1

INTRODUCTION

Jafreezal Jaafar

To understand the concept of Big Data, one should know about Data. Data is anything that may be used by a computer, e.g., number, letter, symbol, voice, or image. Data can be stored or transferred as electrical impulses on magnetic, optical, or mechanical media. The term Big Data can be referred to as massive data collection that grows rapidly over time. Typical data management technologies cannot store or handle information effectively due to its size and complexity. Big Data is generally defined as extraordinarily huge data collections that have outgrown standard data processing technologies. Searching the web for definitions finds an almost common definition accepted by most proponents of Big Data that may be summarised as follows: big data refers to data collections that have become so large that standard information systems can no longer efficiently manage its size, scope, or expansion. In other words, the data collection has gotten too huge to manage and extract valuable information from them. There are several challenges that big data encounter nowadays. For example, the main challenges are capturing, storing, searching, sharing, analysing, and visualisation. There is a lot more to say about Big Data. With time, the term has developed to incorporate data quantity and leverage processes. Big Data is now associated with corporate intelligence, analytics, and data mining.

Big Data can be structured, unstructured, or semi-structured. Structured data is any data that can be stored, retrieved, and processed in a set manner. A database table is a basic example of structured data. Unstructured data has no recognised shape or organisation (e.g., sound or image). Aside from its enormous bulk, unstructured data faces several processing hurdles to extract value from it. Unstructured data includes basic text files, photos, videos, and other diverse data sources. Semi-structured data may comprise both types. Semi-structured data seems structured but is not specified. An XML file is an example of semi-structured data.

1.1 MAKING SENSE OF DATA

Data is everywhere. Social media, e-commerce, banking, industries, and mobile apps are all the tools that continuously generate colossal amounts of data, which keep growing with every new second. For example, on social media platforms, people discuss various things like; politics, events, sports, movies, celebrities, food, and wearing. This data is not always in text form. It can be in image, audio, or video, or it can be in numeric form. Therefore, it is tough to understand and analyse this data and make it valuable. Business companies are interested to know what people think about their products. Political parties want to know the public opinion about their way of government. In that case, they either analyse data manually using human resources or build intelligent systems to generate a sense of the raw data. Of course, this cannot be done by a human; therefore, Artificial Intelligence-based techniques are applied to such big data to convert it into useful information. As computer only understands 0's and 1's. Hence to make the computer learn from such data, data science plays its part in producing intelligent systems.

The data is pre-processed and assigned labels (0 or 1 / A or B) to predict answers to the questions. The data science approach, like supervised learning, is used to train the model, learn the trends from existing data and make predictions. Supervised learning helps answer questions like "Are the customers satisfied with the services provided?." "How will win the next general elections in Malaysia?." Most of the time, data is not as simple as directly assigning labels. To know insights into such data, one needs to answer, "How is data organised?." So, to answer this question and make sense of the data, the data is generally grouped based on similar features using unsupervised learning techniques such as clustering analysis. Customer segmentation is an example of such data. Companies need to understand various consumer groups to build marketing strategies based on the various consumer groups.

1.2 CHARACTERISTICS OF BIG DATA

The Big Data definition does not only depend on the size or volume of the data, but there are some other characteristics generally known as 8V's. The following are the characteristics of Big Data:

1. Volume: At any one moment, the quantity of data may be processed. This might take the form of an accumulation of data over time or a single, large chunk of data that must be handled all at once.
2. Velocity: Velocity is about the quickness of the Big Data processes. For instance, how quickly data comes in and how quickly the data is changed and sent out. Telemetry for a self-driving car needs to be analysed in real-time. This is an example of something that needs to happen quickly. Latency is the adversary of speed.

3. Variety: Variety refers to heterogeneous data sources that must first be recognised and standardised before the computation may take place.
4. Visualisation: Visualisation refers to the illustration of data and relationships among the data. Data must be made accessible in a comprehensible manner after it has been produced and processed.
5. Vocabulary: This term has two meanings. The first definition has to do with the language used to express the intended analysis conclusion. For example, "accuracy" and "performance" may signify different things in structural engineering and animation. The second meaning includes semantic searching and semantic space activities. Controlled vocabularies (ontologies) represent a defined meaning as well as a relationship to another word. For example, the phrase "child" implies "parent." Search, and retrieval are utilised to reveal unknown connections with customers in the artificial intelligence domain. In fact, the ontology's strength determines initiatives' relative success or failure using semantic-based technology.
6. Vagueness: Vagueness is a concept that denotes a problem with understanding the delivered findings.
7. Viability: Viability is a model's capacity to reflect reality accurately, referred to as viability.
8. Value: Value refers to consumer focus and care. In other words, we define values by understanding what is important to consumers.

1.3 OVERVIEW OF DATA SCIENCE

Given the vast volumes of data created today, data science is one of the most hotly contested subjects in Information Technology (IT) circles. It has become trendy nowadays. Almost every business is using data science to develop and improve consumer happiness.

Data science is a branch of study that works with massive amounts of data using cutting-edge tools and methodologies to uncover previously hidden patterns, extract relevant information, and make business choices. Data science is concerned with using advanced machine learning techniques to create prediction models.

Data scientists can extract usable information from organised and unstructured data using scientific approaches, processes, algorithms, and systems. They then put this knowledge to use by applying it to a broad range of real-world challenges. We can formally define data science as *"Data science is the research into the methods for extracting valuable insights from large amounts of data by combining domain knowledge with programming skills and a thorough grasp of mathematics and statistics."* Figure 1.1 shows the overview of the data science field.

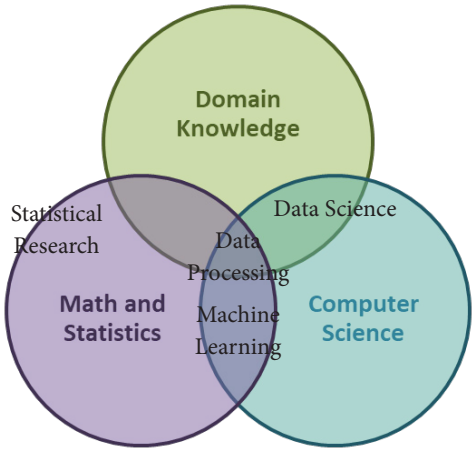

Figure 1.1 Overview of data science [1]

1.4 WHAT IS BIG DATA ANALYTICS?

Big Data requires particular techniques to collect, process, and analyse data in order to extract relevant information for decision-making purposes. "Big Data Analytics" is a term used to describe Big Data. It may also be defined as a methodology for extracting meaning from massive datasets by studying the data in its raw form. It allows us to find patterns in large amounts of unstructured data and then extrapolate useful information from them.

There is a distinct difference between Data Science and Data Analytics when it comes to handling Big Data. Data Science incorporates Data Analytics under one roof. Math, Statistics, Computer Science, Information Science, Machine Learning and Artificial Intelligence play a role in Data Science. It uses ideas like data mining, data inference, predictive modelling, and Machine Learning algorithm development to identify patterns from complicated information and translate them into effective business strategies. Statistical analysis, on the other hand, is the primary focus of data analytics. The purpose of data analytics is to identify the details of the insights that have been gleaned. To put it simply, Data Analytics is a subfield of Data Science that focuses on providing detailed solutions to the issues posed by Data Science.

1.4.1 Data Analytics Process

Before applying analytics, the data passes through following steps,

1. Determine how the data should be grouped:
   A variety of characteristics, for example, age, population, wealth, or sex, may be used to partition data. The data values can be in different formats, for instance, numerical or category.
2. Data collection:
   Data may be acquired via internet sources, computers, individuals, and community sources.
3. Organizing the data:
   After collection, data has to be arranged for analysis. Data organising may be done using a spreadsheet or other statistical software.
4. Data cleansing:
   Data cleansing is a process to detect and remove the incorrect and inaccurate data. All the data must be checked for duplicates and errors. After that, it is checked to ensure it is not missing anything. Before the data is sent to a data analyst for analysis, it is important to clean the data to repair or remove any errors.

1.4.2 Data Analytics Types

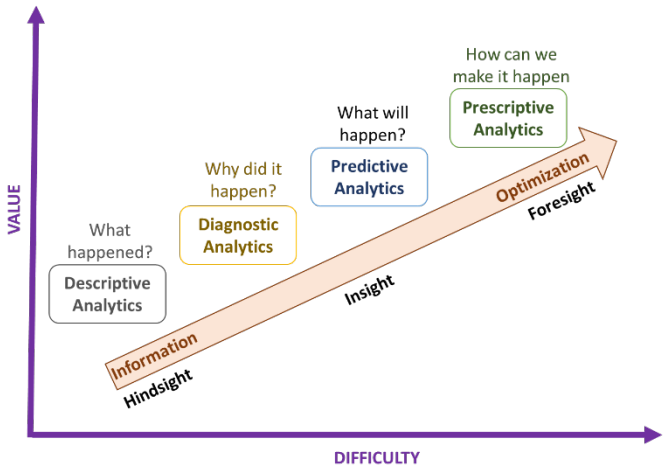

Figure 1.2 Gartner analytics ascendancy model [2]

There are four fundamental types of data analytics, as shown in Figure 1.2 [2]:

1. **Descriptive Analytics** is a basic type of analytics. It deals with "how many, what, where, or when." For instance, how things have changed over time, such as an increase or decrease in views or sales.

2. **Diagnostic Analytics** deal with events and causes. It helps us to identify the reasons why certain things happened. We may learn things like "Did the new advertising campaign effect cloths sales?" by looking at the statistics.
3. **Predictive Analytics** is concerned with occurrences that are projected to occur in the near future. It helps the data scientists to identify the trends, correlations, and causations. For example, with predictive analytics, companies may learn more about what occurred to their sales during the previous summer's scorching temperatures.
4. **Perspective Analytics** is useful to decide what actions should be taken based on the predicted outcomes. In other words, it helps us to know that how to do something. For example, based on the prediction of customer demands, prescriptive analytics can tell a manager to add an evening shift in production to avoid the supply delays.

1.5 WHO IS DATA SCIENTIST?

A Data Scientist is someone who cleans, processes, and interprets data using a mix of Mathematical, Statistical, and Machine Learning methodologies. Data Scientists create sophisticated data modelling methods using prototypes, machine learning algorithms, predictive models, and bespoke (or customised) analyses. Therefore, Data Scientist performs the following activities:

1. To process, clean, and validate the integrity of data.
2. To undertake Exploratory Data Analysis on big datasets in order to learn more about them.
3. To perform data mining by creating Extract, Transform, and Load (ETL) pipelines.
4. To conduct statistical analysis using Machine Learning (ML) techniques.
5. To write code for automation and build resourceful ML libraries.
6. To get business insights using machine learning technologies and algorithms.
7. To uncover new patterns in data that may be used to make business forecasts.

1.6 WHY BIG DATA ANALYTICS

Using big data analytics only for experimentation is no longer the norm in today's world. Many organisations started to see genuine benefits with this approach, and they are now increasing their efforts to use more data and models in their strategies. Customers, sales, and transaction data may be used to make more accurate judgments if they are combined with other information, such as social media data such as stock prices and weather forecasts. In order to get the most out of their data, firms might employ big data analytics. Smarter business decisions and more efficient operations are a result of this. Big Data analytics has spread to all disciplines. As a consequence, Big Data is used in many areas, including finance, healthcare, education, government, retail, manufacturing, and others. For example, Amazon, Netflix, Spotify, LinkedIn, Swiggy, and more employ big data analytics.

Big Data analytics helps conventional and online merchants identify client behaviour and propose interest-based items. This aids in the development of new and enhanced items for the organisation.

1. Cost savings: Data storage costs may be reduced using Big Data technologies like Apache, Hadoop, Spark, etc. These technologies aid businesses in discovering new and better methods to conduct their operations.
2. Time saving: Firms may gather data from various sources using real-time in-memory analytics. They may use Hadoop and other tools to evaluate data quickly and make choices based on what they have learned.
3. Understand the market conditions: Understanding market conditions is easier with the aid of Big Data analysis. For example, it is possible to determine the most popular items by analysing customers' purchase habits. As a result, businesses are able to gain an advantage over their rivals.
4. Social media listening: A company's Big Data technologies may be used to analyse sentiment. For example, they can see who is talking about their firm and what they are saying about it.
5. Boost customer acquisition and retention: Customers are a company's most valuable assets. A strong client base is essential to the success of any organisation. It is impossible to ignore competition in the market even if you have a stable consumer base. The success of businesses will suffer if we do not know what our consumers desire. It will have a negative impact on company development as a consequence of the loss of customers. Customer-related trends and patterns may be discovered using big data analytics. A successful company is a result of analysing customer behaviour.
6. Solve advertiser's problem and offer marketing insights: Big data analytics influences all corporate activities. Customer expectations may be met thanks to this technology. Using big data analytics aids in developing new products for the firm. As a result, successful advertising initiatives are ensured.
7. The driver of Innovations and Product Development: The ability to innovate and reconstruct goods is a key benefit of big data for businesses.

1.7 ISSUES AND CHALLENGES OF BIG DATA ANALYTICS

Analytical techniques are incredibly crucial for businesses nowadays. Businesses use data analytics to enhance decision-making, promote accountability, benefit financial health, and assist workers in predicting losses and monitoring their own performances. Obtaining these advantages, on the other hand, is a lot easier said than done. There are a variety of difficulties that businesses encounter throughout the process of data collecting and the subsequent application of analytics to it. The following are some of the most significant obstacles to overcome:

1. Amount of data
   The volume of data generated by today's data-driven enterprises and big data sometimes overwhelms risk managers and other workers. Every day, an organisation may get data on hundreds of incidents and interactions, presenting analysts with thousands of interlocking data sets.
2. Collecting meaningful and real-time data
   Because so much information is accessible, it is tough to sift through it all to find the nuggets of value. Data analysis might be compromised when personnel are overworked and only concentrate on the simplest to acquire metrics rather than those which bring real value. As a result, it is hard for an employee to acquire real-time insights if they must manually filter through data. Decision-making may suffer greatly from outdated facts.
3. Data from multiple sources
   The next difficulty arises when attempting to examine data from various unconnected sources. Often, disparate bits of data are held in disparate systems. Employees may be unaware of this, resulting in insufficient or wrong analysis. Manually integrating data is time intensive and may restrict insights to those that are readily accessible.
4. Visual representation of data
   Data typically requires visual representation in graphs or charts to be easily understood and remembered. It is impossible to construct these tools manually, even if they are valuable. The tedious and time-consuming process of sifting through data from many sources and putting it into a reporting tool is irritating.
5. Inaccessible data
   Transferring data to a centralised system has little effect if it is not freely available to those needing it. Decision-makers and risk managers must access all of an organisation's data in order to get insight into what is occurring at any given time, even if they are working remotely. Access to data should be the simplest aspect of data analytics.
6. Poor quality data
   Inaccurate data may have a devastating effect on data analytics. The quality of the input determines the quality of the result. Human mistakes often cause inaccurate data during data input. Any judgments that are based on this research will almost certainly have unfavorable outcomes. Furthermore, asymmetric data: when one system does not keep up with the changes made in another system, making the data obsolete.

1.8 BIG DATA ANALYTIC LIFECYCLE

Big data lifecycle can also be known as a big information lifecycle. It refers to the entire period of time from where the data get existence until the visualisation [3]. Figure 1.3 shows steps of Big Data lifecycle:

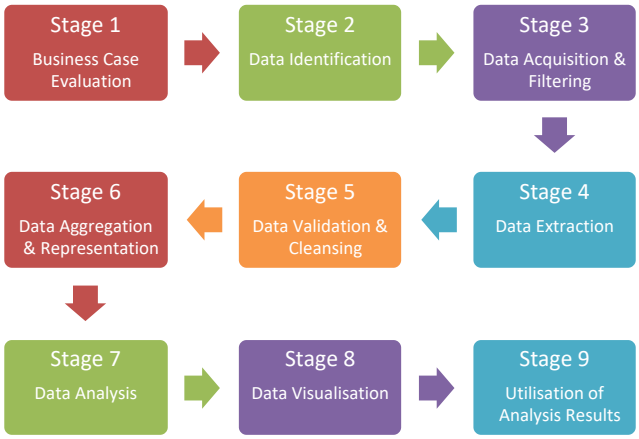

Figure 1.3 Big Data analytics life cycle [3]

1. Business case evaluation
   The business case justifies the project and specifies its specifications and management considerations. The Big Data Lifecycle begins with a solid business case review. Before any Big Data project can begin, the business goals and expected outcomes must be defined. Begin with the end in mind and clearly describe the project's goals.
2. Data identification
   In order to identify a person whose data is being matched with publicly accessible information or auxiliary data, data identification is used. It is crucial to know where the data will come from. Identifying the original data source and the level of reliability of a dataset are especially important when it comes to data obtained from third parties.
3. Data acquisition and filtering
   Analog waveforms are generally converted to digital values for processing using DAS or DAQ acronyms, which stand for Data Acquisition and Filtering. Sensors are used to translate physical properties into electrical signals in data collecting systems. Big Data Lifecycle Stage 2 is a continuation of Stage 1. Information from both within and outside of the firm may be obtained at this point. The first stage in filtering is to remove faulty data once it has been acquired.
4. Data extraction
   Data extraction is the act or process of obtaining information from various sources for subsequent processing or storage. If a Big Data tool is used, it significantly impacts

the complexity and scope of the data transformation. Most contemporary Big Data technologies are able to read relational and non-relational data in the industry-standard formats.

5. Data validation and cleansing
   It is the process of discovering and fixing faulty or erroneous records from a record set, table, or database. Incorrect data leads to invalid outcomes. The Big Data Lifecycle requires the Data Validation and Cleansing step to guarantee that only relevant data is evaluated. During this step, data is verified against a set of specified requirements and rules.
6. Data aggregation and representation
   Data may be scattered over different datasets, needing common attributes like date or ID to link them together. Some data variables, like date of birth, occur in numerous databases. Either a data reconciliation procedure or a dataset providing the right value is needed.
7. Data analysis
   It involves analysing, cleaning, manipulating, and modelling data to find relevant information, draw conclusions, and help decision-making. Depending on the analysis type, data analysis may be simple or complicated. This stage generates the project's "real value." Less meticulous execution of preceding phases will provide factual and accurate outcomes.
8. Data visualisation
   It is an endeavour to help people grasp the value of data by visualising it. Data visualisation tools may help reveal hidden patterns, trends, and correlations in text-based data. The data visualisation stage uses data visualisation methods and technologies to visually convey analytical findings to business users. For example, data points are often plotted as charts or heat maps.
9. Utilisation of analysis
   Results The Big Data Lifecycle concludes with putting the findings into practice. It examines how and where the Big Data Project's outcomes may be used.

1.9 SUMMARY

This chapter briefly introduces what Big Data is and what are its characters. It further discusses about how to analyse Big Data using Big Data Analytics and the role of Data Science. It also briefly described the responsibilities of Data Analyst, followed by the issues and challenges of Big Data. It guides how to make sense out of the Big Data. Finally, it describes the life cycle of Big Data Analytics.