

1

Background of Data Science

1.1 Introduction

Data science is one of the most promising and high-demand career paths for skilled professionals in the 21st century. Currently, successful data professionals understand that they must advance past the traditional skills of analyzing large amounts of data, statistical learning, and programming skills. In order to explore and discover useful information for their companies or organizations, data scientists must have a good grip of the full spectrum of the data science life cycle and have a level of flexibility and understanding to maximize returns at each phase of the process.

Data science is a “concept to unify statistics, mathematics, computer science, data analysis, machine learning and their related methods” in order to find trends, understand, and analyze actual phenomena with data. Due to the Coronavirus disease (COVID-19) many colleges, institutions, and large organizations asked their nonessential employees to work virtually. The virtual meetings have provided colleges and companies with plenty of data. Some aspect of the data suggest that virtual fatigue is on the rise. Virtual fatigue is defined as the burnout associated with the over dependence on virtual platforms for communication. Data science provides tools to explore and reveal the best and worst aspects of virtual work.

In the past decade, data scientists have become necessary assets and are present in almost all institutions and organizations. These professionals are data-driven individuals with high-level technical skills who are capable of building complex quantitative algorithms to organize and synthesize large amounts of information used to answer questions and drive strategy in their organization. This is coupled with the experience in communication and leadership needed to deliver tangible results to various stakeholders across an organization or business.

Data scientists need to be curious and result-oriented, with good knowledge (domain specific) and communication skills that allow them to explain very technical results to their nontechnical counterparts. They possess a strong quantitative background in statistics and mathematics as well as programming knowledge with

focuses in data warehousing, mining, and modeling to build and analyze algorithms. In fact, data scientists are a group of analytical data expert who have the technical skills to solve complex problems and the curiosity to explore how problems need to be solved.

1.2 Origin of Data Science

Data scientists are part mathematicians, statisticians and computer scientists. And because they span both the business and information technology (IT) worlds, they're in high demand and well-paid. Data scientists were not very popular some decades ago; however, their sudden popularity reflects how businesses now think about "Big data." Big data is defined as a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. That bulky mass of unstructured information can no longer be ignored and forgotten. It is a virtual gold mine that helps boost revenue as long as there is someone who explores and discovers business insights that no one thought to look for before. Many data scientists began their careers as statisticians or business analyst or data analysts. However, as big data began to grow and evolve, those roles evolved as well. Data is no longer just an add on for IT to handle. It is vital information that requires analysis, creative curiosity, and the ability to interpret high-tech ideas into innovative ways to make profit and to help practitioners make informed decisions.

1.3 Who is a Data Scientist?

The term "data scientist" was invented as recently as 2008 when companies realized the need for data professionals who are skilled in organizing and analyzing massive amounts of data. Data scientists are quantitative and analytical data experts who utilize their skills in both technology and social science to find trends and manage the data around them. With the growth of big data integration in business, they have evolved at the forefront of the data revolution. They are part mathematicians, statisticians, computer programmers, and analysts who are equipped with a diverse and wide-ranging skill set, balancing knowledge in several computer programming languages with advanced experience in statistical learning and data visualization.

There is not a definitive job description when it comes to a data scientist role. However, we outline here some stuffs they do:

- Collecting and recording large amounts of unruly data and transforming it into a more usable format.

- Solving business-related problems using data-driven techniques.
- Working with a variety of programming languages, including SAS, Minitab, R, and Python.
- Having a strong background of mathematics and statistics including statistical tests and distributions.
- Staying on top of quantitative and analytical techniques such as machine learning, deep learning, and text analytics.
- Communicating and collaborating with both IT and business.
- Looking for order and patterns in data, as well as spotting trends that enables businesses to make informed decisions.

Some of the useful tools that every data scientist or practitioner needs are outlined below:

- *Data preparation*: The process of cleaning and transforming raw data into suitable formats prior to processing and analysis.
- *Data visualization*: The presentation of data in a pictorial or graphical format so it can be easily analyzed.
- *Statistical learning or Machine learning*: A branch of artificial intelligence based on mathematical algorithms and automation. Artificial intelligence (AI) refers to the process of building smart machines capable of performing tasks that typically require human intelligence. They are designed to make decisions, often using real-time data. Real-time data are information that is passed along to the end user immediately it is gathered.
- *Deep learning*: An area of statistical learning research that uses data to model complex abstractions.
- *Pattern recognition*: Technology that recognizes patterns in data (often used interchangeably with machine learning).
- *Text analytics*: The process of examining unstructured data and drawing meaning out of written communication.

We will discuss all the above tools in details in this book. There are several scientific and programming skills that every data scientist should have. They must be able to utilize key technical tools and skills, including R, Python, SAS, SQL, Tableau, and several others. Due to the ever growing technology, data scientist must always learn new and emerging techniques to stay on top of their game. We will discuss the R and Python programming in Chapters 5 and 6.

1.4 Big Data

Big data is a term applied to ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by classical data-processing tools. In particular, it refers to data sets whose

size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low latency. Sources of big data includes data from sensors, stock market, devices, video/audio, networks, log files, transactional applications, web, and social media and much of it generated in real time and at a very large scale.

In recent times, the use of the term “big data” (both stored and real-time) tend to refer to the use of user behavior analytics (UBA), predictive analytics, or certain other advanced data analytics methods that extract value from data. UBA solutions look at patterns of human behavior, and then apply algorithms and statistical analysis to detect meaningful anomalies from those patterns’ anomalies that indicate potential threats. For example detection of hackers, detection of insider threats, targeted attacks, financial fraud, and several others.

Predictive analytics deals with the process of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. Generally, predictive analytics does not tell you what will happen in the future. However, it forecasts what might happen in the future with some degree of certainty. Predictive analytics goes hand in hand with big data: Businesses and organizations collect large amounts of real-time customer data and predictive analytics and uses this historical data, combined with customer insight, to forecast future events. Predictive analytics helps organizations to use big data to move from a historical view to a forward-looking perspective of the customer. In this book, we will discuss several methods for analyzing big data.

1.4.1 Characteristics of Big Data

Big data has one or more of the following characteristics: high volume, high velocity, high variety, and high veracity. That is, the data sets are characterized by huge amounts (volume) of frequently updated data (velocity) in various types, such as numeric, textual, audio, images and videos (variety), with high quality (veracity). We briefly discuss each in detail. **Volume:** Volume describes the quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. **Velocity:** Velocity describes the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in both stored and real-time. Compared to small data, big data are produced more continually (it could be nanosecond, second, minute, hours, etc.). Two types of velocity related to big data are the frequency of generation and the frequency of handling, recording, and reporting. **Variety:** Variety describes the type and formats of the data. This helps people who analyze it to effectively use the resulting insight. Big data draws from different formats and completes missing pieces through data fusion. Data fusion is a term used to describe the technique of integrating multiple data sources to produce more consistent,

accurate, and useful information than that provided by any individual data source. **Veracity:** Veracity describes the quality of data and the data value. The quality of data obtained can greatly affect the accuracy of the analyzed results. In the next subsection we will discuss some big data architectures. A comprehensive study of this topic can be found in the application architecture guide of the Microsoft technical documentation.

1.4.2 Big Data Architectures

Big data architectures are designed to handle the ingestion, processing, and analysis of data that is too large or complex for classical data-processing application tools. Some popular big data architectures are the Lambda architecture, Kappa architecture and the Internet of Things (IoT). We refer the reader to the Microsoft technical documentation on Big data architectures for a detailed discussion on the different architectures. Almost all big data architectures include all or some of the following components:

- *Data sources:* All big data solutions begin with one or more data sources. Some common data sources includes the following: Application data stores such as relational databases, static files produced by applications such as web server log files, and real-time data sources such as the Internet of Things (IoT) devices.
- *Data storage:* Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats. This kind of store is often called a data lake. A data lake is a storage repository that allows one to store structured and unstructured data at any scale until it is needed.
- *Batch processing:* Since data sets are enormous, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis. Normally, these jobs involve reading source files, processing them, and writing the output to new files. Options include running U-SQL jobs or using Java, Scala, R, or Python programs. U-SQL is a data processing language that merges the benefits of SQL with the expressive power of ones own code.
- *Real-time message ingestion:* If the solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing. This might be a simple data store, where incoming messages are stored into a folder for processing. However, many solutions need a message ingestion store to act as a buffer for messages and to support scale-out processing, reliable delivery, and other message queuing semantics.
- *Stream processing:* After obtaining real-time messages, the solution must process them by filtering, aggregating, and preparing the data for analysis. The processed stream data is then written to an output sink.

- *Analytical data store*: Several big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools. The analytical data store used to serve these queries can be a Kimball-style relational data warehouse, as observed in most classical business intelligence (BI) solutions. Alternatively, the data could be presented through a low-latency NoSQL technology, such as HBase, or an interactive Hive database that provides a metadata abstraction over data files in the distributed data store.
- *Analysis and reporting*: The goal of most big data solutions is to provide insights into the data through analysis and reporting. Users can analyze the data using mathematical and statistical models as well using data visualization techniques. Analysis and reporting can also take the form of interactive data exploration by data scientists or data analysts.
- *Orchestration*: Several big data solutions consist of repeated data processing operations, encapsulated in workflows, that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or move the results to a report or dashboard.