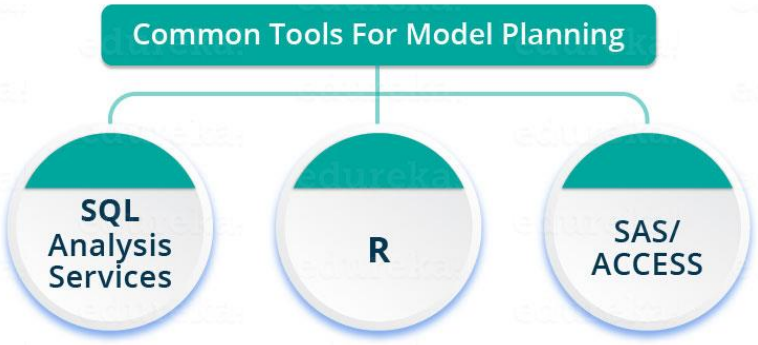
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Vol. 9 Issue 07
July 2022

3.2. Phase 2—Data preparation: In this phase, you require analytical sandbox in which you can perform analytics for the entire duration of the project. You need to explore, preprocess and condition data prior to modeling. Further, you will perform ETLT (extract, transform, load and transform) to get data into the sandbox. Let's have a look at the Statistical Analysis flow below.

You can use R for data cleaning, transformation, and visualization. This will help you to spot the outliers and establish a relationship between the variables. Once you have cleaned and prepared the data, it's time to do exploratory analytics on it.

3.3. Phase 3—Model planning: Here, you will determine the methods and techniques to draw the relationships between variables. These relationships will set the base for the algorithms which you will implement in the next phase. You will apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.

Let's have a look at various model planning tools.



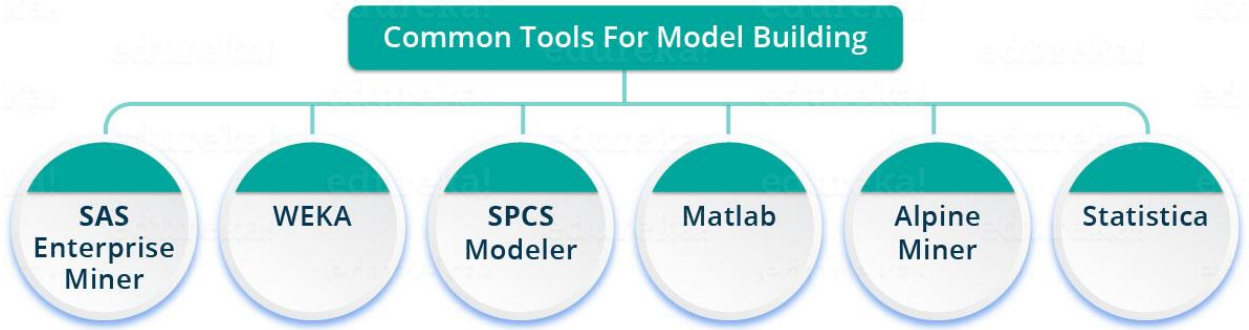
R and Python have a complete set of modeling capabilities and provide good environment for building interpretive models.

SQL Analysis services can perform in-database analytics using common data mining functions and basic predictive models.

SAS/ACCESS can be used to access data from Hadoop and others are used for creating repeatable and reusable model flow diagrams.

3.4. Phase 4—Model building: In this phase, you will develop datasets for training and testing purposes. Here you need to consider whether your existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing). You will analyze various learning techniques like classification, association and clustering to build the model.

You can achieve model building through the following tools.



3.5. Phase 5—operationalize: In this phase, you deliver final reports, briefings, code and technical documents. In addition, sometimes a pilot project is also implemented in a real-time production environment. This will provide you a clear picture of the performance and other related constraints on a small scale before full deployment.

3.6. Phase 6—Communicate results: Now it is important to evaluate if you have been able to achieve your goal that you had planned in the first phase. So, in the last phase, you identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure based on the criteria developed in Phase 1.

4.0. In conclusion: Data science is referred to as the "oil of the 21st century," our digital data carries the most importance in the field. It has incalculable benefits in business, research and our everyday lives. Your route to work, your most recent Google search for the nearest coffee shop, and your Instargram post about what you ate, and even the health data from your fitness tracker are all important to different data scientists in different ways. Sifting through massive lakes of data, looking for connections and patterns, data science is responsible for bringing us new products, delivering breakthrough insights and making our lives more convenient.

References:

Bell, G.; Hey, T.; Szalay, A. (2009). "Computer Science: Beyond the Data Deluge". Science. **323** (5919): 1297–1298.

Cao, Longbing (2017). "Data Science: A Comprehensive Overview". ACM Computing Surveys. ***50*** *(3): 43:1–43:42.*

Dhar, V. (2013). "Data science and prediction". Communications of the ACM. **56** (12): 64–73. Archived from the original on 9 November 2014. Retrieved 2 September 2015.

Jeff, Leek (2013). "The key word in "Data Science" is not Data, it is Science". Simply Statistics. Archived from the original on 2 January 2014. Retrieved 1 January 2014.

Khan, S. (2020). Meta-Analysis Methods for Health and Experimental Studies. Singapore: Springer Nature.

Hayashi, Chikio (1998). "What is Data Science? Fundamental Concepts and a Heuristic Example". Data Analysis, and Knowledge Organization. Springer Japan. pp. 40–51.

Osuolale Peter Popoola and Nicholar Nsowah Nuamah (2018)."New Trend in Modelling Climate Change in the Era of Big Data" Anale. Seria Informatică. Vol. XVI fasc. 2 – 2018.

Sahinoglu, M. and Cueva-Parra, L. (2011). CLOUD computing. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(1), 47-68.

Tony, Hey; Stewart Tansley; Kristin Michele Tolle (2009). The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research. ISBN 978-0-9825442-0-4. Archived from the original on 20 March 2017.

United Nations Economic Commission for Europe (2013). *Classification of Types of Big Data.* http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data.

United Nations Global Pulse (2012). "Big Data for Development: Challenges and Opportunities