Module 1

ENTERPRISE BIG DATA PROFESSIONAL

An introduction to big data and data science for the enterprise

ENTERPRISE BIG DATA FRAMEWORK®

APMG International

Enterprise Big Data Professional

An Introduction to Big Data and Data Science for the Enterprise

Jan-Willem Middelburg

Contents

Colophon

Purpose

This handbook, Enterprise Big Data Professional, is based on a subset of the Enterprise Big Data Framework. The handbook is intended as the official reference guide for students aiming to sit for the APMG Enterprise Big Data Professional (EBDP) examination. Additionally, the materials in this guide provide introductory guidance to anyone who wishes to learn more about Data Science, Data Analysis and Big Data. Readers wishing to know more about the Enterprise Big Data certification scheme are invited to visit the website* of the Enterprise Big Data Framework Alliance.

Important Note

The Enterprise Big Data Framework education and certification scheme has been developed jointly with APMG-International. Only Accredited Training Organizations (ATOs) or their affiliates are allowed to provide courses and examinations in APMG-International's qualification schemes. Readers wishing to know more about APMG-International or the examination requirements are invited to visit the APMG website†.

Trademarks

Figure i: The Enterprise Big Data Framework badges

*https://www.bigdataframework.org
†https://www.apmg-international.com

About the Enterprise Big Data Framework Alliance

The Enterprise Big Data Framework Alliance (EBDFA) is an independent body of knowledge for the development and advancement of big data best practices. The Enterprise Big Data Framework Alliance aims to inspire, promote, and develop excellence in big data practices and applications across the globe, leading to outstanding business value through data analysis, machine learning and big data for every individual member of the community.

The Enterprise Big Data Framework education and certification scheme is a vendor-neutral program that strives to lay the groundwork for any professional in the big data industry. The curriculum has been developed to provide in-demand role-based qualifications, and to set a global standard for excellence in Big Data and Data Science.

The Big Data Certification Scheme

The Enterprise Big Data Framework education and certification scheme consists of five certifications for different role-based specialisations, and discusses fundamental knowledge, concepts, and techniques of big data environments. The comprehensive curriculum enables professionals to obtain real-world proficiency in big data techniques, processes and practices, and educates practitioners about the possibilities of big data.

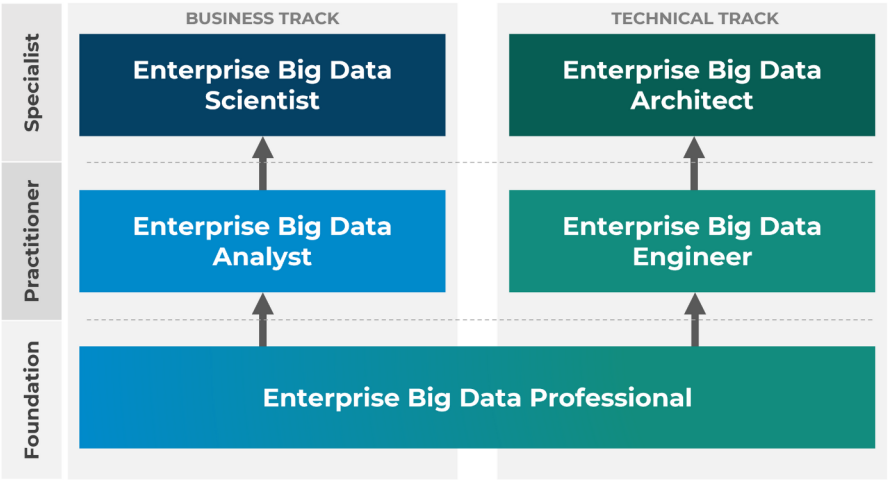

Figure ii: The Enterprise Big Data Framework certification scheme

All the information, guides and study materials of the Enterprise Big Data Framework are made available for free to the community to enhance the knowledge and application of Big Data.

This guide *Enterprise Big Data Professional* contains all the theory and content to pass the APMG-International Big Data Professional examination.

Forewords

Foreword by Richard Pharro

Why do we need guidance, training and qualifications on Big Data?

According to research, big data has the ability to transform virtually every organization. A report by McKinsey suggests that governments could save more than $150 million in operational efficiencies. This is a compelling reason for every manager to have a better understanding of the impact it will have on their industry, market and business.

Each year we are producing a phenomenal amount of data. Without skilled people, many organizations cannot appreciate the value of their data and more importantly how to access it, so big data analysts can unlock significant value and it is likely to become a clear competitive advantage in the very near future as existing organizations develop strategies to innovate and compete by leveraging both historic and real time information.

By using this information, we are likely to see productivity benefits and be able to deliver more cost effective and high-quality services to our customers. However, to achieve this organizations will need access to skilled talent. Like many other areas in the digital economy, competent people are hard to find. Research suggests that in the United States alone there is a shortage of nearly 200,000 people with the appropriate analytical skills and it's suggested that between 1-2 million managers and analysts lack the knowhow to use the information that can be generated to make effective decisions.

Using Big Data won't be easy. With the growth of regulations regarding cyber security, use of intellectual property and personal privacy, there are some very challenging issues to be addressed. However, without the appropriate people and technology and an understanding of how the enterprise works, the information will remain unused and it is likely that nimbler and more efficient organizations will capture market share.

This guidance and the training and certification programs that support it, are aimed at helping organizations understand both the value of their data and the opportunities they offer, together with the need to educate and develop their managers and leaders throughout the business in order to maximize returns. This applies to commercial companies, government and not for profit organizations.

You have the data; the question is do you know how to maximize its use?

This guide will help.

Richard Pharro
CEO, APMG-International

Foreword by Jan-Willem Middelburg

The speed with which data is being created has never been as great as in the 21st century. All experts agree that 98% of data was created in the 21st century, and that roughly 80% of that data has never been analysed at all.

All this data contains a wealth of information that could lead to a brighter, more sustainable future. However, in order to retrieve value from these enormous data sets, we will need to have many qualified data analysts, data scientists and data professionals. People who, when confronted with a problem can use a variety of different approaches and techniques to solve a problem. People who have a fundamental understanding of data structures and algorithms. And, above all else, people who are able to think creatively and help solve the world's most pressing problems.

The very nature of data - and the beauty of it - is that it is universal and factual. Data created in China can be analysed with the same approaches, techniques and tools as data that was created in Brazil, or any other part of the world. No matter in which industry data was created, data problems are similar. Classification algorithms can be used to predict credit card fraud in the finance industry. Yet these same classification algorithms can be used to find diseases in photographs in medical domains. Yet in order to understand this, you will require knowledge about classification algorithms in the first place.

And this is exactly why the Enterprise Big Data Framework was created. With the Big Data Framework, we aim to bring a structured approach to enterprise organizations in order to obtain value from massive quantities of data. We found that - although there are many good programs available around the world – these programs are mainly vendor driven or rooted strongly in the academic domain. The enterprise domain, however, is different. In enterprise organizations, the return on investment matters. Enterprises cannot be dependent on single individuals, and they need to comply to local rules and regulations. In order to systematically retrieve value from data, we need structure. We need an Enterprise Big Data Framework.

The Enterprise Big Data Frameworks was created by Big Data Professionals for Big Data Professionals. We aspire to make knowledge about Big Data practices and Data Science solutions available to everyone in the world. This is why most of the materials to get started are provided for free. Everyone who has helped put the Enterprise Big Data Framework together believes in this important mission.

Together with APMG-International, we have created the Enterprise Big Data certification scheme, that provides a fast track for anyone who wants to learn more about Big Data and Data Science techniques. In creating this certification scheme, we hope to inspire corporations and individual to start analysing data. To open pandora's box and make better decisions towards the future.

Jan-Willem Middelburg
Lead Author, The Enterprise Big Data Framework

1. Introduction to Big Data

1.1 What is Big Data?

Big Data and its analysis techniques are at the centre of modern science and business. Every day, millions of transactions, emails, photos, video files, posts and search queries are generated that result in terabytes of data. All that data is stored in databases on various places across the planet.

All that data potentially contains a wealth of information. By analysing the data that is generated every day, governments, researchers and companies might discover knowledge that they could use to their benefit. For governments, this might be to prevent tax fraud or to increase the economic interests of the nation. For researchers, discoveries in data might help to develop new medicines. And for companies, data analysis might help determine the best location to open a new store in order to obtain a competitive advantage. Although these are different examples, and the value of the information is different for every type of organization, the process to extract insights out of data is very similar.

Extracting valuable knowledge out of massive quantities of data is, however, more difficult than it sounds. Due to the sheer volume of data that is generated every day, databases grow towards billions of records, and data analysis becomes increasingly more difficult. Specifically, the larger the data sets become that any organization produces, the more difficult it becomes to capture, store, manage, share, analyse and visualize data sets.[1] For this specific reason, the knowledge and skills that are required to translate data into valuable information has become a domain of growing importance. This domain, which includes the skills, technology and techniques required to analyse large data sets, is collectively referred to as "Big Data."

Although the importance of Big Data has been recognized over the last decade, people and organizations still uphold different opinions on its definition.[2] In general, the term Big Data is used when data sets cannot be managed or processed by traditional commercial software (and its underlying IT infrastructure) within a tolerable amount of time. The domain of Big Data is however more all-encompassing than the speed of data transfer, or its required technology, tools, or processes. Over time, Big Data has gradually evolved in an entire domain of study that interfaces with data science, machine learning and artificial intelligence.

Although there are many good definitions of Big Data, this guide will use a definition that focuses on Big Data as a knowledge domain. In the rest of this guide, we will therefore adhere to the following definition of Big Data:

> **Big Data is the knowledge domain that explores the techniques, skills and technology to deduce valuable insights out of massive quantities of data.**

The objective of this guide is to discuss these techniques, skills and technologies in a structured approach. With the Enterprise Big Data Framework, we aim to equip every reader with the knowledge and skills to deduce valuable insights out of massive quantities of data. These skills will empower you to obtain fact-based, data-driven information in order to support your future decisions. In order to accomplish this goal, this first chapter will introduce some fundamental concepts and terminology about data, data structures and the characteristics of Big Data. In chapter 2, we will subsequently introduce the Enterprise Big Data Framework, a holistic model of six capabilities to increase Big Data proficiency in enterprises. Each of the remaining sections will subsequently build upon the capabilities of the Enterprise Big Data Framework.

1.2 Value of Big Data

The primary reason why Big Data has developed rapidly over the last years is because it provides long-term enterprise value. This value is captured by organizations through revenue expansion, cost reductions and increased profit margins. As a result, Big Data provides companies with the opportunity to use their enterprise data as a competitive advantage. Due to its wide range of applications, Big Data is embraced by all types of industries, ranging from healthcare, finance and insurance, to the academic and non-profit sectors.

There are various ways in which Big Data skills and technology can help enterprises to capture value. These organizations use data analysis tools and techniques to increase corporate performance and facilitate growth. For most enterprises, the quest to become 'data-driven' is coupled to a 'digital transformation' program and is supported by senior leadership. The realization that data and technology can help to obtain a competitive advantage, is a key business driver for most organizations. However, most organizations have found that this is not an easy task, and that digital transformation requires profound changes to the way the organization need to be designed, organized and managed.

The possibilities of Big Data continue to evolve rapidly, driven by innovation and the reduced cost of data storage and processing capabilities in organizations. In general terms, most organizations leverage their data to create value in any of the following 5 ways:[3]

Creating Transparancy

Making Big Data more easily accessible across the enterprise in a timely way creates tremendous value. In most organizations, information is not easily accessible across different functional departments, resulting in siloed insights and decision-making. Through increasing transparency and sharing Big Data across departments, organizations can improve performance, reduce the amount of work that is done repetitiously across multiple departments and identify inefficiencies.

In many enterprise organizations, major operational improvements can be made when information is more easily available. For example, integrating data from Reasearch & Development, engineering, and manufacturing units, potentially across several enterprises, can enable concurrent engineering

techniques that can greatly reduce the waste caused from having to rework designs, and thus accelerate time to market.

Data-Driven Discovery

Through Internet-of-Things (IoT) technologies, more and more products contain sensors, which capture data about enterprise processes, customer behaviours and the use of products and services. By analysing the data that is generated by these sensors, companies can alter their decision-making processes, and adjust their service offerings. Through Big Data, companies will know exactly how their products will flow through their supply chain, making it possible to plan improvement projects accordingly. Additionally, Big Data offers new insights in the way customers are using products and services, offering companies insights that might not have been identified previously. In the insurance industry, for example, Big Data can help to determine profitable products and provide improved ways to calculate insurance premiums based on previously submitted claims of customers.

Customer Segmentation and Customized Marketing

Customer segmentation and targeted marketing are concepts which have been widely adopted by enterprises who sell products or services to consumers. Big Data brings customer segmentation and customized marketing to entirely new levels, providing improved opportunities to customize product-market offerings to specific segments of customers.

Data about user or customer behaviour makes it possible to build different customer profiles that can be targeted accordingly. Data that is captured on social media enhances these capabilities, enabling companies to target products and services to highly targeted customer profiles. Location based data, which can be captured through Bluetooth or GPS, adds a completely new dimension to targeted marketing and will provide a further focus on Big Data practices.

Support Decisions with Automated Algorithms

Data-driven decision making is one of the key drivers for Big Data. Through analytics and algorithms, decision-making can be significantly improved. Organizations that utilize Big Data techniques can discover patterns, detect anomalies and minimize risk. Through Big Data algorithms (which we will further discuss in Chapter 5), organizations can automate processes that will lead to more accurate decisions.

In the banking industry, for example, Big Data algorithms can help bank employees to minimize risk when offering financial products to their customers. Similarly, the accounting industry can use Big Data algorithms to detect anomalies in audits or highlight cases that need further inspection. With the assistance of algorithms, people can make more accurate decisions supported by their enterprise data.

Product Development and Innovation

Big Data can unearth patterns that identify the need of new products or increase the design of current products or services. Product development and Innovation is tightly coupled with Data Driven discovery and can help companies discover new business opportunities. Through the analysis of search queries, product usage data and user-experience metrics, organizations can identify demand for products that the organization was previously unaware of. Universities or colleges, for example, might study their website traffic and search volumes to forecast class enrolment and allocate teaching resources accordingly.

Besides the five ways to capture value from Big Data as outlined above, there are many other potential business gains or ways to capture value with Big Data. Many examples and business cases in this area already exist and more are designed almost every day. Because there are so many potential ways to capture value from Big Data, it is of paramount importance that the organization has a clear plan to capture value from its Big Data initiatives. In Chapter 3, we will therefore further discuss how to formulate a Big Data strategy.

1.3 A Short History of Big Data

The term 'Big Data' has been in use since the early 1990s. Although it is not exactly known who first used the term, most people credit John R. Mashey (who at the time worked at Silicon Graphics) for making the term popular.[4] Big Data is now a well-established knowledge domain, both in academics as well as in industry.

To best understand how Big Data was able to grow to such popularity, it is important to place Big Data into its historic perspective. From a knowledge domain perspective, Big Data is the combination of the very mature domain of statistics with the relatively young domain of computer science. As such, it builds upon the collective knowledge of mathematics, statistics, and data analysis techniques in general.

Ever since the early beginnings of civilization, people have tried to use 'data' towards better decision making, or to gain a competitive (or military) advantage. This quest can even be dated back to the ancient Egyptians and the Roman Empire. The famous Library of Alexandria, which was established around 300 B.C., can be considered as a first attempt by the ancient Egyptians to capture all 'data' within the empire. It is estimated that the library consisted of 40,000 to 400,000 scrolls (which would be the equivalent of around 100,000 books).[5] Even the ancient leaders of the world realized that combining different data sources could result in an advantage over other competing empires.

Other well documented use cases of the first forms of data analysis come from the Roman empire. The ancient Roman military utilized very detailed statistical analysis to 'predict' at which border the chance of an enemy insurgency would be the most prevalent. Based on these analyses, they were able to deploy their armies in the most efficient way possible. It is not a far stretch to consider these calculations one of the earliest forms of 'predictive' data analysis. And again, these analysis techniques provided the Roman military with an advantage over other armies.

To understand the world of Big Data, it is therefore important to realize that most techniques that are used today (from predictive algorithms to classification techniques) have been developed centuries ago, and that Big Data continues to build on the work of some of the greatest minds in history. The key aspect that has changed, of course, is the availability and accessibility to massive quantities of data. Whereas up until the 1950s, most data analysis was done manually and on paper, we now have the technology and capability to analyse terabytes of data within split seconds.

Especially since the beginning of the 21st century, the volume and speed with which data is generated has changed beyond measures of human comprehension. The total amount of data in the world was 4.4 zettabytes in 2013. That is set to rise steeply to 44 zettabytes by 2020.[6] To put that in perspective, 44 zettabytes are the equivalent to 44 trillion gigabytes. Even with the most advanced technologies today, it is impossible to analyse all this data. The need to process these increasingly larger (and unstructured) data sets is how traditional data analysis transformed into 'Big Data' in the last decade.

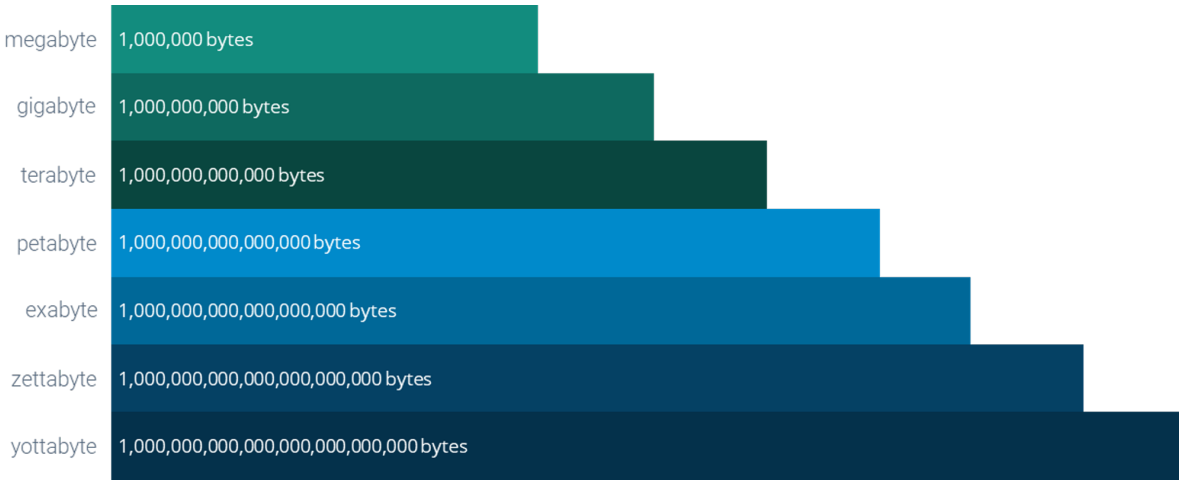

Figure 1.1: Data volumes and Big Data terminology

The evolution of Big Data can roughly be subdivided into three main phases.[7] Each phase was driven by technological advancements and has its own characteristics and capabilities. To understand the context of Big Data today, it is important to understand how each of these phases contributed to the modern meaning of Big Data.

Big Data Phase 1 - Structured Content

Data analysis, data analytics and Big Data originate from the longstanding domain of database management. It relies heavily on the storage, extraction, and optimization techniques that are common in data that is stored in Relational Database Management Systems (RDBMS). The techniques that are used in these systems, such as structured query language (SQL) and the extraction, transformation, and loading (ETL) of data, started to professionalize in the 1970s.

Database management and data warehousing systems are still fundamental components of modern-day Big Data solutions. The ability to quickly store and retrieve data from databases or find information in large data sets, is still a core requirement for the analysis of Big Data. Relational

database management technology and other data processing technologies that were developed during this phase, are still strongly embedded in the Big Data solutions from leading IT vendors, such as Microsoft, Google and Amazon. A number of core technologies and characteristics of this first phase in the evolution of Big Data is outlined in figure 1.3.

Big Data Phase 2 - Web-based Unstructured Content

From the early 2000s, the internet and corresponding web applications started to generate tremendous amounts of data. In addition to the data that these web applications stored in relational databases, IP-specific search and interaction logs started to generate web based unstructured data. These unstructured data sources provided organizations with a new form of knowledge: insights into the needs and behaviours of internet users. With the expansion of web traffic and online stores, companies such as Yahoo, Amazon and eBay started to analyse customer behaviour by analysing click-rates, IP-specific location data and search logs, opening a whole new world of possibilities.

From a technical point of view, HTTP-based web traffic introduced a massive increase in semi-structured and unstructured data (further discussed in Chapter 1.6. Besides the standard structured data types, organizations now needed to find new approaches and storage solutions to deal with these new data types to analyse them effectively. The arrival and growth of social media data greatly aggravated the need for tools, technologies and analytics techniques that were able to extract meaningful information out of this unstructured data. New technologies, such as networks analysis, web-mining and spatial-temporal analysis, were specifically developed to analyse these large quantities of web based unstructured data effectively.

Big Data Phase 3 - Mobile and Sensor-based Content

The third and current phase in the evolution of Big Data is driven by the rapid adoption of mobile technology and devices, and the data they generate. The number of mobile devices and tablets surpassed the number of laptops and PCs for the first time in 2011.[8] In 2020, there are an estimated 10 billion devices that are connected to the internet. And all of these devices generate data every single second of the day.

Mobile devices not only give the possibility to analyse behavioural data (such as clicks and search queries), but they also provide the opportunity to store and analyse location-based GPS data. Through these mobile devices and tablets, it is possible to track movement, analyse physical behaviour and even health-related data (for example the number of steps you take per day). And because these devices are connected to the internet almost every single moment, the data that these devices generate provide a real-time and unprecedented picture of people's behaviour.

Simultaneously, the rise of sensor-based internet-enabled devices is increasing the creation of data to even greater volumes. Famously coined the 'Internet of Things' (IoT), millions of new TVs, thermostats, wearables and even refrigerators are connected to the internet every single day, providing massive additional data sets. Since this development is not expected to stop anytime soon, it could be stated that the race to extract meaningful and valuable information out of these new data

sources has only just begun. A summary of the evolution of Big Data and its key characteristics per phase is outlined in figure 1.2.

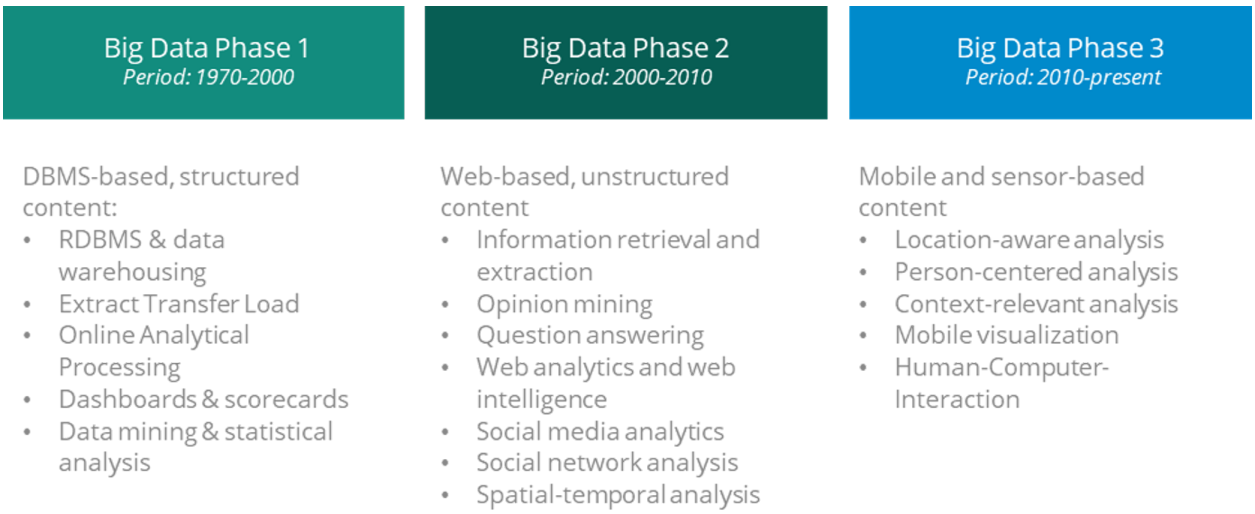

Figure 1.2: The three major phases in the evolution of Big Data

1.4 Big Data Characteristics

The term Big Data is generally used to indicate a data set that is 'massive' in size, and therefore difficult to store, process and analyse with traditional computing resources. From the definition of Big Data that was presented in Chapter 1.1, the question remains how the term 'massive' can be defined. In other words, what elements make Big Data truly 'Big'?

Although there is no universally accepted answer to this question, it is common practice to define Big Data through several key characteristics.[9] The most widely accepted characteristics of Big Data are denoted by the 4V model, which is depicted in figure 1.3.[10] The 4V model considers the nature and necessity of Big Data by considering core properties of massive data sets: volume, velocity, variety, and veracity.

1. **Volume** – The volume of data refers to the inherent size of the data sets that need to be analysed and processed, which are now frequently larger than terabytes or petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities. Typically, the volume characteristic indicates that the data set of interest is too large to process with a regular laptop or desktop processor, and that specific Big Data technology is required to run the analysis. An example of a high-volume data set would be all credit card transactions on a day for a particular credit card company.
2. **Velocity** – Velocity refers to the speed with which data is generated as well as the speed with which the data can be analysed or processed. The fact that data is generated at high speeds requires no further explanation, when you consider the amount of content that is posted on social media platforms in a single minute. In many cases, the value of Big Data is not only the

ability of companies to analyze large data sets, but also to execute this within an acceptable time frame. Particularly with time-sensitive information, such as stock price information, the ability to process and analyse data quickly can provide a competitive advantage.

3. **Variety** – Variety refers to the different types of data collected through sensors, smart phones or social media. Most of these data generating devices will capture data in different formats. A smartphone, for example, will capture data into a variety of different formats. Messages could be stored as text files, photos in jpeg format, and videos will be converted to an mp4 format. The ability to analyse different data formats is inherent to obtain value from data sets. The variety of different data types frequently requires distinct processing capabilities and specialist algorithms. A classification of different data structure will be further discussed in Chapter 1.6.
4. **Veracity** – Veracity refers to the quality of the data that is being analyzed. High veracity data has many records that are valuable to analyze and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data. The non-valuable in these data sets is referred to as noise. An example of a high veracity data set would be data from a medical experiment or trial.

Figure 1.3: The Four Big Data Characteristics - 4V Model

Data that is characterized by high volume, high velocity, high variety, and high veracity must be processed with advanced technological solutions to reveal meaningful information. Because of these characteristics of the data, the knowledge domain that deals with the storage, processing, and analysis of these data sets has been labeled Big Data.

1.5 Data Analysis, Analytics, Data Science, BI and Big Data

In corporate environments, the use of data to support decision-making has been well established. Most corporations utilize data analysis technologies daily and have structured reporting and dashboards in place. For more advanced problems, organization have created analytics or data science teams to help find the best predictions.

Because many 'data terms' are uses interchangeably, we will start this section with providing an overview of the most common terms that are used in enterprise environments. We will discuss the following five concepts:

1. Data Analysis
2. Data Analytics
3. Data Science
4. Business Intelligence (BI)
5. Big Data

Although all five definitions are closely related, there are some subtle differences between the terms that have an impact for the design of Big Data solutions.

Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains. Data analysis – in the literal sense – has been around for centuries (as discussed in Chapter 1.2).

The primary purpose of data analysis is to review existing data to describe patterns that have happened in the past. It is therefore also frequently referred to as descriptive data analysis. An example of data analysis would be to review the sales patterns of different stores over the past years.

With data analysis, there is a strong focus on the process. What are the best steps in which data can be analysed, and what constitutes a sound reproducible process? As a result, data analysis typically a very process-oriented exercise or, as the famous statistician John Tukey defined it:

> **Data analysis included all procedures for analysing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise, or more accurate, and all the machinery and results of (mathematical) statistics which apply to analysing data.**[11]

Figure 1.4: Data analysis - sales patterns in stores

Data Analytics

Data Analytics is the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.

Data Analytics encompasses a growing field of data science capabilities including statistics, mathematics, machine learning, predictive modelling, data mining, cognitive computing, and artificial intelligence. There are four categories of data analytics that organizations need to consider:

1. **Descriptive analytics**: Descriptive analytics or data mining are at the bottom of the big data value chain, but they can be valuable for uncovering patterns that offer insight. A simple example of descriptive analytics would be reviewing the number of people that visited the company's website over the past few months. Descriptive analytics can be useful in the sales cycle, for example, to spot seasonal trends and to adjust purchasing decisions accordingly.
2. **Diagnostic analytics**: Diagnostic analytics are used for discovery or to determine why something happened. In a social media marketing campaign for example, diagnostic analytics can be used to determine why certain advertisements resulted in increased conversion rates. Diagnostic analytics provide valuable insights for organizations because it helps them understand which decisions impact the company's performance.
3. **Predictive analytics**: Predictive analytics use Big Data to identify past patterns to predict the future. From trends or patterns in existing data sets, predictive algorithms calculate the probability that a certain event will occur. For example, some companies are using predictive analytics for sales lead scoring, indicating which incoming sales leads will have the highest chance of converting into an actual customer. Properly tuned predictive analytics can be used to support sales, marketing, or for other types of complex forecasts.

4. **Prescriptive analytics**: Prescriptive analytics is the last and most valuable level of analytics. While Big Data analytics in general sheds light on a subject, prescriptive analytics gives you a laser-like focus to answer specific questions. For example, in the health care industry, you can better manage the patient population by using prescriptive analytics to measure the number of patients who are clinically obese, then add filters for factors like diabetes and LDL cholesterol levels to determine where to focus treatment. The same prescriptive model can be applied to almost any industry target group or problem.

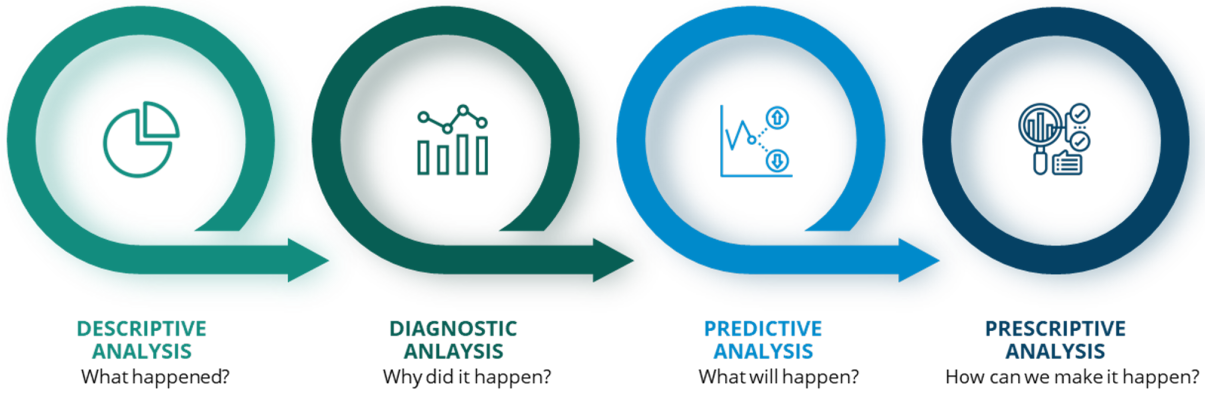

Figure 1.5: Four different types of analytics and their increased value

Whereas data analysis aims to support decision-making by reviewing past data (i.e., descriptive, or diagnostic analytics), analytics in the context of Big Data is primarily concerned with optimizing the future (i.e., predictive, or prescriptive analytics). For this purpose, analytics makes use of (complex) algorithms to find patterns in data to provide advice on the best possible course of action for an organization (i.e., recommendations). An example of a popular and widely used analytics tool is Google Analytics that organizations can use to predict website traffic and optimize online advertisements.

Business Intelligence

Business Intelligence (BI) comprises of the strategies and technologies used by enterprises for the data analysis of business information. Business Intelligence uses both data analysis and analytics techniques to consolidate and summarize information that is specifically useful in an enterprise context.

The key challenge with Business Intelligence is to consolidate the different enterprise information systems and data sources into a single integrated data warehouse on which analysis or analytics operations can be performed. A data warehouse is a (large) centralized database in an organization that combines a variety of different databases from different sources. An example of Business Intelligence would be to build a management dashboard that visualizes key enterprise KPIs across different division in the world.

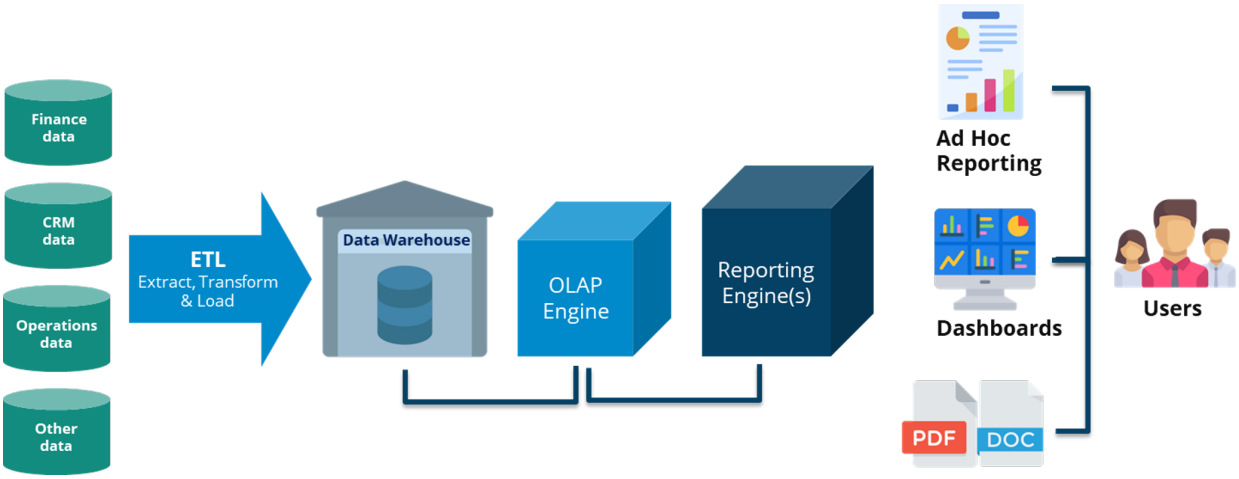

Figure 1.6: Structure of a Business Intelligence environment

Business Intelligence (BI) operations are executed by a wide variety of enterprise applications and technologies for collecting, storing, analysing, and providing access to data to help the enterprise users make better decision making. Most BI applications support querying and reporting, online analytical processing (OLAP), statistical data analysis, forecasting and data mining, as depicted in figure 1.6.

Data Science

Out of all definitions that are discussed in this section, Data Science has grown to great popularity in the last decade. 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, and apply knowledge from data across a broad range of application domains.[12]

The National Institute of Standards and Technology (NIST) defnes Data Science as follows:

> **Data science is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process.**

From the definition of data science, it is important to note that there is no reference to the size, speed or structure of the underlying data sets. Rather, data science focuses on the processs to obtain knowledge from data. For that reason, data science is more concerned with the data lifecycle process, and requires a fundemental understanding of the way in which actionable knowledge is obtained from data. In Chapter 5, we will cover the fundamental aspects of data science in greater detail.

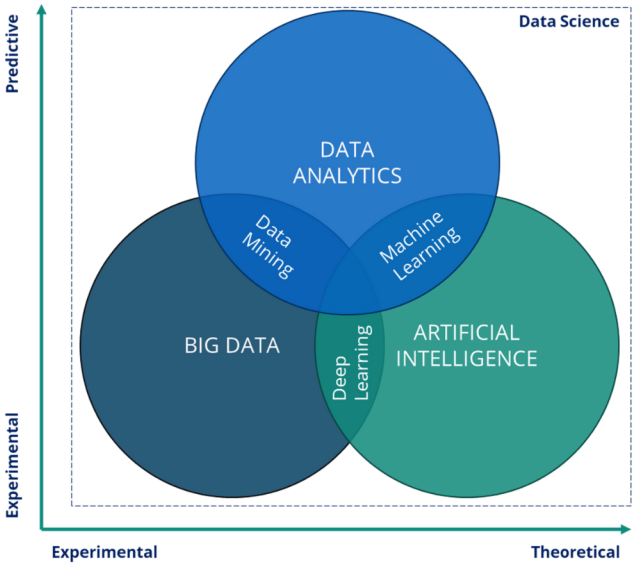

Figure 1.7: The Data Science knowledge domain

Data Science has become a 'container term' for more advanced data processing techniques, than encompass Data Analytics, Big Data and Artificial Intelligence (depicted in figure 1.7). More importantly, data science has grown to become the job title of choice for people involved in the data domains. Practitioners that are involved with data analysis, data analytics, business intelligence or Big Data, prefer to call themselves data scientists. As a result, the demand and interest in Data Science has been skyrocketing.

Big Data

As discussed in Chapter 1.4, Big Data is characterized by four key characteristics — the four V's. Big Data makes use of both data analysis and analytics techniques and frequently builds upon the data in enterprise data warehouses (as used in BI). As such, it can be considered the 'next step' in the evolution of Business Intelligence. Big Data, however, requires a different approach than Business Intelligence for a number of key reasons.

- The data that is analyzed in Big Data environments is larger than what most traditional BI solutions can cope with, and therefore requires distinct and *distributed storage and processing* solutions.
- Big Data is characterized by the variety of its data sources and *includes unstructured or semi-structured data.* Big Data solutions need, for example, to be able to process images of audio files.

The difference between Big Data and Business Intelligence is depicted in figure 1.8:

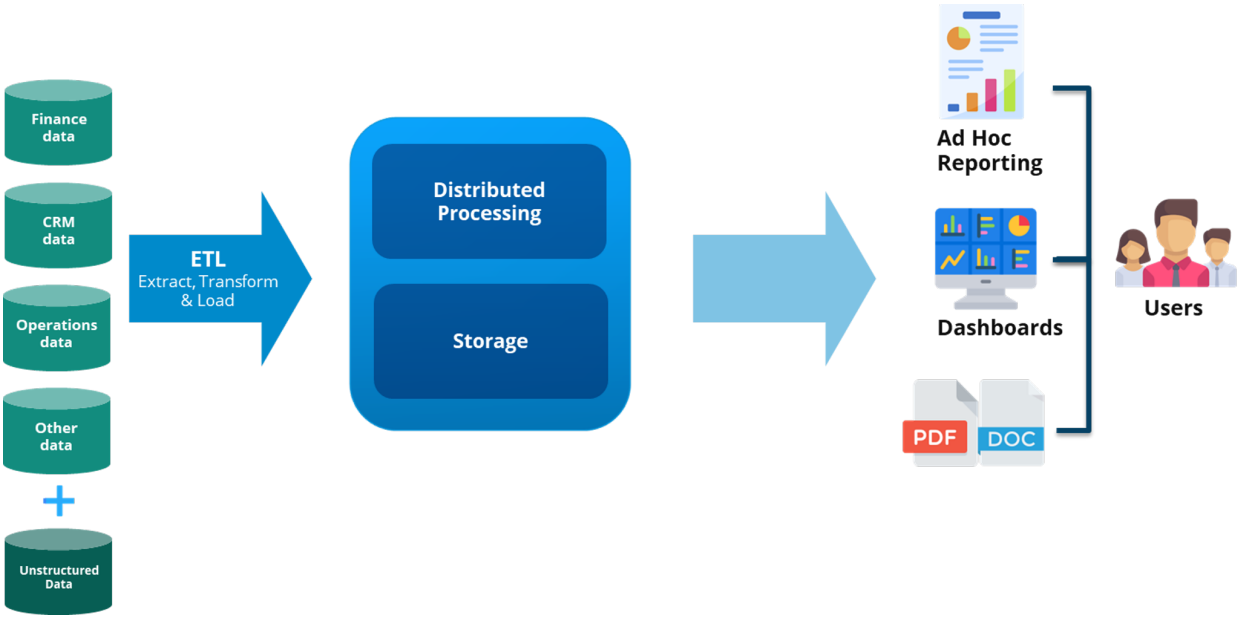

Figure 1.8: Big Data includes unstructured data and requires distributed storage and processing

Figure 1.8 provides a very high-level overview of Big Data solutions. The technical structure and architecture of Big Data environments will be further discussed in detail in Chapter 4.

1.6 Data Structures

In computer science, a data structure is a particular way of organizing and storing data in a computer such that it can be accessed and modified efficiently. More precisely, a data structure is a collection of data values, the relationships among them, and the functions or operations that can be applied to the data.

For the analysis of data, it is important to understand that there are three common types of data structures:

1. **Structured data**: Structured data is data that adheres to a pre-defined data model and is therefore straightforward to analyze. Structured data conforms to a tabular format with relationship between the different rows and columns. Common examples of structured data are Excel files or SQL databases. Each of these have structured rows and columns that can be sorted.
2. **Unstructured data**: Unstructured data is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in structures databases. Common examples of unstructured data include audio, video files or No-SQL databases.

3. **Semi-structured data**: Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contain tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure.[13] Examples of semi-structured data include JSON and XML are forms of semi-structured data.

Most 'traditional' data analysis and analytics techniques (including most Business Intelligence solutions) have the ability to process structured data. Processing unstructured or semi-structured data is however much more complex and requires distinct solutions for analysis.

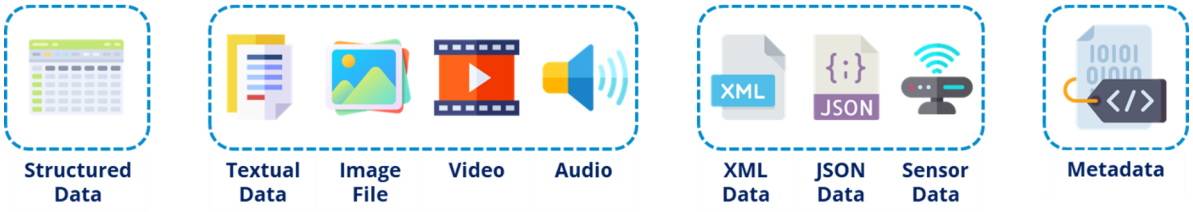

Figure 1.9: Different data structures

A last category of data type is **metadata**. From a technical point of view, this is not a separate data structure, but it is one of the most important elements for Big Data analysis and big data solutions. Metadata is data about data. It provides additional information about a specific set of data. In a set of photographs, for example, metadata could describe when and where the photos were taken. The metadata then provides fields for dates and locations which, by themselves, can be considered structured data. Because of this reason, metadata is frequently used by Big Data solutions for initial analysis.

1.7 Data Products and Big Data Solutions

Working with data and performing data analysis and analytics operations requires specialized knowledge. In most organizations, business users are only interested in finding the outcomes and solutions to certain questions — they require a data product. A data product is an application that runs data analysis or analytics operations upon a certain input, and mostly have an easy-to-understand user interface. The users of data products do not exactly understand all underlying algorithms and are only able to run certain queries to find specific answers. Building data products can therefore be considered one of the key objectives of Big Data.

Because of the growing interest in Big Data and its increased use in enterprise organizations, many data products have been developed. Commercial software companies bundle many data products together, and license these as Big Data solutions to enterprise organizations. Big Data solutions are a

quick way for enterprises to start leveraging the potential of Big Data analysis, because enterprises do not need to develop all required data products in-house. The downside of (commercial) Big Data solutions is that they are often expensive, and it is difficult to alter any of the underlying algorithms of the Big Data solution.

There are many Big Data solutions available on the market and almost every large Enterprise IT provider (Google, Amazon, Microsoft, SAP, etc.) now offers one or more Big Data solutions. Additionally, start-ups play a very important role in the development of Big Data solutions because they come up with new and innovative data products. The Enterprise Big Data Framework has been developed from a vendor-independent perspective and therefore does not recommend any specific Big Data solutions.

Figure 1.10: Examples of popular Big Data vendors

Hadoop

It would not be possible to discuss Big Data solutions without mentioning Hadoop. Hadoop is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model. It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework.[14]

Distributed storage, distributed processing and the Hadoop framework will be explained in more detail in Chapter 4. However, it is important to note that most Big Data solutions make use of the Hadoop framework as their underlying software framework. The term 'Hadoop' has therefore also become known as the ecosystem that connects different Big Data solutions (and commercial vendors) together.

1.8 Artificial Intelligence

Artificial Intelligence (AI) is intelligence displayed by machines, in contrast to the natural intelligence (NI) displayed by humans and other animals. The domain of Artificial Intelligence was first envisioned by a handful of computer scientists at the Dartmouth conferences in 1956, and has seen an explosive growth, especially since 2015.[15]

Whereas Artificial Intelligence can be considered a complete domain of science by itself, it is strongly interwoven with Big Data because the volume and variety of data sources are often massive (in terms of volume) and diverse (in terms of sensors). Additionally, many of the statistical and machine learning algorithms that are used to analyze Big Data sets, are similar to the ones used in Artificial Intelligence.

The knowledge domain of Artificial Intelligence has evolved over the years to include Machine Learning algorithms (discussed in the next section) and finally Deep Learning, which is driving today's AI explosion. The evolution of AI, Machine Learning and Deep Learning is depicted in figure 1.11:

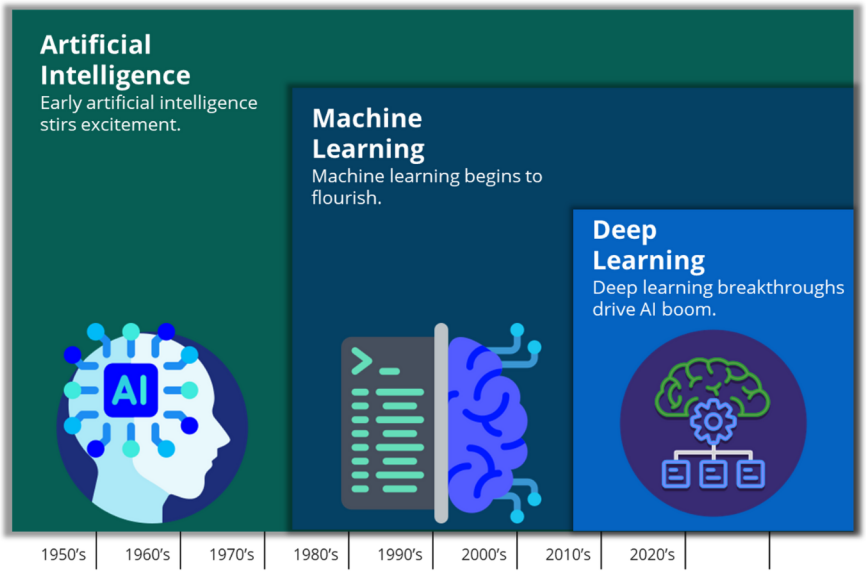

Figure 1.11: The evolution of AI, Machine Learning and Deep Learning

In the course of the evolution of Artificial Intelligence, the underlying algorithms have become more complex and omnipotent. Besides its technical challenges and complexity, Artificial Intelligence also raises many sociological and ethical questions that makes the subject even more complex. The foundational concepts of Artificial Intelligence, Machine Learning and Deep Learning are further explained in Chapter 8 of this guide.

A popular example of the application of AI is self-driving cars. The final objective of self-driving cars is to mimic the exact same behaviors as 'natural' people would make whilst-driving (or preferably

even better behavior without any accidents). The input data that have to be processed need to come from different sensors (high variety) and needs to process thousands of signals every single second (high velocity and high volume) as traffic situations change.

Many large tech organizations envision Artificial Intelligence as a new wave of economic potential that provides growth potential for large enterprises.[16] AI builds on the capabilities and knowledge of Big Data and it is therefore important to have a solid foundation of this topic.

1.9 Machine Learning

The domains of Big Data and Machine Learning are very closely related and have become more interwoven in recent years. Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data.[17]

Machine Learning aims to 'teach' computers to perform certain operations (by running machine learning algorithms), so that the computer is able to make improved decisions in the future and can 'learn' from previous situations. Machine Learning is widely used for the purpose of data mining, which is the subject of shifting through large amounts of data to find unknown or hidden patterns.

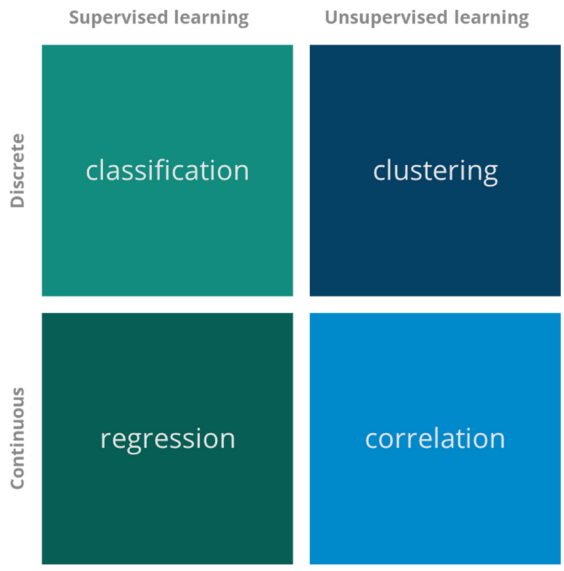

Figure 1.12: The four main Machine Learning domains

Supervised and Unsupervised Machine Learning

Machine Learning algorithms can be subdivided into two main classes (supervised and unsupervised) based on whether or not the input data is labelled, or completely unknown:

1. **Supervised Machine Learning**: In supervised machine learning, a computer learns a certain task because it is fed labeled training data.[18] In other words, the computer is first confronted with a number of 'sample cases', from which it learns what decision to make. When new data than enters the system, the system subsequently 'knows' what decision to make. For this reason, supervised machine learning is mostly associated with classification and regression techniques. An example of supervised machine learning is the sorting of email messages into a spam folder or as regular mail. The computer first needs to 'learn' which type of messages should be considered spam, by feeding a set of training data. After the computer 'understands' this training set and derived certain rules from it, it can classify future emails by itself.
2. **Unsupervised Machine Learning**: In unsupervised machine learning, a computer is fed data and needs to infer relationships in the data, without any prior knowledge about the data set. Any set of data can be fed into the computer, after which the machine will try to find certain patterns and relationships within the data. Unsupervised machine learning is therefore ideal for the purpose of data mining. The techniques associated with machine learning are clustering and correlation. An example of unsupervised machine learning would be to feed large amount of insurance claims into a computer. Based on unsupervised learning algorithms, the computer might find that certain claims do not fit within a regular pattern and therefore might be fraudulent. These outliers would then need to be evaluated and validated by insurance agents.

Classification, regression, clustering and correlation are further explained in Chapter 5. Although it is technically not necessary to have 'big' data sets in order to perform machine learning operations, much value can be generated when the two are paired. In the rest of this guide, we will therefore consider machine learning in the context of Big Data.

Case Study: Big Data at UPS

UPS is no stranger to big data, having begun to capture and track a variety of package movements and transactions since as early as the 1980s. The company now tracks data on 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking requests from customers per day. The company stores over 16 petabytes of data.

Much of its recently acquired big data, however, comes from telematics sensors in over 46,000 vehicles. The data on UPS package cars (trucks), for example, include their speed, direction, braking, and drive train performance. The data is not only used to monitor daily performance, but also to drive a major redesign of UPS drivers' route structures. This initiative, called ORION (OnRoad Integrated Optimization and Navigation), is arguably the world's largest operations research project.

It also relies heavily on online map data, and will eventually reconfigure a driver's pickups and drop-offs in real time of more than 8.4 million gallons of fuel by cutting 85 million miles off of daily routes. UPS estimates that saving only one daily mile driven per driver saves the company $30 million, so the overall dollar savings are substantial.

Notes

1 Kurzweil, R., Richter, R., Kurzweil, R. and Schneider, M.L., 1990. The age of intelligent machines (Vol. 579). Cambridge: MIT press.

2 Turing, A.M., 2009. Computing machinery and intelligence. In Parsing the Turing Test (pp. 23-65). Springer, Dordrecht.

3 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice Hall.

4 Chui, M., 2017. Artificial intelligence the next digital frontier?. McKinsey and Company Global Institute, p.47.

5 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice hall.

6 Baral, C. and De Giacomo, G., 2015, January. Knowledge Representation and Reasoning: What's Hot. In AAAI (pp. 4316-4317).

7 Aron, A. and Aron, E.N., 1994. Statistics for psychology. Prentice-Hall, Inc.

8 Chambers, J.M., 2018. Graphical methods for data analysis. CRC Press.

9 White, T., 2012. Hadoop: The definitive guide. O'Reilly Media, Inc.

10 Borthakur, D., 2008. HDFS architecture guide. Hadoop Apache Project, 53.

11 Brown, B., Chui, M. and Manyika, J., 2011. Are you ready for the era of 'big data'. McKinsey Quarterly, 4(1), pp.24-35.

12 Dhar, V., 2013. Data science and prediction. Communications of the ACM, 56(12), pp.64-73.

13 Buneman, P., 1997, May. Semistructured data. In Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (pp. 117-121). ACM.

14 Zikopoulos, P. and Eaton, C., 2011. Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.

15 Michael Copeland. 2016. What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?. Online available [Accessed 4 April 2018].

16 Clark, J., 2015. Why 2015 was a breakthrough year in artificial intelligence. Bloomburg Business, 8.

17 IDG Enterprise. 2016. 2016 Data & Analytics Research. Online available

18 Mohri, M., Rostamizadeh, A. and Talwalkar, A., 2012. Foundations of machine learning. MIT press.

2. The Enterprise Big Data Framework

2.1 Why an Enterprise Big Data Framework?

Frameworks provide structure. The core objective of the Big Data Framework is to provide a structure for enterprise organizations that aim to benefit from the potential of Big Data. In order to achieve long-term success, Big Data is more than just the combination of skilled people and technology — it requires structure and capabilities.

The Enterprise Big Data Framework was developed because — although the benefits and business cases of Big Data are apparent — many organizations struggle to embed a successful Big Data practice in their organization. The structure provided by the Big Data Framework provides an approach for organizations that takes into account all organizational capabilities of a successful Big Data practice. All the way from the definition of a Big Data strategy, to the technical tools and capabilities an organization should have.

The main benefits of applying the Enterprise Big Data framework include:

- The Enterprise Big Data Framework provides a structure for organizations that want to start with Big Data or aim to develop their Big Data capabilities further.
- The Enterprise Big Data Framework includes all organizational aspects that should be taken into account in a Big Data organization.
- The Enterprise Big Data Framework is vendor independent. It can be applied to any organization regardless of choice of technology, specialization or tools.
- The Enterprise Big Data Framework provides a common reference model that can be used across departmental functions or country boundaries.
- The Enterprise Big Data Framework identifies core and measurable capabilities in each of its six domains so that the organization can develop over time.

Big Data is a people business. Even with the most advanced computers and processors in the world, organizations will not be successful without the appropriate knowledge and skills. The Enterprise Big Data Framework therefore aims to increase the knowledge of everyone who is interested in Big Data. The modular approach and accompanying certification scheme aims to develop knowledge about Big Data in a similar structured fashion.

The Enterprise Big Data framework provides a holistic structure toward Big Data. It looks at the various components that enterprises should consider while setting up their Big Data organization. Every element of the framework is of equal importance and organizations can only develop further if they provide equal attention and effort to all elements of the Enterprise Big Data Framework.

2.2 The Structure of the Enterprise Big Data Framework

The Enterprise Big Data Framework is a structured approach that consists of six core capabilities that organizations need to take into consideration when setting up their Big Data organization. The Enterprise Big Data Framework is depicted in figure 2.1:

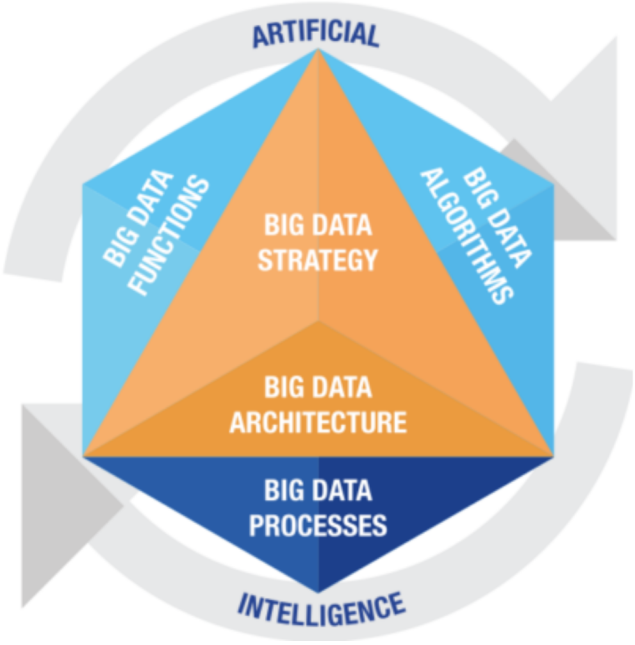

Figure 2.1: The Enterprise Big Data Framework

The Enterprise Big Data Framework consists of the following six main capabilities:

Big Data Strategy

Data has become a strategic asset for most organizations. The capability to analyze large data sets and discern pattern in the data can provide organizations with a competitive advantage. Netflix, for example, looks at user behavior in deciding what movies or series to produce. Alibaba, the Chinese sourcing platform, became one of the global giants by identifying which suppliers to loan money and recommend on their platform. Big Data has become Big Business.

In order to achieve tangible results from investments in Big Data, enterprise organizations need a sound Big Data strategy. How can return on investments be realized, and where to focus effort in Big Data analysis and analytics? The possibilities to analyze are literally endless and organizations can easily get lost in the zettabytes of data. A sound and structured Big Data strategy is the first step to Big Data success. In Chapter 3, we explore the business drivers of Big Data and discuss how to formulate a Big Data strategy.

Big Data Architecture

In order to work with massive data sets, organizations should have the capabilities to store and process large quantities of data. In order to achieve this, the enterprise should have the underlying IT infrastructure to facilitate Big Data. Enterprises should therefore have a comprehensive Big Data architecture to facilitate Big Data analysis. How should enterprises design and set up their architecture to facilitate Big Data? And what are the requirements from a storage and processing perspective?

The Big Data Architecture capability of the Enterprsie Big Data Framework considers the technical capabilities of Big Data environments. It discusses the various roles that are present within a Big Data Architecture and looks at the best practices for design. In line with the vendor-independent structure of the Framework, Chapter 4 will consider the Big Data reference architecture of the National Institute of Standards and Technology (NIST).

Big Data Algorithms

A fundamental capability of working with data is to have a thorough understanding of statistics and algorithms. Big Data professionals therefore need to have a solid background in statistics and algorithms to deduce insights from data. Algorithms are unambiguous specifications of how to solve a class of problems. Algorithms can perform calculations, data processing and automated reasoning tasks. By applying algorithms to large volumes of data, valuable knowledge and insights can be obtained.

The Big Data algorithms element of the framework focuses on the (technical) capabilities of everyone who aspires to work with Big Data. It aims to build a solid foundation that includes basic statistical operations and provides an introduction to different classes of algorithms. Chapter 5 provides an elementary introduction to the topic of Big Data algorithms. Advanced statistical and machine learning techniques are discussed in the Enterprise Big Data Analyst* and Enterprise Big Data Scientist† guides.

Big Data Processes

In order to make Big Data successful in any enterprise organization, it is necessary to consider more than just the skills and technology. Processes can help enterprises to focus their direction. Processes bring structure, measurable steps and can be effectively managed on a day-to-day basis. Additionally, processes embed Big Data expertise within the organization by following similar procedures and steps, embedding it as ‘a practice’ of the organization. Analysis becomes less dependent on individuals and thereby, greatly enhancing the chances of capturing value in the long term.

*https://www.bigdataframework.org/big-data-certification/enterprise-big-data-analyst/
†https://www.bigdataframework.org/big-data-certification/enterprise-big-data-scientist/

Chapter 6 provides an overview of three fundamental Big Data processes that are applicable to every organization. It discusses the benefits of every process and provides a step-by-step description of the process activities to embed a Big Data practice in the organization.

Big Data Functions

Big Data functions are concerned with the organizational aspects of managing Big Data in enterprises. This element of the Big Data framework addresses how organizations can structure themselves to set up Big Data roles and discusses roles and responsibilities in Big Data organizations. Organizational culture, organizational structures and job roles have a large impact on the success of Big Data initiatives. We will therefore review some 'best practices' in setting up enterprise Big Data.

In the Big Data Functions chapter, we discuss the non-technical aspects of Big Data. Chapter 7 discusses the practical aspect of setting up a Big Data Center of Excellence (BDCoE) and provides detailed guidance on the elements of the BDCoE. Additionally, it also addresses critical success factors for starting Big Data project in the organization.

Artificial Intelligence

The last element of the Big Data Framework addresses Artificial Intelligence (AI). One of the major areas of interest in the world today, AI provides a whole world of potential. In this part of the framework, we address the relation between Big Data and Artificial Intelligence and outline key characteristics of AI.

Many organizations are keen to start Artificial Intelligence projects, but most are unsure where to start their journey. This guide takes a functional view of AI in the context of bringing business benefits to enterprise organizations. Chapter 8 therefore showcases how AI follows as a logical next step for organizations that have built up the other capabilities of the Big Data Framework. The last element of the Big Data Framework has been depicted as a lifecycle on purpose. Artificial Intelligence can start to continuously learn from the Big Data in the organization in order to provide long lasting value.

2.3 Working with the Enterprise Big Data Framework

The Enterprise Big Data Framework has been set up as an open standard for every person or organization to use it the way that they see fit. The framework builds further on existing theories and summarizes major themes and streams within Big Data. There is, however, a logical sequence to the six elements of the framework for organizations that wish to implement the Big Data Framework. The sequence starts from the middle of the framework, and follows the six elements of the framework clockwise as depicted in figure 2.2:

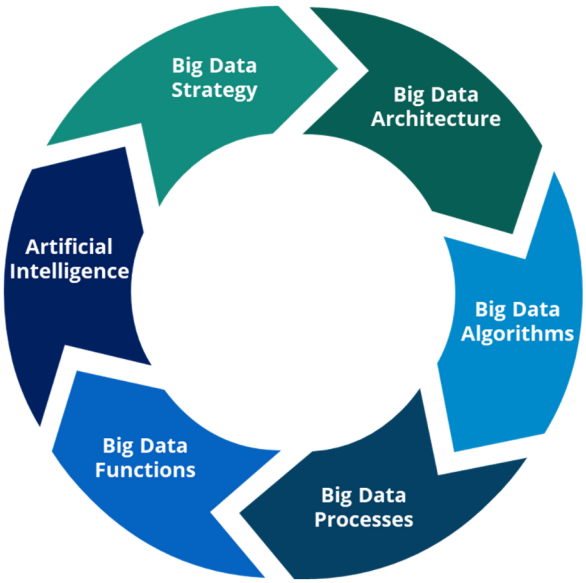

Figure 2.2: Logical sequence to start with the Enterprise Big Data Framework

Although the sequencing is straightforward, it is important to understand that Artificial Intelligence (the last element of the Big Data Framework) can only be achieved if all other elements of the framework are functioning properly. Artificial Intelligence requires a sufficiently developed Big Data organization in order to 'learn' how to make decisions and provide long lasting enterprise value. As an analogy, children also first need to have a basic education (through schools) before they can reason effectively and efficiently.

2.4 Big Data Maturity Assessment

In order to systematically improve the capabilities of their Big Data organization, enterprises can conduct regular (typically annually) Big Data maturity assessments. Big Data maturity models are the artifacts used to measure Big Data maturity. These models help organizations to create structure around their Big Data capabilities and to identify where to start.

A Big Data maturity assessment provides tools that assist organizations to define goals around their Big Data program and to communicate their Big Data vision to the entire organization. The underlying maturity models also provide a methodology to measure and monitor the state of a company's Big Data capability, the effort required to complete their current stage or phase of maturity and to progress to the next stage.[19] Additionally, the Big Data maturity assessment measures and manages the speed of both the progress and adoption of Big Data programs in the organization.

The Enterprise Big Data Framework Maturity Assessment

The Big Data Framework maturity assessment is an example of such a maturity assessment and is based on the six dimensions of the Big Data framework (as discussed in Chapter 2.1). It measures the maturity of Enterprise Big Data over each of these components based on the five point Capability Maturity Model (CMM) scale developed by the Carnegie Mellon Software Engineering Institute. The CMM offers guidelines for organizations to determine their current process maturity and develop a strategy for improving software quality and processes. It consists of the following five stage, as depicted in figure 2.3:

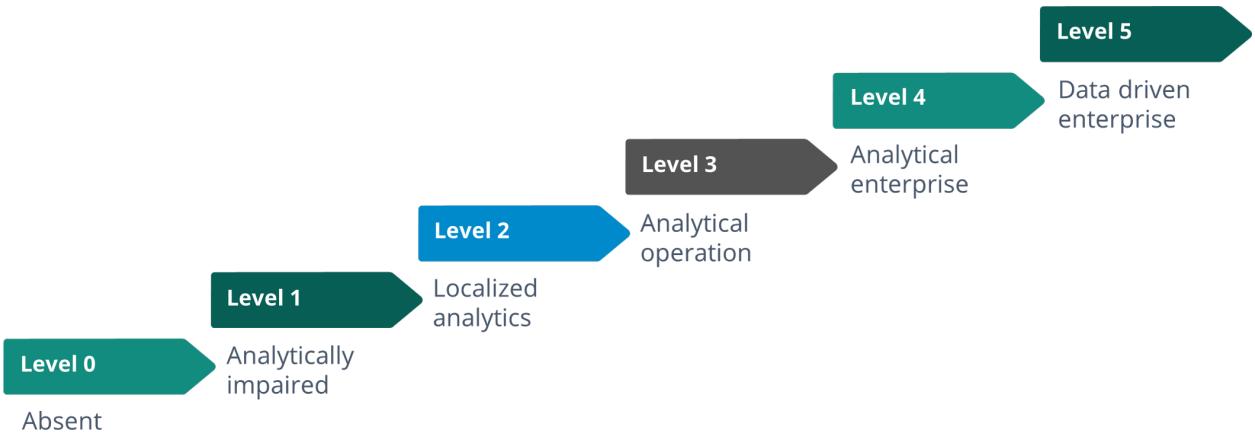

Figure 2.3: Maturity levels of the Enterprise Big Data Framework

The five levels of Big Data maturity are based on the five-scale CMM-levels:

1. **Analytically Impaired (chaotic and ad hoc activity)** - Minimal analytics activities and infrastructure across the enterprise, with ambiguous data and analytics strategy.
2. **Localized Analytics (initial activity)** – Pockets of analytics across the enterprise, however functioning in silos and no overarching data or analytics strategy.
3. **Analytical Operation (repeatable activity)** – Expanding siloed functional analytics to shared operational level analytics with support and commitment from the C-suite.
4. **Analytical Enterprise (managed activity)** – Data and analytics are viewed as an enterprise priority. The organization is developing enterprise wide analytics capabilities across all domains to create meaningful content and ideas.
5. **Data Driven Enterprise (optimized activity)** – Trusted insight created by enterprises with analytics that support strategic decision making. The enterprise is reaping the benefits and is focused on optimization of analytics.

Every area of the Enterprise Big Data Framework is subsequently assessed to determine the level of capability. The outcome of the Enterprise Big Data Framework maturity assessment is depicted in figure 2.4 below and provides valuable information on the potential improvement areas for the organization.

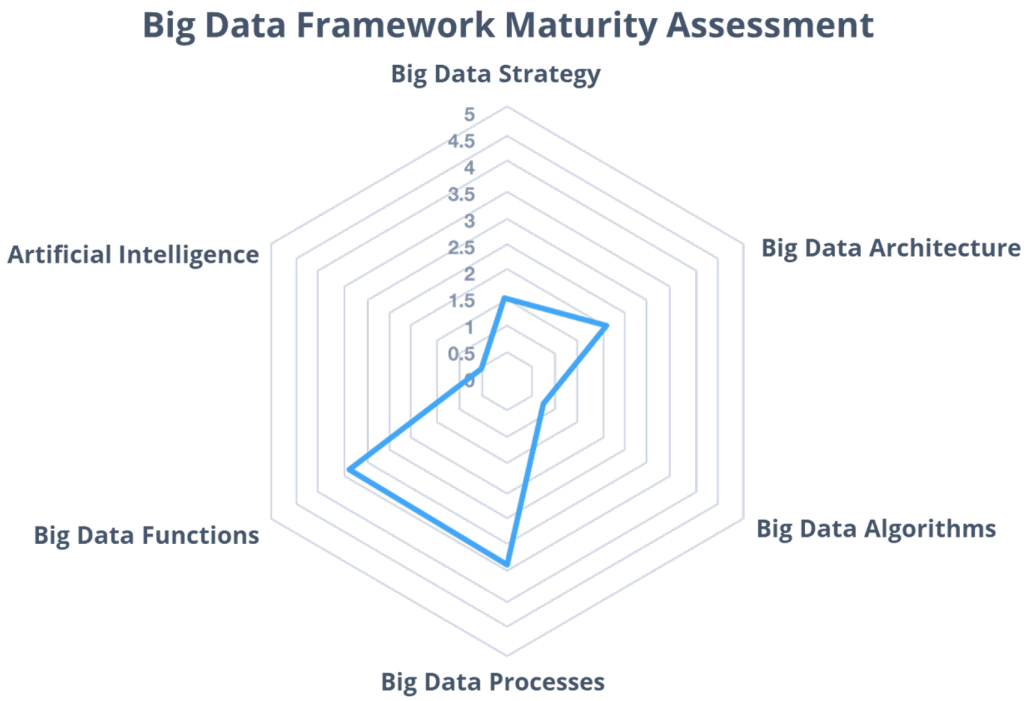

Figure 2.4: The Enterprise Big Data Framework maturity assessment

More information about the criteria for performing a Enterprise Big Data Maturity Assessment, together with detailed guidance is available on the Big Data Framework* website.

*https://www.bigdataframework.org/resources/

Notes

19 Kurzweil, R., Richter, R., Kurzweil, R. and Schneider, M.L., 1990. The age of intelligent machines (Vol. 579). Cambridge: MIT press.

3. Big Data Strategy

3.1 Big Data as Competitive Strategy

Ever since the rise of the modern enterprise, information has been a key strategic asset. Organizations that have more precise customer information, industry insights or marketing information can leverage this information to outperform their competitors. Over the years, groundbreaking systems from companies such as American Airlines (electronic reservations), Otis Elevator (predictive maintenance), and American Hospital Supply (online ordering) have dramatically boosted their creators' revenues and reputations.[20]

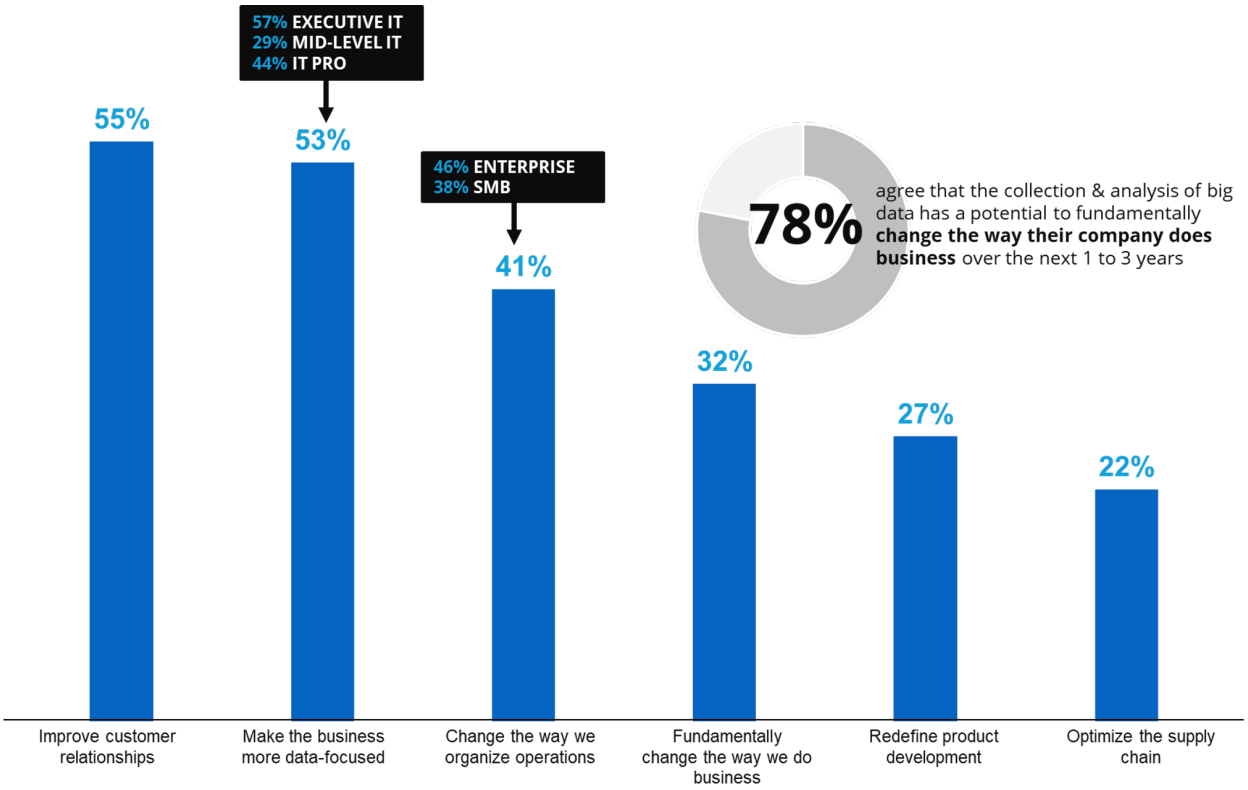

Figure 3.1: Big Data provides organizations with a competitive advantage

The growth of Big Data and Big Data solutions has brought the concept of 'information as a strategic asset' to a whole new level. Organizations now not only compete on the preciseness of their data, but also on their ability to process large quantities (volume), at great speed (velocity), and from disparate data sources. Having new information quicker than competitors provides a small window

of opportunity to act before the competition dives in as well. In a recent survey from IDG, 78% of enterprises agree that data strategy, collection and analysis of Big Data has the potential to fundamentally change the way their company does business over the next one to three years.[21]

Although most enterprises agree that Big Data provides a competitive advantage, many organizations remain poorly behind the curve. Cross-industry studies show that on an average, less than half of an organization's structured data is actively used in making decisions—and less than 1% of its unstructured data is analyzed or used at all. More than 70% of employees have access to data they should not, and 80% of analysts' time is spent simply discovering and preparing data.[22]

The reason why so many companies are struggling to realize their competitive advantage through Big Data is because they have not (adequately) defined a Big Data strategy. In many organizations, Big Data is still project-based, instead of being embedded into the veins of the organization.

In order to avoid these pitfalls and realize a long term competitive advantage, the Enterprise Big Data Framework starts with defining and formulating a Big Data Strategy. Every other activity or process that is further discussed throughout the Enterrprise Big Data Framework should relate back to the Big Data Strategy.

3.2 Business Drivers for Big Data

Big Data emerged in the last decade from a combination of business needs and technology innovations. A number of companies that have Big Data at the core of their strategy have become very successful at the beginning of the 21st century. Famous examples include Apple, Amazon, Facebook and Netflix.

A number of business drivers are at the core of this success and explain why Big Data has quickly risen to become one of the most coveted topics in the industry. Six main business drivers can be identified:

1. The plummeting of technology costs;
2. The digitization of society
3. Connectivity through cloud computing;
4. Increased knowledge about data science;
5. Social media applications;
6. The upcoming Internet-of-Things (IoT).

In this section, we will explore a high-level overview of each of these business drivers. Each of these adds to the competitive advantage of enterprises by creating new revenue streams by reducing the operational costs.

The plummeting of technology costs

Technology related to collecting and processing massive quantities of diverse (high variety) data has become increasingly more affordable. The costs of data storage and processors keep declining, making it possible for small businesses and individuals to become involved with Big Data. For storage capacity, the often-cited Moore's Law still holds that the storage density (and therefore capacity) still doubles every two years. The plummeting of computer memory and storage cost is well been depicted in the figure 3.2 below.

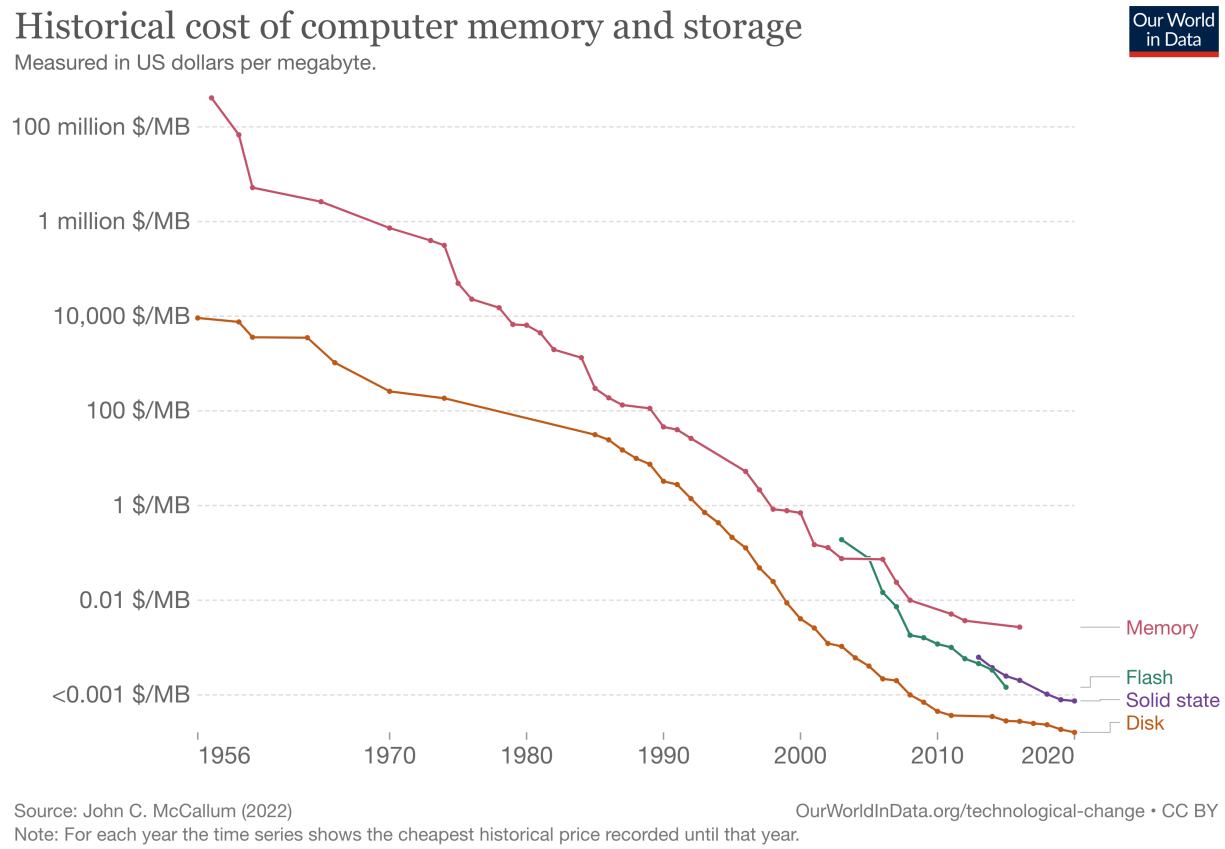

Figure 3.2: Historical cost of computer memory and storage. Source: Our World in Data

Besides the plummeting of the storage costs, a second key contributing factor to the affordability of Big Data has been the development of open source Big Data software frameworks. The most popular software framework (nowadays considered the standard for Big Data) is Apache Hadoop for distributed storage and processing. Hadoop will be further explained in Chapter 4. Due to the high availability of these software frameworks in open sources, it has become increasingly inexpensive to start Big Data projects in organizations.

The digitization of society

Big Data is largely consumer driven and consumer oriented.[23] Most of the data in the world is generated by consumers, who are nowadays 'always-on'. Most people now spend 4-6 hours per day consuming and generating data through a variety of devices and (social) applications. With every click, swipe or message, new data is created in a database somewhere around the world. Because everyone now has a smartphone in their pocket, the data creation sums to incomprehensible amounts. Some studies estimate that 60% of data was generated within the last two years, which is a good indication of the rate with which society has digitized.[24]

Connectivity through cloud computing

Cloud computing environments (where data is remotely stored in distributed storage systems) have made it possible to quickly scale up or scale down IT infrastructure and facilitate a pay-as-you-go model.[25] This means that organizations that want to process massive quantities of data (and thus have large storage and processing requirements) do not have to invest in large quantities of IT infrastructure. Instead, they can license the storage and processing capacity they need and only pay for the amounts they actually used. As a result, most of Big Data solutions leverage the possibilities of cloud computing to deliver their solutions to enterprises.

Increased knowledge about data science

In the last decade, the term data science and data scientist have become tremendously popular. In October 2012, Harvard Business Review called the data scientist "sexiest job of the 21st century" and many other publications have featured this new job role in recent years.[26] The demand for data scientist (and similar job titles) has increased tremendously and many people have actively become engaged in the domain of data science.

As a result, the knowledge and education about data science has greatly professionalized and more information becomes available every day. While statistics and data analysis mostly remained an academic field previously, it is quickly becoming a popular subject among students and the working population.

Social media applications

Everyone understands the impact that social media has on daily life. However, in the study of Big Data, social media plays a role of paramount importance. Not only because of the sheer volume of data that is produced everyday through platforms such as Twitter, Facebook, LinkedIn and Instagram, but also because social media provides nearly real-time data about human behavior.[27]

Social media data provides insights into the behaviors, preferences and opinions of 'the public' on a scale that has never been known before. Due to this, it is immensely valuable to anyone who is able to derive meaning from these large quantities of data. Social media data can be used to identify

customer preferences for product development, target new customers for future purchases, or even target potential voters in elections.[28] Social media data might even be considered one of the most important business drivers of Big Data.

The upcoming Internet-of-Things (IoT)

The Internet of things (IoT) is the network of physical devices, vehicles, home appliances and other items embedded with electronics, software, sensors, actuators, and network connectivity which enables these objects to connect and exchange data. It is increasingly gaining popularity as consumer goods providers start including 'smart' sensors in household appliances. Whereas the average household in 2010 had around 10 devices that connected to the internet, this number is expected to rise to 50 per household by 2020.[29] Examples of these devices include thermostats, smoke detectors, televisions, audio systems and even smart refrigerators.

Figure 3.3: Big Data generated by the Internet-of-Things

Each of these connected devices generates data that is exchanged over the internet and which can be analyzed to retrieve value. Similar to social media, the data that is generated through IoT devices is massive in terms of quantity and can provide insights into the behavior of consumers. As such, it is extremely valuable.

3.3 Formulating a Big Data strategy

In this section, we will discuss a practical approach to formulate a Big Data strategy. As discussed in Chapter 3.1, a comprehensive enterprise-wide Big Data strategy can provide enterprises with a significant competitive advantage in the marketplace. A Big Data strategy, however, cannot be seen as something separate from the organizational strategy, and should be firmly embedded. When we are discussing a Big Data strategy, this effectively means a business strategy that includes Big Data.

A Big Data strategy defines and lays out a comprehensive vision across the enterprise and sets a foundation for the organization to employ data-related or data-dependent capabilities. A well-defined and comprehensive Big Data strategy makes the benefits or Big Data actionable for the organization. It sets out the steps that an organization should execute in order to become a "Data-Driven Enterprise" (as discussed in Chapter 2.4). The Big Data strategy incorporates some guiding principles to accomplish the data-driven vision, directs the organization to select specific business goals and is the starting point for data driven planning across the enterprise.[30]

Besides the gains of realizing a competitive advantage, enterprises require a Big Data strategy because it transcends organizational boundaries. Without a Big Data strategy, enterprises will be forced to deal with a variety of data related activities that will most likely be initiated by different business units. Various departments are likely to start up their own analytics, Business Intelligence or data management programs, without taking into account the overall long-term strategic objectives.

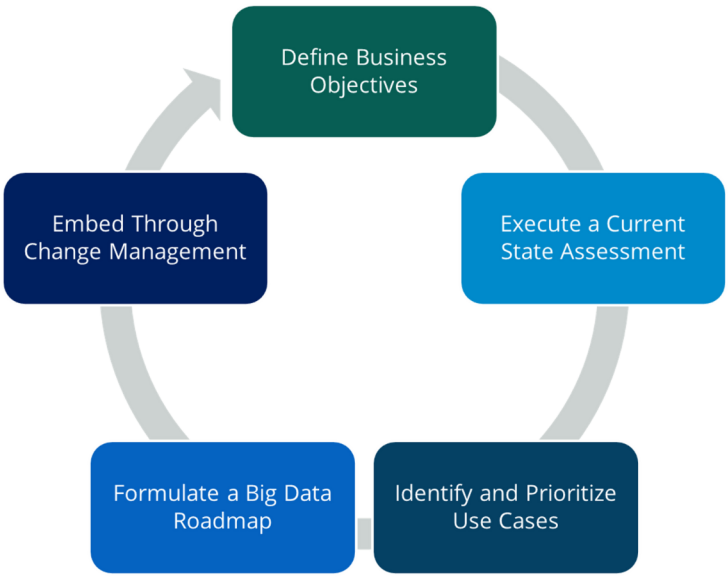

Figure 3.4: The five steps to formulating a Big Data strategy

The driving force behind the formulation of an enterprise Big Data strategy should be the combination of either the CEO/CIO (when Big Data defines the enterprise) or the COO/CIO (when Big Data optimizes the enterprise). This recognizes that the data is not only an IT asset, but also an organization wide corporate asset.

A well-defined enterprise Big Data strategy should be actionable for the organizations. In order to achieve this, organizations can follow the following 5-step approach to formulate their Big Data strategy:

1. Define business objectives
2. Execute a current state assessment
3. Identify and prioritize Use Cases

4. Formulate a Big Data Roadmap
5. Embed through Change Management

Each of the steps to formulate a Big Data strategy is explained in further detail in the sections below. The continous and repetitive nature of these five steps is depicted in figure 3.4.

Step 1: Define business objectives

In order to leverage Big Data in any organization, it is first necessary to fully understand the corporate business objectives of the enterprise. What makes an organization successful? Revenues and profits are often the result of meeting or exceeding business Key Performance Indicators (KPIs). Start with understanding how an organization is successful, before exploring how Big Data technologies and solution might enhance the future performance. The Big Data strategy should align to the corporate business objectives and address key business problems, as the primary purpose of Big Data is to capture value by leveraging data. One way to accomplish this is to align with the enterprise strategic planning process, as most organizations already have this process in place.

Examples of frequently occurring business objectives from a recent survey have been listed in figure 3.5.

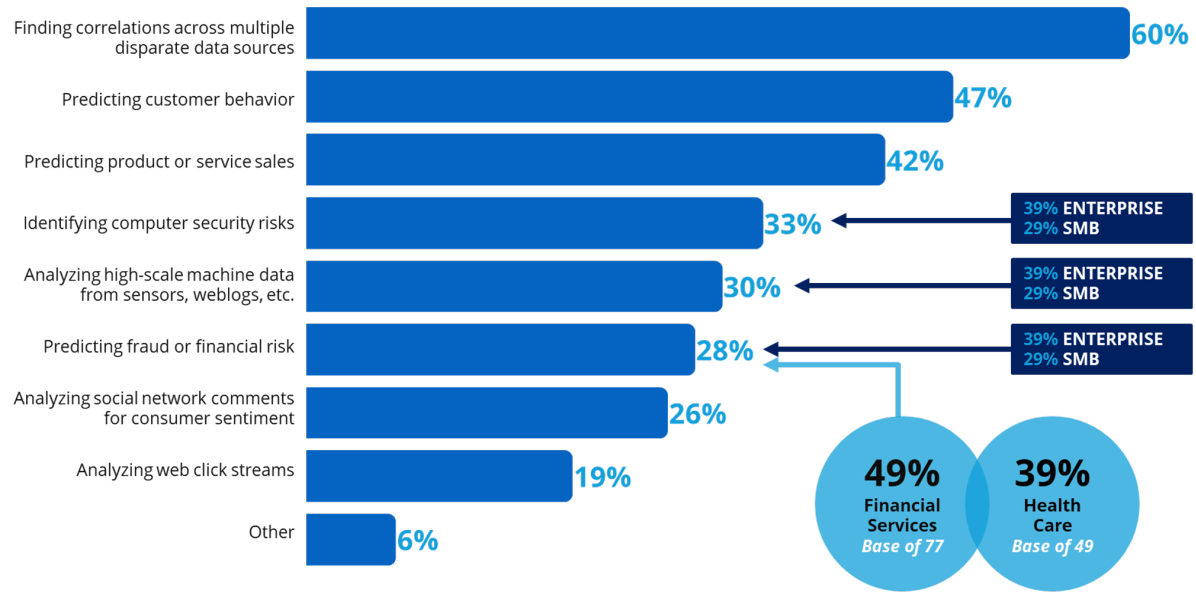

Figure 3.5: Common Big Data business objectives

In order to identify business objectives, involvement of key business stakeholders is of paramount importance. Ensure that these stakeholders are involved right from the start and provide key input on a continuous basis. Key stakeholders to consider in this first step are:

- **Executive sponsors**. The importance of finding and aligning with executive sponsors cannot be underestimated. Their support is essential throughout the ups and downs of formulating the Data Strategy and implementing it.
- **Right talent on the team**. Involving people with the right talent and skill sets is essential in determining the right business objectives. Explore both internal talent as well as external consultants.
- **Potential trouble makers**. Every project or initiative will have some 'stakeholders' who are either deliberately or unintentionally opposed to change. Knowing who they are, and their motivations upfront will help later in the process.

Step 2: Execute a current state assessment

In this step, the primary focus is to assess the current business processes, data sources, data assets, technology assets, capabilities, and policies or the enterprise. The purpose of this exercise is to help with gap analysis of existing state and the desired future state.

As an example, if the scope of the data strategy is to get a 360-degree view of customers and potential customers, the current state assessment would include any business process, data assets including architecture, capabilities (business and IT), and departmental policies that touch customers. Current state assessment is typically conducted with a series of interviews with employees involved in customer acquisition, retention, and processing.

In this stage, it is also important to identify and nurture some data evangelists. These people truly believe in the power of data in making decisions and may already be using the data and analytics in a powerful way. By involving these people, asking for their input, it becomes easier to formulate the roadmap in a later stage.

Step 3: Identify and prioritize Use Cases

In step 3, envision how predictive analytics, prescriptive analytics and ultimately cognitive analytics (further discussed in Chapter 8) can help the organization to accelerate, optimize and continuously learn, by developing Use Cases that align with the business objectives from step 1. Document each of the Use Cases to understand how Big Data can realize the business objective, as per figure 3.6:

BUSINESS OBJECTIVE: PRODUCT LAUNCHES IN EMERGING MARKETS	
Identify which products can be launched most successfully in which market by analyzing buying patterns and correlations between product categories in order to find the optimal product-market combinations.	
BUSINESS POTENTIAL	**IMPLEMENTATIONS RISKS**
• Development of targeted product launches that are aligned with the requirements of each market, so that profit margins are optimized. • Reduction of failed product launches, so that costs are reduced.	• Accuracy and availability of historical purchase data. • Ability to set-up and coordinate localized product launches. • Languages barriers. • Policy and government barriers in emerging markets.
FINANCIAL GOALS IMPACT	**IMPLEMENTATINO CONSIDERATIONS**
• Revenue Growth – 3/5 • Customer Acquisition – 5/5 • Customers Retention – 4/5 • Market Basket Margin – 3/5 • Product Cross Sell – 4/5 • New Product – 4/5 • Financial Goals Score = 3.8/5	• Capturing and governing data models and data sets. • Ability to use predictive models to forecast demand. • Analysis capabilities and knowledge. • Implementation Feasibility = 4/5

Figure 3.6: Defining Big Data use cases

Well-defined Use Cases provide a clear and effective way to define how Big Data technologies and solutions can realize business goals. After the Use Cases have been developed, the next step is to prioritize all of the Use Cases based on their business impact, budget and resource requirements. By conducting this exercise, enterprises can identify which Big Data initiatives provide most business value.

One of the most effective ways to prioritize Use Cases is by using a Prioritization Matrix. Prioritization Matrix facilitates the discussion and debate between the Business and IT stakeholders in identifying the "right" Use Cases to start a Big Data initiative — those Use Cases with both meaningful business value (from the business stakeholders' perspectives) and reasonable feasibility of successful implementation.[31]

The Prioritization Matrix in the figure 3.7 below is an excellent management tool for driving organizational alignment and commitment around the organization's top priority Use Cases.

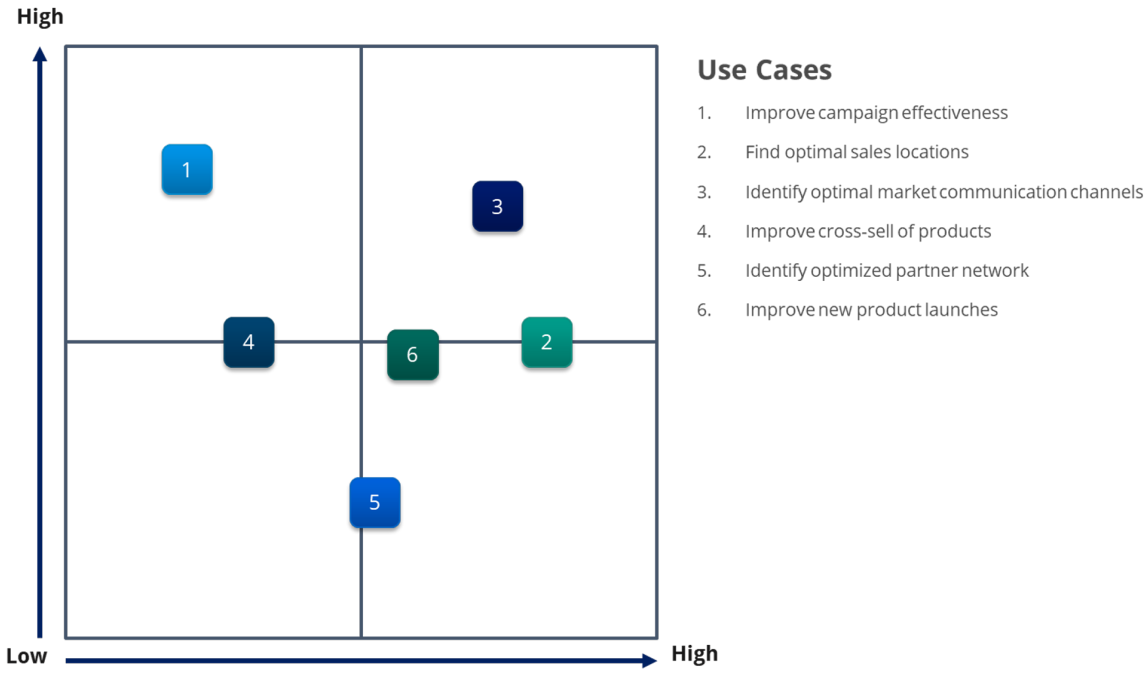

Figure 3.7: Defining a Prioritization Matrix for Big Data Use Cases

Step 4: Formulate a Big Data Roadmap

The next step is probably the most intense and contentious phase and without a doubt will account for majority of the time in formulating data strategy. Based on the current capability state assessment (step 2) and the identified and prioritized Big Data Use Cases (step 3), the Roadmap can be developed.

The Big Data Roadmap outlines which projects (or Use Cases) will be executed first and what capabilities (knowledge, tools and data) will be increased in the next 3-5 years.

With the desired future state in mind, the Roadmap should focus on identifying gaps in data architecture, technology and tools, processes and of course people (skills, training , etc.). The current state assessment and Use Cases will present multiple strategic options for initiatives and the next task is to prioritize these options based on complexity, budget and potential benefits.

The sponsors and stakeholders will have a key role to play in prioritizing these initiatives. The end result of this phase is a roadmap to roll out the prioritized Big Data initiatives.

Step 5: Embed through Change Management

Although technically not a part of the Big Data Strategy formulation, Change Management (involving the hearts and minds of people) will have a profound impact on the success or failure of a Big Data strategy.

Change management should encompass organizational change, cultural change, technological change, and changes in business processes. Data Governance, which deals with the overall management of availability, usability, integrity, and security of data, becomes a crucial component of change management.

Appropriate incentives and ongoing metrics should be key part of any change management program. Further guidance on the Change Management aspect of Big Data is further discussed in Chapter 7.

8.4 The Big Data Strategy Document Checklist

In the previous section, we discussed how organizations can formulate a Big Data strategy. In the end, the Big Data strategy will be written down in a document so that it can be approved and shared with the rest of the organization.

The Big Data Strategy documents should include the following sections, as outlined in the table below:

Topic	Content
Background / Context	This section should articulate background that necessitated the Data Strategy in the first place. Examples could be: Corporate strategic direction, Digital Transformation initiative, or mergers and acquisition related context, etc.
Business Case	The sole purpose of Data Strategy is to unlock business value and this section should articulate the value being unlocked both quantitatively and qualitatively. The business case is probably the toughest one but a necessary one.
Goals	This section identifies specific Data Strategy related goals and ideally in a SMART fashion (Specific, Measurable, Agreed upon, Realistic, Time-based).
Implementation Roadmap	This section connects the strategy with tactics with a roadmap on how the strategy will be implemented over a period of time.
Risk and Succes Factors	Strategy should directly address various risk factors and success enablers (or accelerators). Time and time again, change management is either a major risk or success enabler if not thought through in a detailed fashion so make sure to address it head on in this section.
Budget Estimates	What good is a strategy if it doesn't have budget estimates. My advice is to be realistic and as comprehensive as possible. If you take short cuts to get a strategy approved, it's just a matter of time before it comes back to bite you.
KPIs and Metrocs	To ensure that the strategy is either on track or needs to be adjusted, identify KPIs that need to be tracked on a short term and long term basis.

An example of a Big Data strategy document is available on the Big Data Framework website.

Case Study - Big Data at United Healthcare

United Healthcare, like many large organizations pursuing Big Data, has been focused on structured data analysis for many years, and even advertises its analytical capabilities to consumers ("Health in Numbers"). Now, however, it is focusing its analytical attention on unstructured data—in particular, the data on customer attitudes that is sitting in recorded voice files from customer calls to call centers. The level of customer satisfaction is increasingly important to health insurers, because consumers have more choice about what health plans they belong to. Service levels are also being monitored by state and federal government groups, and published by organizations such as Consumer Reports.

In the past, that valuable data from calls couldn't be analyzed. Now, however, United is turning the voice data into text, and then analyzing it with "natural language processing" software. The analysis process can identify — though it's not easy, given the vagaries of the English language — customers who use terms suggesting strong dissatisfaction. A United representative can then make some sort of intervention—perhaps a call exploring the nature of the problem. The decision being made is the same as in the past — how to identify a dissatisfied customer — but the tools are different.

To analyze the text data, United Healthcare uses a variety of tools. The data initially goes into a "data lake" using Hadoop and NoSQL storage, so the data doesn't have to be normalized. The natural language processing—primarily a "singular value decomposition," or modified word count—takes place on a database appliance. A variety of other technologies are being surveyed and tested to assess their fit within the future state architecture. United also makes use of interfaces between its statistical analysis tools and Hadoop.

Notes

20 Kurzweil, R., Richter, R., Kurzweil, R. and Schneider, M.L., 1990. The age of intelligent machines (Vol. 579). Cambridge: MIT press.

21 Turing, A.M., 2009. Computing machinery and intelligence. In Parsing the Turing Test (pp. 23-65). Springer, Dordrecht.

22 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice Hall.

23 Chui, M., 2017. Artificial intelligence the next digital frontier?. McKinsey and Company Global Institute, p.47.

24 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice hall.

25 Baral, C. and De Giacomo, G., 2015, January. Knowledge Representation and Reasoning: What's Hot. In AAAI (pp. 4316-4317).

26 Aron, A. and Aron, E.N., 1994. Statistics for psychology. Prentice-Hall, Inc.

27 Chambers, J.M., 2018. Graphical methods for data analysis. CRC Press.

28 White, T., 2012. Hadoop: The definitive guide. O'Reilly Media, Inc.

29 Borthakur, D., 2008. HDFS architecture guide. Hadoop Apache Project, 53.

30 Brown, B., Chui, M. and Manyika, J., 2011. Are you ready for the era of 'big data'. McKinsey Quarterly, 4(1), pp.24-35.

31 Dhar, V., 2013. Data science and prediction. Communications of the ACM, 56(12), pp.64-73.

4. Big Data Architecture

4.1 Introduction to Big Data architecture

This chapter provides an overview of fundamental and essential topic areas pertaining to Big Data architecture. We will start by introducing an overview of the NIST Big Data Reference Architecture (NBDRA), and subsequently cover the basics of distributed storage/processing. The chapter will end with an overview of the Hadoop open source software framework.

Everyone presently studying the domain of Big Data should have a basic understanding of how Big Data environments are designed and operated in enterprise environments, and how data flows through different layers of an organization. Understanding the fundamentals of Big Data architecture will help system engineers, data scientists, software developers, data architects, and senior decision makers to understand how Big Data components fit together, and to develop or source Big Data solutions.

In this chapter, we will only cover the fundamentals of Big Data architecture that apply to every enterprise. A more in-depth overview is provided in the Enterprise Big Data Engineer publication.

4.2 The NIST Big Data Reference Architecture

In order to benefit from the potential of Big Data, it is necessary to have the technology in place to analyze huge quantities of data. Since Big Data is an evolution from 'traditional' data analysis (as discussed in section 1.4), Big Data technologies should fit within the existing enterprise IT environment. For this reason, it is useful to have common structure that explains how Big Data complements and differs from existing analytics, Business Intelligence, databases and systems. This common structure is called a reference architecture.

A reference architecture is a document or set of documents to which a project manager or other interested party can refer to for best practices.[32] Within the context of IT, a reference architecture can be used to select the best delivery method for particular technologies and documents such things as hardware, software, processes, specifications and configurations, as well as logical components and interrelationships. In summary, a reference architecture can be thought of as a resource that documents the learning experiences gained through past projects.

The objective of a reference architecture is to create an open standard, one that every organization can use for their benefit. The National Institute of Standards and Technology (NIST) — one of the leading organizations in the development of standards — has developed such a reference architecture: the NIST Big Data Reference Architecture.[33]

The benefits of using an 'open' Big Data reference architecture include:

1. It provides a common language for the various stakeholders;
2. It encourages adherence to common standards, specifications, and patterns;
3. It provides consistent methods for implementation of technology to solve similar problem sets;
4. It illustrates and improves understanding of the various Big Data components, processes, and systems, in the context of a vendor- and technology-agnostic Big Data conceptual model;
5. It facilitates analysis of candidate standards for interoperability, portability, reusability, and extendibility.

The NIST Big Data Reference Architecture is a vendor-neutral approach and can be used by any organization that aims to develop a Big Data architecture. The Big Data Reference Architecture, is shown in Figure 4.1 and represents a Big Data system composed of five logical functional components or roles connected by interoperability interfaces (i.e., services). Two fabrics envelop the components, representing the interwoven nature of management and security and privacy with all five of the components. In the next few paragraphs, each component will be discussed in further detail, along with some examples.

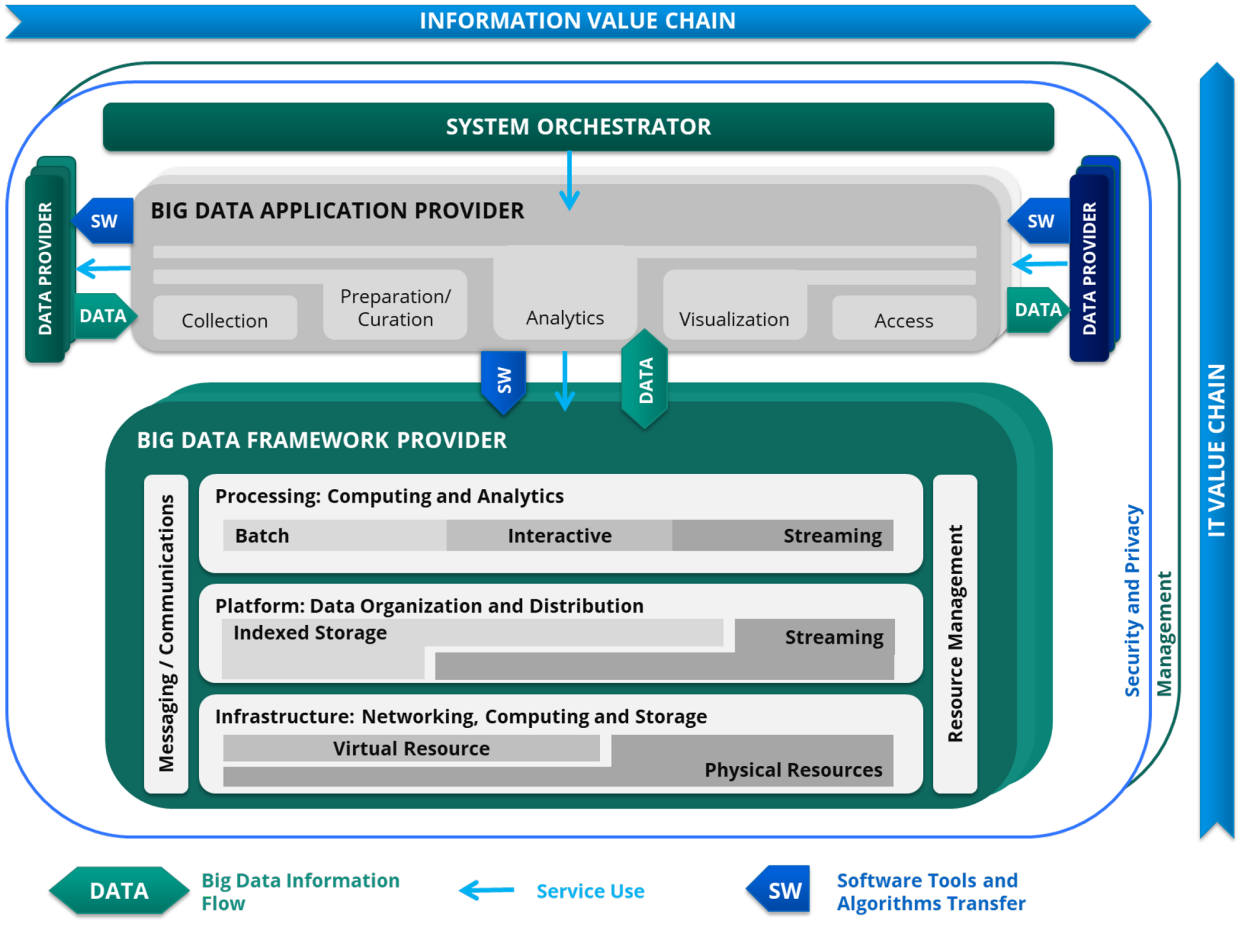

Figure 4.1: The NIST Big Data Reference Architecture (NBDRA)

The NIST Big Data Reference Architecture is organized around five major roles and multiple sub-roles aligned along two axes representing the two Big Data value chains: the Information Value (horizontal axis) and the Information Technology (IT; vertical axis). Along the Information Value axis, the value is created through data collection, integration, analysis, and applying the results following the value chain. Along the IT axis, the value is created through providing networking, infrastructure, platforms, application tools, and other IT services for hosting of and operating the Big Data in support of required data applications. At the intersection of both axes is the Big Data Application Provider role, indicating that data analytics and its implementation provide the value to Big Data stakeholders in both value chains.

The five main **roles** of the NIST Big Data Reference Architecture, shown in figure 4.1 represent the logical components or roles of every Big Data environment, and present in every enterprise:

1. System Orchestrator;
2. Data Provider;
3. Big Data Application Provider;
4. Big Data Framework Provider;
5. Data Consumer.

The two **dimensions** shown in figure 4.2 encompassing the five main roles are:

1. Management;
2. Security & Privacy.

These dimensions provide services and functionality to the five main roles in the areas specific to Big Data and are crucial to any Big Data solution. The Management and Security & Privacy dimension are further discussed in Chapter 6.

System Orchestrator

System Orchestration is the automated arrangement, coordination, and management of computer systems, middleware, and services.[34] Orchestration ensures that the different applications, data and infrastructure components of Big Data environments all work together. In order to accomplish this, the System Orchestrator makes use of workflows, automation and change management processes.

A much cited comparison to explain system orchestration — and the explanation of its name — is the management of a music orchestra. A music orchestra consists of a collection of different musical instruments that can all play at different tones and at different paces. The task of the conductor is to ensure that all elements of the orchestra work and play together in sync. System orchestration is very similar in that regard. A Big Data IT environment consists of a collection of many different applications, data and infrastructure components. The System Orchestrator (like the conductor) ensures that all these components work together in sync.

Data Provider

The Data Provider role introduces new data or information feeds into the Big Data system for discovery, access, and transformation by the Big Data system. The data can originate from different sources, such as human generated data (social media), sensory data (RFID tags) or third-party systems (bank transactions).

One of the key characteristics of Big Data (see Chapter 1.3) is its variety aspect, meaning that data can come in different formats from different sources. Input data can come in the form of text files, images, audio, weblogs, etc. Sources can include internal enterprise systems (ERP, CRM, Finance) or external system (purchased data, social feeds). Consequently, data from different sources may have different security and privacy considerations.

As depicted in figure 4.1, data transfers between the Data Provider and the Big Data Application Provider. This data transfer typically happens in three phases: initiation, data transfer and termination. The initiation phase is started by either of the two parties and often includes some level of authentication. The data transfer phase pushes the data towards the Big Data Application Provider. The termination phase checks whether the data transfer has been successful and logs the data exchange.

Big Data Application Provider

The Big Data Application Provider is the architecture component that contains the business logic and functionality that is necessary to transform the data into the desired results. The common objective of this component is to extract value from the input data, and it includes the following activities:

- Collection;
- Preparation;
- Analytics;
- Visualization;
- Access.

The extent and types of applications (i.e., software programs) that are used in this component of the reference architecture vary greatly and are based on the nature and business of the enterprise. For financial enterprises, applications can include fraud detection software, credit score applications or authentication software. In production companies, the Big Data Application Provider components can be inventory management, supply chain optimization or route optimization software.

Big Data Framework Provider

The Big Data Framework Provider has the resources and services that can be used by the Big Data Application Provider, and provides the core infrastructure of the Big Data Architecture. In

this component, the data is stored and processed based on designs that are optimized for Big Data environments.

The Big Data Framework Provider can be further sub-divided into the following three sub-roles:

1. **Infrastructure**: networking, computing and storage
2. **Platforms**: data organization and distribution
3. **Processing**: computing and analytic

Most Big Data environments utilize distributed storage and processing (discussed in Chapter 4.3) and the Hadoop open source software framework (discussed in Chapter 4.4) to design these sub-roles of the Big Data Framework Provider.

The **infrastructure layer** concerns itself with networking, computing and storage needs to ensure that large and diverse formats of data can be stored and transferred in a cost-efficient, secure and scalable way. At its very core, the key requirement of Big Data storage is that it is able to handle very massive quantities of data and that it keeps scaling with the growth of the organization, and that it can provide the input/output operations per second (IOPS) necessary to deliver data to applications. IOPS is a measure for storage performance that looks at the transfer rate of data.

The **platform layer** is the collection of functions that facilitates high performance processing of data. The platform includes the capabilities to integrate, manage and apply processing jobs to the data. In Big Data environments, this effectively means that the platform needs to facilitate and organize distributed processing on distributed storage solutions. One of the most widely used platform infrastructure for Big Data solutions is the Hadoop open source framework (as discussed in Chapter 4.3). The reason Hadoop provides such a successful platform infrastructure is because of the unified storage (distributed storage) and processing (distributed processing) environment.

The **processing layer** of the Big Data Framework Provider delivers the functionality to query the data. Through this layer, commands are executed that perform runtime operations on the data sets. Frequently, this will be through the execution of an algorithm that runs a processing job. In this layer, the actual analysis takes place. It facilitates the 'crunching of the numbers' in order to achieve the desired results and value of Big Data.

Data Consumer

Similar to the Data Provider, the role of Data Consumer within the Big Data Reference Architecture can be an actual end user or another system. In many ways, this role is the mirror image of the Data Provider. The activities associated with the Data Consumer role include the following:

- Search and Retrieve;
- Download;
- Analyze Locally;
- Reporting;

- Visualization;
- Data to Use for Their Own Processes.

The Data Consumer uses the interfaces or services provided by the Big Data Application Provider to get access to the information of interest. These interfaces can include data reporting, data retrieval and data rendering.

4.3 Distributed Data Storage and Processing

Traditional Data Analysis — Local Storage and Processing

What is it that differentiates a Big Data environment from a traditional enterprise IT environment? Before we dive into the distributed architecture of Big Data solutions, we first take a look at the way 'traditional' data analysis is performed. By looking at this structure first, we can make a comprehensive comparison with the way Big Data environments are set up.

Traditional data analysis — as performed by millions or organizations every day — has a fairly straightforward and static design. Most enterprises create structured data with stable data models via a variety of enterprise applications, such as CRM, ERP and various financial systems.[35] Various data integration tools subsequently use extract, transform and load (ETL) operations to load the data from these enterprise applications to a centralized data warehouse.

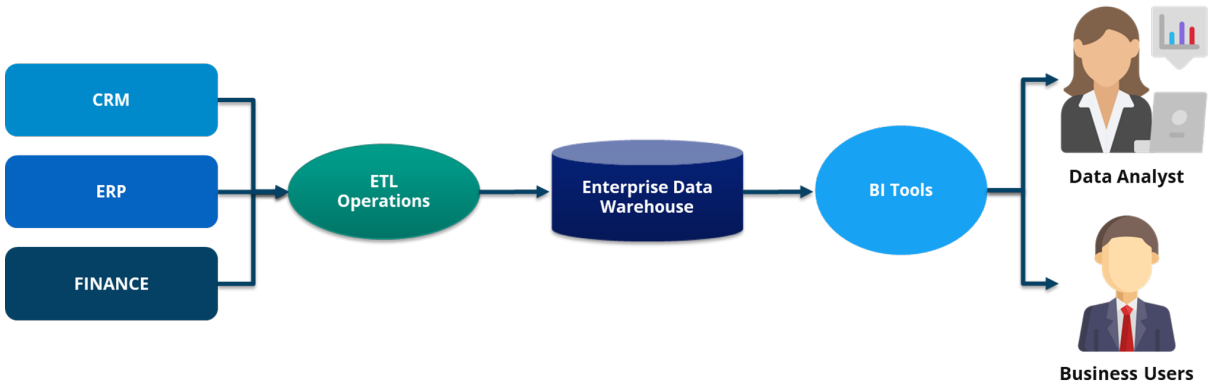

Figure 4.2: Traditional data analysis environment

In the data warehouse, the different data (originating from the different applications) are neatly stored into databases with structured rows and columns (similar to a large Excel sheet). Due to this neat structuring, a Business Intelligence analysis tool (examples are SAP Business Objects or IBM Cognos) can subsequently run queries and provide reports that provide the requested information and insights. The data volumes in data warehouses rarely exceed multiple terabytes as large data volumes degrade performance.

Big Data Environments — Distributed Storage and Processing

As discussed in Chapter 1.3, the volume and variety characteristics differentiate Big Data from traditional data analysis. Most data originates from a variety of different sources (not only enterprise data) and does not conform to relational database structures. These data sources exceed multiple terabytes.

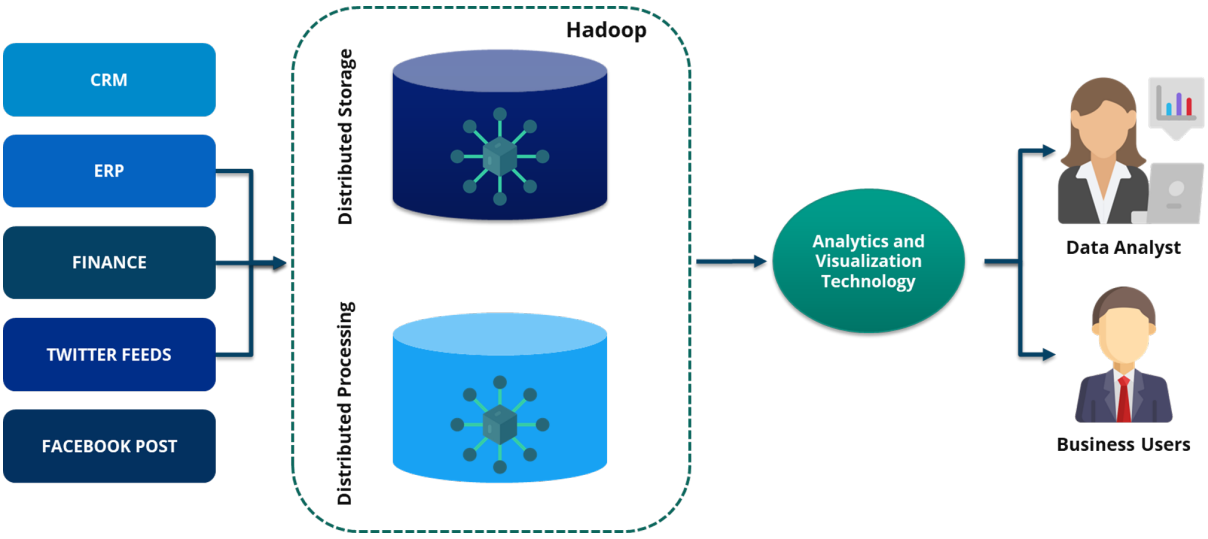

Figure 4.3: Big Data analysis environment

In order to deal with the size (volume) and disparity (variety) of this data, a different architecture is necessary to ensure that the performance levels are maintained, and the processing of Big Data brings actual value to the enterprise. To achieve these objectives, Big Data architectures commonly adhere to the following four core design principles:

1. The use of **commodity hardware nodes** to enable scale out;
2. The use of **distributed storage** to store structured, unstructured or semi-structured data;
3. The use of **distributed processing techniques** to enable parallel processing;
4. The use of advanced **analytics and visualization** technology.

These four design principles of a Big Data architecture ensure that large amounts of data can be processed efficiently and effectively. To analyze Big Data, structured data (CRM, ERP and financial data) and unstructured data (Twitter feeds and Facebook posts) are broken into "parts," which are then loaded into a distributed storage system made up of multiple nodes running on commodity hardware.[36]

In order to process the distributed data, each "part" is subsequently analyzed within the cluster itself. Rather than bringing all the data to one central location, processing occurs at each node simultaneously and therefore parallel to each other. The local processing of data within the nodes is called distributed processing. Finally, analytics and visualization technology can be used to display the end result.

The four design principles are embedded in the Hadoop open-source software framework, which will be further discussed in the next section.

4.4 Big Data Storage

The massive volume and growth of data imposes advanced requirements on storage and management. In this section, we will further discuss the storage of Big Data (a.k.a. the infrastructure layer from the NIST architecture). Big Data storage refers to the storage and management of large scale data sets while achieving reliability and availability of data accessing.[37] Large scale data sets have a distinct impact on the design of storage systems as well as storage mechanisms.

As discussed in Chapter 4.3), in traditional data analysis environments, the data storage that is used to store and retrieve data from CRM, ERP or finance systems are structured Relational Data Base Management Systems (RDBMS). However, because of the unstructured nature of Big Data, different storage systems are required.

Storage in Big Data environment is a complex subject because of two opposing forces that apply. On one hand, the storage infrastructure needs to provide information storage with reliable storage space. On the other hand, it must provide a powerful access interface for query and analysis of Big Data sets.

Storage Systems for Massive Data Sets

Various storage systems exist to meet the demands of massive data. Existing technologies can be classified as either Direct Attached Storage (DAS) or network storage, whereas network storage can be further subdivided into Network Attached Storage (NAS) and Storage Area Networks (SAN).

- **Direct Attached Storage (DAS)**. Direct attached storage is digital storage directly attached to the computer accessing it, as opposed to storage accessed over a computer network. Examples of DAS include hard drives, solid-state drives, optical disc drives, and storage on external drives.
- **Network Attached Storage (NAS)**. Network-attached storage is a file-level computer data storage server connected to a computer network providing data access to a heterogeneous group of clients. NAS is specialized for serving files either by its hardware, software, or configuration. It is often manufactured as a computer appliance — a purpose-built specialized computer.
- **Storage Area Network (SAN)**. A Storage Area Network is a network which provides access to consolidated, block level data storage. SANs are primarily used to enhance storage devices, such as disk arrays, tape libraries, and optical jukeboxes, accessible to servers so that the devices appear to the operating system as locally attached devices.

Direct Attached Storage is only suitable on a small scale (the storage needs to be physically attached to the computer). Most enterprises therefore use network storage (in data centres) to accommodate their storage requirements.

Components of Distributed Storage

Distributed storage system uses computer networks to store information on more than one node, often in a replicated fashion.[38] It is usually specifically used to refer to either a distributed database where users store information on a number of nodes, or a computer network wherein users store information on a number of peer network nodes. In this section, we summarize the key components that are required to realize distributed storage.

Components of distributed storage for Big Data may be classified into three bottom-up levels:

1. **File systems**. A file system is used to control how data is stored and retrieved. Without a file system, information placed in a storage medium would be one large body of data with no way to tell where one piece of information stops and the next begins. By separating the data into pieces and giving each piece a name, the information is easily isolated and identified. Taking its name from the way paper-based information systems are named, each group of data is called a "file". The structure and logic rules used to manage the groups of information and their names are called "file systems". Important files systems in Big Data are the Google File System (GFS) and the Hadoop Distributed File System (HDFS), both of which are expandable distributed files systems.
2. **Databases**. Traditional relational databases (RDBS) cannot meet the challenges on categories and scales that are required for Big Data. For this reason, NoSQL databases are becoming more and more popular as the core technology for Big Data storage. NoSQL database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.[39] NoSQL systems are also sometimes called "Not only SQL" to emphasize that they may support SQL-like query languages. Important examples of NoSQL databases include DynamoDB (Amazon), Voldemort (LinkedIN), BigTable (Google), Cassandra (Facebook), Azure DB (Microsoft), HBase (open source) and MongoDB (open source).
3. **Programming models**. Big Data are generally stored on hundreds or thousands of commercial servers. In order to access the data stored on these servers, parallel programming models have been developed that increase the performance of NoSQL databases. The most important examples in Big Data are MapReduce (further discussed in Chapter 4.7), Dryad (used by Microsoft) and Pregel (used by Google).

4.5 Big Data Analysis Architecture

In order to process large volumes of data, different architectures can be designed for Big Data analysis. The most important distinction (from an architecture point of view) is the difference between real-time analysis and offline analysis:

1. **Real-time analysis**. Real-time analysis is predominantly used in e-commerce and finance applications, where changes in data need to be processed imminently. Since data changes constantly,

updated results are required on a real-time basis. The analysis of credit-card transactions, for example, would require this type of architecture. The main existing architectures of real-time analysis include parallel processing clusters using traditional relational databases and memory-based computing platforms. Examples of real-time processing capability include Greenplum (EMC) and SAP HANA.

2. **Offline analysis**. Offline analysis is used for applications that are less time sensitive and for which the real-time value of data is less urgent. Offline processing (also known as batch processing) imports data on set times and subsequently processes it at time intervals. Most enterprises utilize the offline analysis architecture based on Hadoop in order to reduce costs and improve efficiency of data processing. Examples of offline analysis tools include Scribe (Facebook), Kafka (LinkedIN), TimeTunnel (Taobao) and Chukwa (Hadoop open source).

Although both real-time analysis and offline analysis provide adequate results, most enterprises utilize offline processing if the timeliness of data is not key requirement.

4.6 Hadoop Open Source Software Framework

Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.[40] Hadoop was originally created by Doug Cutting at Yahoo! and inspired by the MapReduce function developed by Google in the early 2000s for indexing web traffic. By now, Hadoop has grown to become the standard software framework for processing Big Data and is used by most major Big Data solutions providers.

The Origin of the Name Hadoop

The name Hadoop is not an acronym; it's a made-up name. The project's creator, Doug Cutting, explains how the name came about:

> "**The name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere: those are my naming criteria. Kids are good at generating such. Googol is a kid's term**".

Sub-projects and "contrib" modules in Hadoop also tend to have names that are unrelated to their function, often with an elephant or other animal theme ("Pig," for example). Smaller components are given more descriptive (and therefore more mundane) names. This is a good principle, as it means you can generally work out what something does from its name. For example, the JobTracker keeps track of MapReduce jobs.

As discussed in Chapter 4.2, the main benefit of the Hadoop software framework is that it incorporates the four design principles of a Big Data architecture. Since Hadoop is so essential to understanding how Big Data solutions work, this section will highlight its main components.

Hadoop Distributed File System (HDFS)

In order to analyze massive quantities of structured and unstructured data, data needs to be broken into "parts," which are then loaded into a distributed storage system made up of multiple nodes running on commodity hardware (see Chapter 4.2). The Hadoop Distributed File System (HDFS) is the file system that enables these data parts to be stored on different machines (commodity hardware) in a cluster. HDFS therefore enables distributed storage.

One of the core properties of the HDFS is that each of the data parts is replicated multiple times and distributed across multiple nodes within the cluster. If one node fails, another node has a copy of that specific data package that can be used for processing.[41] Due to this, data can still be processed and analyzed even when one of the nodes fails due to a hardware failure. This makes HDFS and Hadoop a very robust system.

NameNode

Since HDFS stores multiple copies of the data parts across different nodes in the cluster, it is very important to keep track of where the data parts are stored, and which nodes are available or have failed. The NameNode performs this task. It acts as a facilitator that communicates where data parts are stored and if they are available.

The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files in the file system, and tracks where across the cluster the file data is kept. It does not store the data of these files itself.

MapReduce

Once the data parts are stored across different nodes in the cluster, it can be processed. The MapReduce framework ensures that these tasks are completed by enabling the **parallel distributed processing** of the data parts across the multiple nodes in the cluster.

The first operation of the MapReduce framework is to perform a "Map" procedure. One of the nodes in the cluster requests the Map procedure — usually in the form of a Java query — in order to process some data. The node that initiates the Map procedure is labelled the Job Tracker (discussed next). The Job Tracker then refers to the NameNode to determine which data is needed to execute the request and where the data parts are located in the cluster. Once the location of necessary data parts is established, the Job Tracker submits the query to the individual nodes, where they are processed. The processing thus takes place locally within each node, establishing the key characteristic of distributed processing.

The second operation of the MapReduce framework is to execute the "Reduce" method. This operation happens after processing. When the Reduce job is executed, the Job Tracker will locate the local results (from the Map procedure) and aggregate these components together into a single final result. This final result is the answer to the original query and can be loaded into any number of **analytics and visualization** environments.

The steps of the MapReduce operations are depicted in figure 4.4.

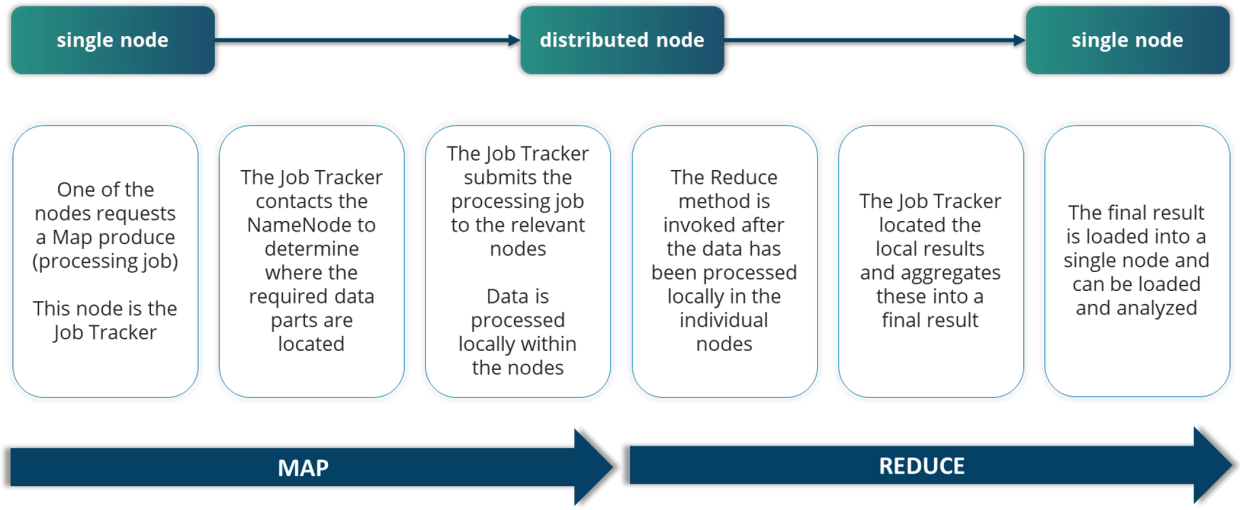

Figure 4.4: Schematic overview of MapReduce

Slave Node

Slave Nodes are the nodes in the cluster that follow directions from the Job Tacker. Unlike the NameNode, the Slave Nodes do not keep track of the location of the data.

Job Tracker

The Job Tracker – introduced in the MapReduce section — is the node in the cluster that initiates and coordinates processing jobs. Additionally, the Job Tracker invokes the Map procedure and the Reduce method.

Notes

32 Kurzweil, R., Richter, R., Kurzweil, R. and Schneider, M.L., 1990. The age of intelligent machines (Vol. 579). Cambridge: MIT press.

33 Turing, A.M., 2009. Computing machinery and intelligence. In Parsing the Turing Test (pp. 23-65). Springer, Dordrecht.

34 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice Hall.

35 Chui, M., 2017. Artificial intelligence the next digital frontier?. McKinsey and Company Global Institute, p.47.

36 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice hall.

37 Baral, C. and De Giacomo, G., 2015, January. Knowledge Representation and Reasoning: What's Hot. In AAAI (pp. 4316-4317).

38 Aron, A. and Aron, E.N., 1994. Statistics for psychology. Prentice-Hall, Inc.

39 Chambers, J.M., 2018. Graphical methods for data analysis. CRC Press.

40 White, T., 2012. Hadoop: The definitive guide. O'Reilly Media, Inc.

41 Borthakur, D., 2008. HDFS architecture guide. Hadoop Apache Project, 53.

5. Big Data Algorithms

5.1 Introduction to Algorithms

Big Data is the knowledge domain that explores the techniques, skills and technology to deduce valuable insights out of massive quantities of data. In this chapter, the focus will be primarily on the techniques that are necessary to deduce these insights from the data.

In order to find 'value' in datasets, data scientists apply algorithms. Algorithms are unambiguous specifications on how to solve a class of problems. Algorithms can perform calculation, data processing and automated reasoning tasks. By applying algorithms to large volumes of data, valuable knowledge and insights can be obtained. A very basic example of an algorithm – that finds the maximum value in a set of data - is depicted in the the code box below.*

```
numbers = [9, 34, 11, -4, 27]

# find the maximum number
max_number = max(numbers)
print(max_number)

# Output: 34
```

Algorithms can vary from very simple with only a few lines of code, to very sophisticated and complex, with millions of lines of code. In this chapter, we start with the basic operations behind algorithms. More advanced and complex examples are discussed in the Enterprise Big Data Scientist Guide.

The application of algorithms, and its subsequent use for Big Data, is grounded in the scientific domain of statistics. Everyone involved in data science should therefore have a fundamental knowledge about statistical operations and how they could be applied in algorithms. This chapter will therefore discuss essential statistical operations and provide common algorithms that are used in Big Data analysis and analytics solutions.

5.2 Desciptive Statistics

Descriptive statistics are summary statistics that quantitatively describe or summarize features of a collection of information.[42] Descriptive statistics provide key values that quickly summarize datasets and are understandable for everyone that is working with the data.

*We will be using Python to illustrate examples in this publication, because it is the most easy language to learn, and most widely used language in Data Science. If you want to follow along, you can use Google Colab or any other Python editor.

For example, the shooting percentage in basketball is a descriptive statistic that summarizes the performance of a player or a team. This number is the number of shots made divided by the number of shots taken. For example, a player who shoots 33% is making approximately one shot in every three. The percentage summarizes or describes multiple discrete events, and everyone can compare the statistic to the shooting percentages of other players.

In this guide, we will explain the following descriptive statistics:

1. Central Tendency Statistics
2. Dispersion Statistics
3. Distribution Shapes

Each descriptive statistic is illustrated with a short example that explains how the statistic should be calculated, so that it can subsequently be used in the development of Big Data Algorithms.

Central Tendency Statistics

Central tendency statistics (or measures of central tendency) are typical for defining values in data sets. These statistics describe how various data points are organized around its central point. The most common measures of central tendency are the mean, the median, and the mode.

Mean

The arithmetic mean (or simply "mean") of a sample is of the sum of the sampled values divided by the number of items in the sample. In the example below, the mean is calculated for a group of basketball players.*

```
# A list with the age of basketball players
playerAge = [28, 24, 27, 34, 19, 28]

# Sum each value and divide by the total values
mean = sum(playerAge) / len(playerAge)
print(mean)

# Output: 26.666666666666668
```

Median

The median is the value separating the higher half of a data sample, a population, or a probability distribution, from the lower half. For a data set, it may be thought of as the "middle" value. In the example below, the median is calculated for another group of basketball players.†

*Note that it is obviously way easier to use commonly available functions, like the `mean()` function in the `statistics` package. However, since this is important to udnerstand the fundamentals, we will showcase how to actually calculate these statistics.

†Similar to the previous comment, you can als use the `median()` function fromt the statistics package.

```
# A list with the age of basketball players
playerAge = [28, 24, 27, 34, 19, 28, 27]

# Sort the data low-to-high or high-to-low
playerAge.sort()
print(playerAge)
# Output: [19, 24, 27, 27, 28, 28, 34]

# Select the middle value from the list
middleIndex = int((len(playerAge) - 1)/2)
print(middleIndex)

median = playerAge[middleIndex]
print(median)

# Output middelIndex: 3
# Output middelIndex: 27
```

If the number of variables is even, the median is calculated by selecting the middle two values and dividing these by two. One of the core properties of the median is that is not greatly affected by outliers in the data set – extreme values will not have a large impact on the median.

Mode

The mode of a set of data values is the value that appears most often. In other words, it is the value that is most likely to be sampled. Similar to the mean and median, the mode can provide key information about a data set. In the example below, the mode is determined for a group of basketball players.*

```
# Import the Counter function
from collections import Counter

# A list with the age of basketball players
playerAge = [28, 24, 28, 34, 19, 28, 27]

def my_mode(sample):
  c = Counter(sample)
  return [k for k, v in c.items() if v == c.most_common(1)[0][1]]

my_mode(playerAge)

# Output: 28
```

*This is a slightly more difficult implementatation, so try if you can follow. Ultimately, you need to count how many times a value appears in a list (hence the counter).

In case two variables are both appearing as most frequent, the dataset can be classified as bimodal. If more than two variables are appearing as most frequent, the dataset can be labeled as multi-modal.

Dispersion Statistics

In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed. Dispersion statistics indicate how data points are distributed around its centre values. Common examples of measures of statistical dispersion are the range, interquartile range, variance and standard deviation.

Range

The range of a set of data is the difference between the largest and smallest value in the dataset.

```
# A list with the age of basketball players
playerAge = [28, 24, 28, 34, 19, 28, 27]

# The list is sorted from smallest to highest value
playerAge.sort()
print(playerAge)
# Output: [19, 24, 27, 28, 28, 28, 34]

# Find the smallest (first) and highest (last) value in the list
firstValue = playerAge[0]
lastValue = playerAge[-1]
print(firstValue)    # Output: 19
print(lastValue)     # Output: 34

# Determine the range from the list
range = lastValue - firstValue
print(range)         # Output: 15
```

Interquartile Range

The interquartile range (IQR), also called the mid-spread or middle 50%. To calculate the IQR, the data set is divided into four quartiles, or four rank-ordered even parts via linear interpolation. These quartiles are denoted by Q1 (also called the lower quartile), Q2 (the median), and Q3 (also called the upper quartile). The lower quartile corresponds with the 25th percentile and the upper quartile corresponds with the 75th percentile, so IQR = Q3 − Q1.

In other words, the IQR is a statistic that indicates where the middle 50% of values are located, as per the example below:

```
# Import the numpy package
import numpy as np

# A list with the age of basketball players
playerAge = [19, 19, 20, 24, 27, 28, 28, 31, 34]

# The IQR is determined by subtracting Q1 from Q3
Q2 = np.percentile(playerAge, 50)
print(Q2)       # Output: 27

Q1 = np.percentile(playerAge, 25)
print(Q1)       # Output: 20

Q3 = np.percentile(playerAge, 75)
print(Q3)       # Output: 28

# The IQR is the difference between Q3 and Q1
IQR = Q3 - Q1
print(IQR)      # Output: 8
```

The IQR range is a very useful statistic in data science, because it does not include extreme values (outliers). Extreme values in data sets are commonly generated by corrupt data. Consequently, the IQR is an adequate measure to eliminate outliers.

Variance

Variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are spread out from their average value. The closer the variance is to zero, the more closely the data points are clustered together.

The official formula for variance is as follows: $\sigma^2 = \frac{\Sigma(X=mu)^2}{N}$

The variance of a data set is calculated by following these steps:

1. Subtract the mean from each value in the data. This gives you a measure of the distance of each value from the mean.
2. Square each of these distances (so that they are all positive values), and add all of the squares together.
3. Divide the sum of the squares by the number of values in the data set.

In the example below, the variance is calculated for the age of a group basketball players:

```python
# Importing the statistics module
import statistics

# A list with the age of basketball players
playerAge = [19, 19, 20, 25, 27, 28]

# In step 1, we subtract the mean from each value
mean = statistics.mean(playerAge)
print(mean)  # Output: 23

playerAgeMinusMean = [x - mean for x in playerAge]
print(playerAgeMinusMean)  # Output: [-4, -4, -3, 2, 4, 5]

# In step 2, we square each value and sum the values
playerAgeMinusMeanSquared = [x * x for x in playerAgeMinusMean]
print(playerAgeMinusMeanSquared) # Output: [16, 16, 9, 4, 16, 25]

sumOfSquaredValues = sum(playerAgeMinusMeanSquared)
print(sumOfSquaredValues)  # Output: 86

# In step 3, the variance is the sum of squared value divided by the number
# of observations in the list
variance = sumOfSquaredValues / len(playerAge)
print(variance)  # Output: 14.333333333333334
```

Standard Deviation

The standard deviation (SD, also represented by the Greek letter sigma σ or the Latin letter s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.[43]

A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values. The difference between a low- and a high standard deviation is depicted in figure 5.1.

The standard deviation can be calculated by taking the square root of the variance. A useful property of the standard deviation is that, unlike the variance, it is expressed in the same units as the data.

The official formula for standard deviation is: $\sigma = \sqrt{\frac{\Sigma(X=mu)^2}{N}}$

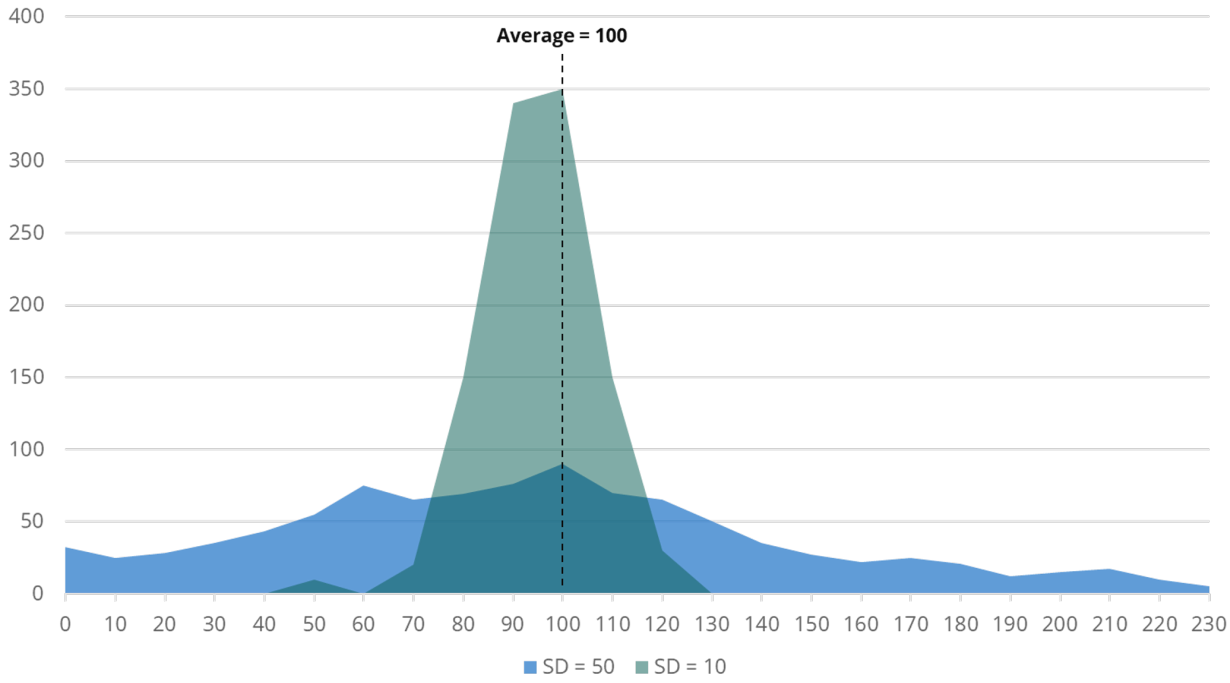

Figure 5.1: Difference between low and high standard deviation

The way to calculate the standard deviation is exactly the same as the variance, with the difference of taking the square root of the variance. The example below calculates the standard deviation for the same data set of basketball players.

```
# Importing the statistics module
import statistics
import math

# A list with the age of basketball players
playerAge = [19, 19, 20, 25, 27, 28]

# In step 1, we subtract the mean from each value
mean = statistics.mean(playerAge)
playerAgeMinusMean = [x - mean for x in playerAge]

# In step 2, we square each value and sum the values
playerAgeMinusMeanSquared = [x * x for x in playerAgeMinusMean]
sumOfSquaredValues = sum(playerAgeMinusMeanSquared)

# In step 3, the variance is the sum of squared value divided by the number
# of observations in the list
variance = sumOfSquaredValues / len(playerAge)

```

```
# As a final step the standard deviation is the square root of the variance
sd = math.sqrt(variance)
print(sd)  # Output: 3.7859388972001824
```

Distribution Shapes

A frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval. In other words, it shows how values of a variable are distributed.

In Big Data analysis and analytics, a number of common distributions are used:

- Frequency distribution
- Probability distribution
- Sampling distribution
- Normal distribution

Frequency Distribution

A frequency distribution is a table or a graph displaying the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

In Big Data, a frequency distribution provides a visual representation of the distribution of observations within a particular samples. Big Data Analysts often use a frequency distribution to visualize or illustrate the data collected in a sample. For example, the age of basketball players can be split into several different categories or ranges.

Group	Frequency
15 or younger	15
16-19	12
20-23	40
24-26	20
27-30	34
31-34	15
35 or older	6

The above frequency distribution can be visually displayed as depicted in figure 5.2.

Age of Basketball Players	
< 15	5
16 - 19	12
20 - 23	40
24 - 26	20
27 - 30	34
31 - 34	15
> 35	6

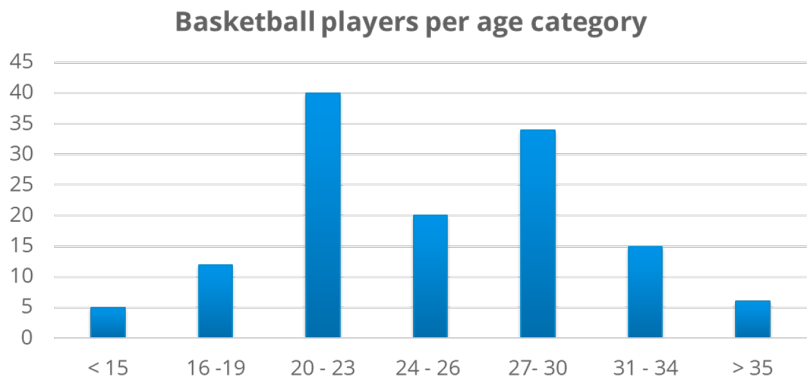

Figure 5.2: Example of a frequency distribution

Probability distribution

A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. A probability is the 'chance' or likelihood that a certain outcome will happen. The probability that a coin flip will be tails is 0.5 (for a fair coin), which indicates that there is a 50% that the coin will show tails in a future coin flip.

A probability distribution is a summary graph that depicts the likelihood of all potential outcomes. It is a mathematical function that can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. The distribution here shows probabilities on the y-axis.

For example, the following probability distribution tells us the probability that a certain soccer team scores a certain number of goals in a given game:

Goals	Probability
0	0.12
1	0.22
2	0.38
3	0.10
4	0.09
5	0.07
6	0.02

The above probability distribution can be visually displayed as depicted in figure 5.3.

Chance of scoring in match	
Player 1	15%
Player 2	20%
Player 3	5%
Player 4	25%
Player 5	20%
Player 6	15%

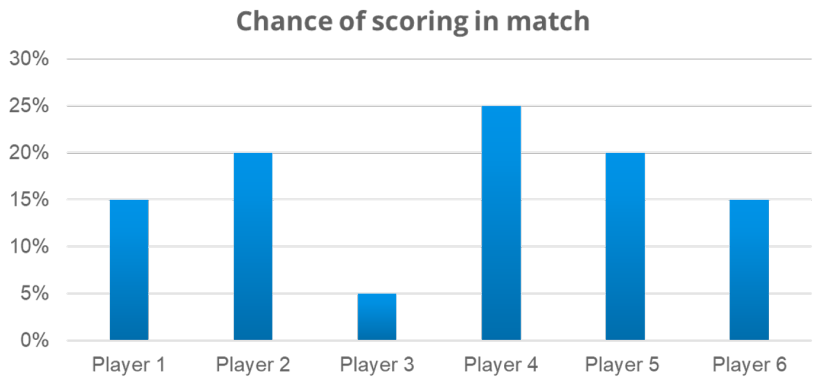

Figure 5.3: Example of a probability distribution

One of the core properties of the probability distribution is that all potential outcomes should sum to 100 percent. In other words, the area under the curve of the probability distribution should always be 1.0 (which is the same as 100% of all probabilities).

Probability distributions are widely used in Big Data, because one of the primary aims is to predict certain outcomes. When banks, for example, provide new credit cards to potential clients, they aim to minimize the risk that the client will default, whilst optimizing the chance that the client will become a profitable customer.

Sampling Distribution

A sampling distribution is the probability distribution of a given statistic based on a random sample. A sample is a subset of a population. Sampling distributions are important in Big Data because they provide a major simplification that can be used for predictive analytics. More specifically, sampling distributions (e.g., Big Data set) allow population inferences to be based on the Big Data set, rather than on the joint probability distribution of all the individual sample values.

Normal Distribution

The normal (or better known as Gaussian) distribution is the most important and common continuous probability distribution. Normal distributions are important in statistics and are often used in data science to represent real-valued random variables whose distributions are not known.[44]

A normal distribution represents data that occurs commonly where most values are the same as the average value and only few values are found at the extremities. In a normal distribution, approximately 99.7% of the values are within three standard deviations of the mean, and the area under the curve is equal to one.

A normal distribution has the same mean, median, and mode.

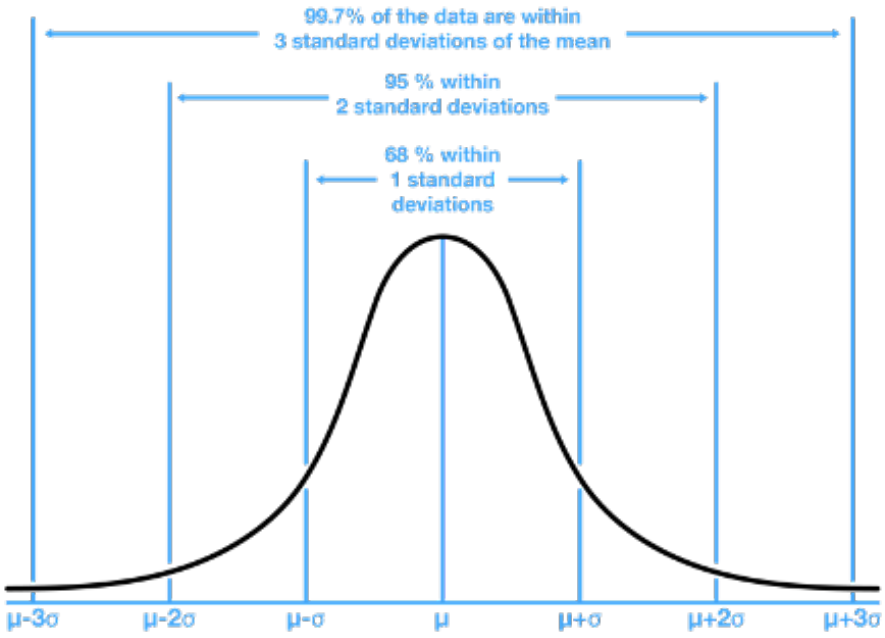

Figure 5.4: Properties of the normal distribution

Skewness

Skewness is a measure of the asymmetry of a probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined.[45]

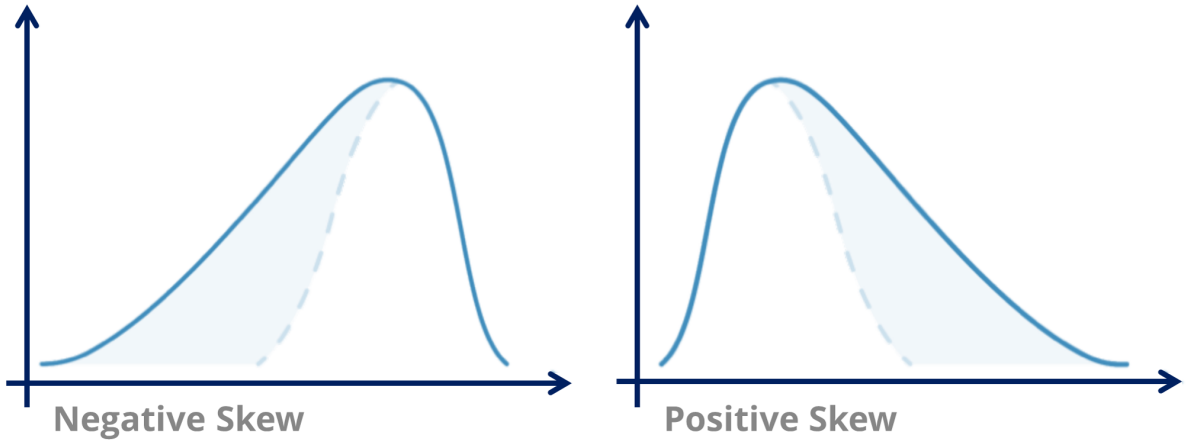

Figure 5.5: Skewness in data sets

Consider the two distributions in the figure just above. Within each graph, the values on the right side of the distribution taper differently from the values on the left side. These tapering sides are

called tails, and they provide a visual means to determine which of the two kinds of skewness a distribution has:

1. **Negative skew**: A distribution is negatively skewed when the tail of the curve is longer on the left side or skewed to the left, and the mean is less than the median and mode. The majority of the values exist on the right side of the curve.
2. **Positive skew**: A distribution is positively skewed where when the tail of the curve is longer on the right side or skewed to the right, and the mean is greater than the median and mode. The majority of the values exist on the left side of the curve.

Skewness in distributions is important in data science because the skewness can indicate potential bias (i.e., not an adequate representation the actual data) in data sets.

Standardization

Because of the properties of the standard normal distribution, it is possible to convert other distributions in terms of their number of standard deviations below or above the mean. Conversion of these distributions is called standardization. Data points of distributions are converted into standard scores (better known as z-scores).

The formula for calculating z-scores is: $z = \frac{(x-\mu)}{\sigma}$

Where μ is the mean of the population and σ is its standard deviation. An example of the process of standardization is outlined in the table below.

Player Age	Player Age (normal)	Player Age (standardized)
Player 1	19	-1.29
Player 2	19	-1.29
Player 3	20	-1.09
Player 4	24	-0.30
Player 5	27	0.28
Player 6	28	0.46
Player 7	28	0.46
Player 8	31	1.07
Player 9	34	1.66
Mean	25.56	0
SD	5.10	1

As can be seen from the table above, the mean of the standardized value is always zero and the standard deviation of the standardized values is always 1. Additionally, z-scores are dimensionless and can easily be compared with other z-scores. Because of these properties, standardized values can be allowed in algorithms.

Suppose for example that there is a dataset with different kinds of properties of basketball players (height, age, average score per game, number of passes, etc.). When comparing the means and standard deviations of each of these properties, they are impossible to compare because they have

different dimensions. However, when standardized values are used, all values are displayed equally.

Standardization is one of the most important processes when analyzing Big Data, because they allow different variables to be combined together. Standardized values are almost always used in the design and execution of algorithms. An in-depth use of the application of standardization is discussed in the Enterprise Big Data Scientist guide.

5.3 Statistical Inference

Statistical inference is the process of deducing properties of a sample of the data (i.e., a probability distribution) in order to make predictions about the entire group of data.[46] Inferential statistics deduces that if certain characteristics of a sample can be proven, these characteristics are also likely also to be present in the entire population.

If for example, a study of 500 basketball athletes shows that 99% of all basketball players in the NBA are taller than 1.95 meters, it could be inferred that 99% of all basketball player in the NBA are taller than 1.95 meters. This would be a statement based on inferential statistics. Whether this statement is true depends on the questions whether the sample data is a representative subset of the entire population.

Populations and Samples

A population is a set of similar items or events which is of interest for some experiment. It is the entire group that is of interest to the experiment. From a data collection perspective, however, it is frequently not possible to have the entire population.

A sample is a subset of the population that is being analyzed, and about which the data is available. The elements of a sample are known as sample points, sampling units or observations. Statistical analysis and algorithms are applied on the sample data in order to make assumptions and statements about the entire population.

In the example of our basketball players, the difference could be as follows:

- The population is determined as all the basketball players in the world.
- A sample has been collected of 10.000 people who play basketball by conducting a survey at basketball clubs.

One of the most important considerations in sampling is to ensure that the sample is a representative subset of the entire population. Otherwise, there could be bias.

Bias

If the sample that has been selected is not an adequate representation of the entire population, it is called a biased sample. Bias will result in inadequate or wrong predictions about the future, because in inferential statistics, assumptions are made about the entire population based on the sample.

Consider the following example about a biased sample:

- In order to determine shirt size for basketball players around the world, a sports apparel company is interested to know the average player height of basketball players in the world.
- One of the datasets that is readily available is the dataset of player characteristics of the professional basketball players in the National Basketball Association (NBA) in the United States.
- Based on this sample, it is concluded that 99% of basketball players in the NBA are above 1.95 meters tall.

Although all the calculations have been made correctly, the sports apparel company would make a big mistake to produce only shirt sizes for players that are 1.95 meters tall, because the sample is biased. The sample only considered NBA players, which is a very small subset of the most successful basketball players in the world. Within this small group, the average length might be 1.95 meters, but this does not represent the average height of the average basketball player in the world. Most incorrect predications in statistics are made when the sample data is biased.

Populations, Samples and Bias in Big Data

As discussed in the previous section, bias can result in inadequate predictions because the sample does not always represent the population. The only way to completely eliminate bias is when the sample is the same as the actual population.

With Big Data, it becomes possible to analyze massive quantities of data. The larger the data set becomes, the closer it is to the actual population, and the less likely it is that the data set becomes biased. In other words, because of Big Data, predictions about the future become more and more accurate.

In the same example that we used before, instead of analyzing the dataset of NBA players, the sports apparel company could also choose to analyze the data of all members of basketball clubs throughout the world. If the company would have access to this massive quantity of data, it could fairly accurately predict player shirt sizes by looking at the average length.

5.4 Correlation

Dependence (or association) is any statistical relationship, whether causal or not, between two random variables or bivariate data. Correlation is any of a broad class of statistical relationships involving dependence, although it is mostly used to indicate whether two variables have a linear relationship. An example of correlation is the relationship between the height of basketball players and their selection for tryouts in the NBA.

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, in order to make purchasing decisions or forecast future sales. However, the presence of a correlation is not sufficient to infer the presence of a causal relationship (i.e., correlation does not imply causation).

In correlation, two (or more) variables are compared to each other. These variables can either be dependent or independent:

- **Independent variables** are not changed or affected by changes in the other variable. They operate independently, and are frequently changed to test the effect on the dependent variable. Common examples of independent variables are temperature, age, or the height of basketball players.
- **Dependent variables** are the variables that change based on the fluctuations of the independent variable. The dependent variables represent the output or outcome whose variation is being studied. In the example above, the chance of selection for NBA tryouts is the dependent variable that we would like to know (dependent on the independent variable of 'player height').

Correlation refers to a specific relationship between two variables, and a number of different correlation coefficients. In this guide, we only discuss the most frequently used correlation — the Pearson correlation — which is sensitive to a linear relationship between two variables.

Pearson Correlation Coefficient

The Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is widely used in Big Data and Data Science to detect relationships between variables.

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient.

The formula for ρ is: $\rho = \frac{cov(X,Y)}{\sigma(x)*\sigma(y)}$

Where cov is the covariance, $\sigma(x)$ is the standard deviation of X, and $\sigma(y)$ is the standard deviation of Y.

Correlations that are close to either -1 of +1 are considered strong correlations, because the variables tend to move in similar directions. Even with strong correlations, though, always keep in mind that correlation does not imply causation.

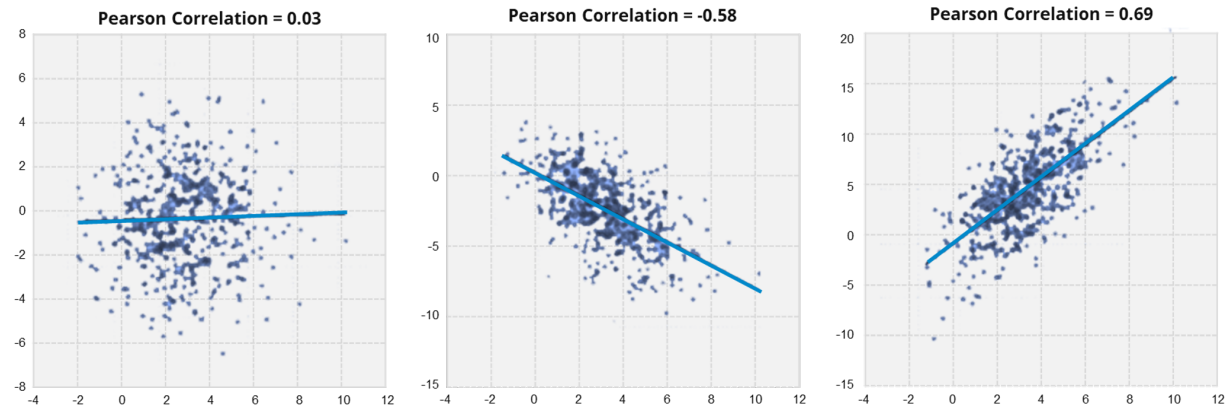

Figure 5.6: Different values of the Pearson correlation coefficient

5.5 Regression

Regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

The most important characteristic of regression is that it always aims to estimate a function of the independent variables — the **regression function**. In other words, in regression, we are trying to find the best fit line in order to make predictions (or forecasts) about the relationship between variables. Because of its predictive nature, it is widely used in machine learning in order to find relationships in datasets.

Simple Linear Regression

In simple linear regression, the objective is to find a linear relation between the dependent variable y and the independent variable x based on the following simple regression function:

Simple linear regression is expressed as: $y = \alpha\dot{x} + \beta$

where α is the slope of the best fit line and β is equal to the y-intercept. The goal is to find the values of and that would provide the best "fit" through all available data points. These values can be found by finding the minimum distance between the regression line and the actual data point (e.g.,— minimize the sum of squared errors). In the case of simple linear regression, α will be the Pearson correlation coefficient.

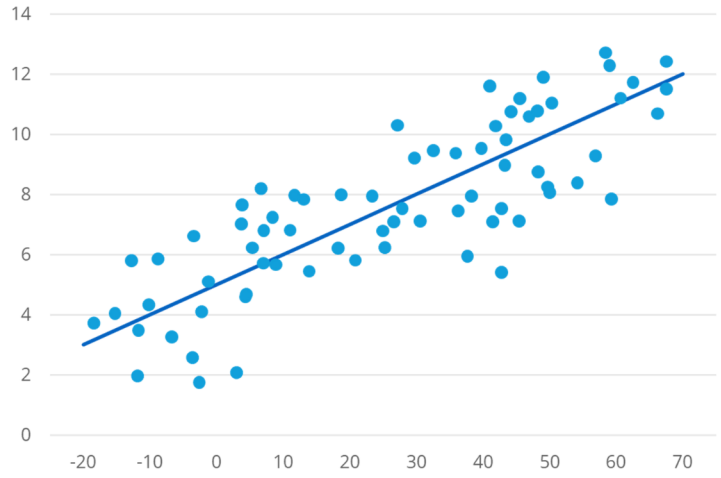

Figure 5.7: Example of simple linear regression

Similarities and Differences between Correlation and Regression

Although correlation and regression might have some elements in common, please keep in mind that both have completely different objectives. Correlation only indicates if a relationship exists, whereas regression aims to estimate the extent of this relationship for predictive purposes. We have outlined the most import differences below.

First, some important similarities between correlation and regression:

- The standardized regression coefficient is the same as Pearson's correlation coefficient;
- The square of Pearson's correlation coefficient is the same as the R2 in simple linear regression;
- Neither simple linear regression nor correlation answer questions of causality directly.

Second, some important differences between correlation and regression:

- While correlation typically refers to the linear relationship, it can refer to other forms of dependence, such as polynomial or truly nonlinear relationships;
- While correlation typically refers to Pearson's correlation coefficient, there are other types of correlation, such as Spearman's.

5.6 Classification

Classification is the problem of identifying to which of a set of categories a new observation belongs, based on a training set of data containing observations whose category membership is known. Because the computer is 'fed' sample data, classification is a form of supervised machine learning.

A classification algorithm — simplistically stated — executes the following steps:

1. A computer is fed sample data that contains information about the class of each data point. For example, it learns to classify carrots as “vegetables” and oranges as “fruits”.
2. After the ‘training’ of the machine, new data or observations are provided to the computer.
3. The computer now starts to classify by itself. In the example, edibles that have similar characteristics to carrots will be labeled “vegetables,” whereas edibles that have similar characteristics to oranges will be labeled as “fruits”.

An example would be assigning a given email into “spam” or “non-spam” classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.).

An example of the process of classification is depicted in the figure below.

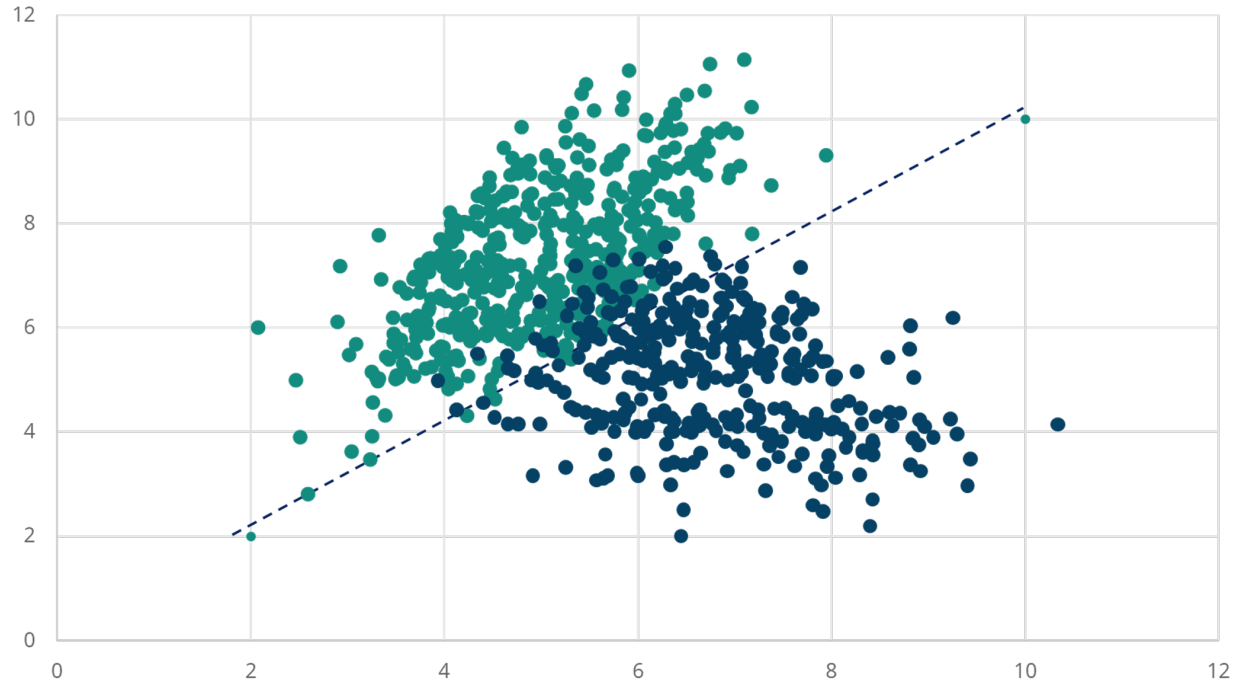

Figure 5.8: Example of binary classification

An algorithm that implements classification, especially in a concrete implementation, is known as a **classifier**. The term classifier” sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category.

5.7 Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). Unlike classification (discussed in Chapter 5.6), clustering is an example of

unsupervised learning. There is no sample data that is first "fed" into the machine, but the computer starts formulating clusters based on similarities between groups.

In order to arrive at a cluster, the computer needs to run a clustering algorithm. There are many known clustering algorithms available, depending on characteristics of the problem to be solved. A commonality is that most clustering algorithms look at the 'similarity' between data points.[47]

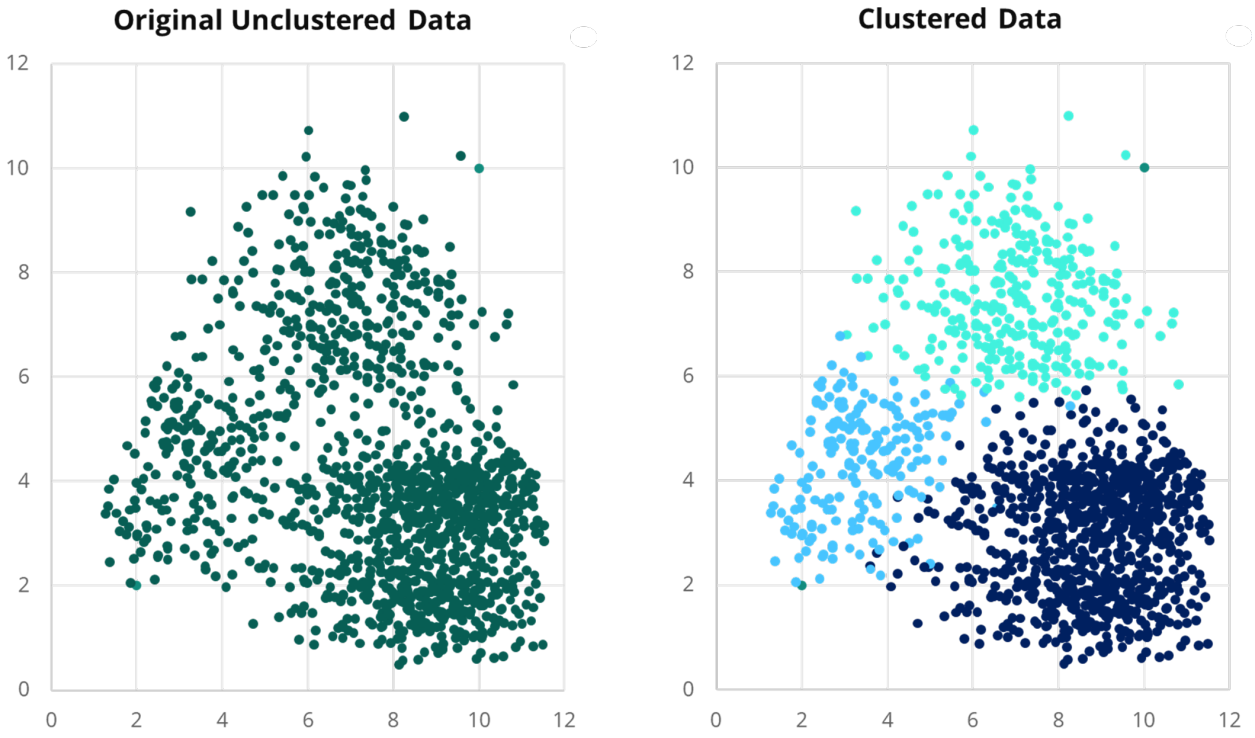

Figure 5.9: Example of a clustering algorithm

5.8 Outlier Detection

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate an error in the data. Especially in the analysis of Big Data sets, outlier detection is a frequently used technique in order to detect erroneous or false data points.

Outliers are generally data points that appear to be unexpected in comparison with the rest of the data — they do not fit into the pattern of the other data points. As discussed in Chapter 5.2, the standard normal distribution can be used to detect outliers. Remember that within the standard distribution, 99% of data point fit within three standard deviations of the mean. If one or more data points are therefore more than three standard deviations from the mean, this might be an indication that these points are incorrect or contain flawed data.

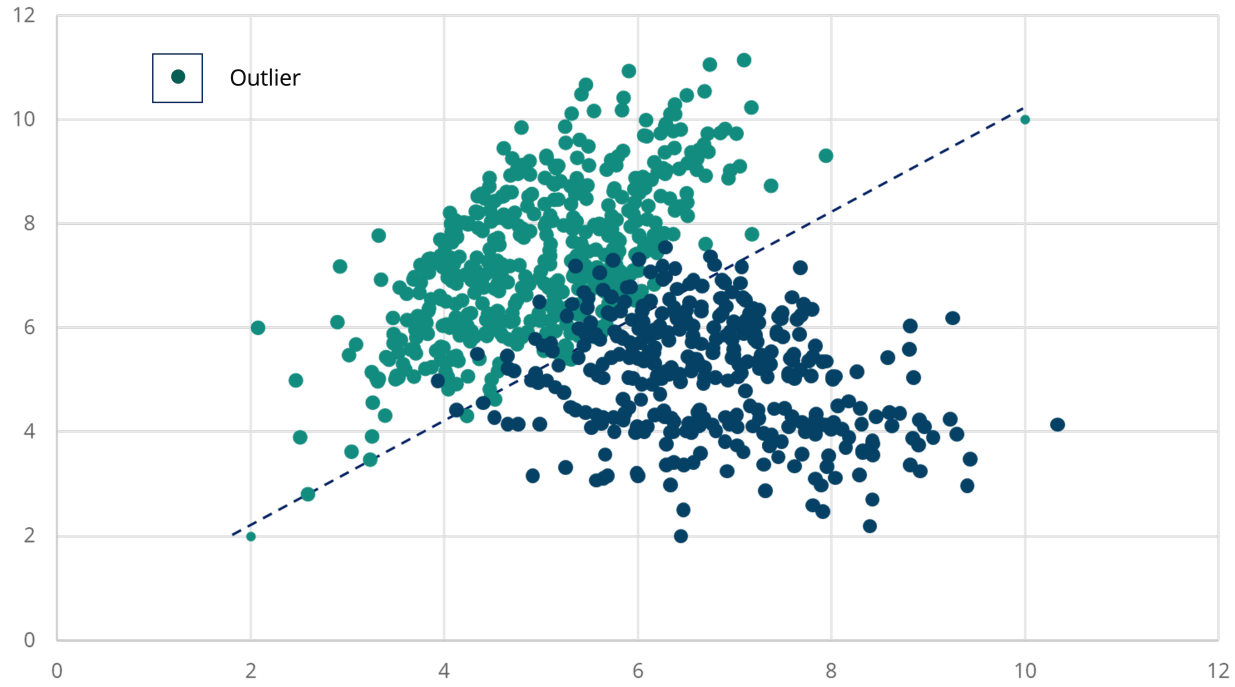

Figure 5:10: Outlier detection algorithms detect observations that do not fit the regular pattern

In machine learning algorithms — that frequently use standardized values — outlier detection can be a very useful operation, because it indicates that there is a very small possibility that this data point would be occurring.

Outlier detection is a widely used technique, especially in the context of Big Data. Insurance and credit card companies use outlier detection that detect fraudulent claims or transactions by looking at data that does not fit within the regular pattern. Similarly, outlier detection algorithms are used by intelligence agencies to detect anomalies in individual behaviors that might pose a threat to national security.

5.9 Data Visualization

Statistical graphics, also known as graphical techniques, are graphics in the field of statistics used to visualize quantitative data. Data visualization is widely used in the domain of Big Data, because it condenses large data sets to summary graphs that are easy to understand and easy to discuss.

Especially in an enterprise context, it is important to use data visualization techniques, because not everyone has a background in statistics and algorithms. In this section, we will briefly discuss the most common data visualization techniques and their properties:

1. Bar charts
2. Histograms
3. Scatter plots

4. Bi-plots
5. Box plots
6. Q-Q plots
7. Pie charts
8. Radar charts

The purpose of this section is to familiarize you with some of the most common visuzalization techniques that are used in Big Data. More detailed data visualization techniques are discussed in the Enterprise Big Data Analyst guide.

Bar Charts

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a line graph.

A bar graph depicts comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value. Some bar graphs present bars clustered in groups of more than one, showing the values of more than one measured variable.

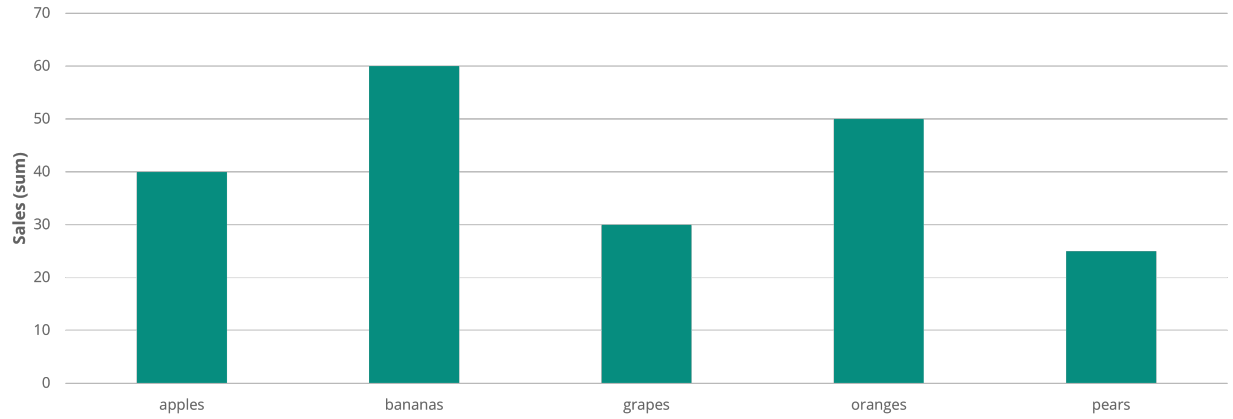

Figure 5.11: Example of a bar chart

Histograms

A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable). A histogram is similar to a bar chart, but each of the bars is connected to the other.

To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent and are often (but are not required to be) of equal size.[48]

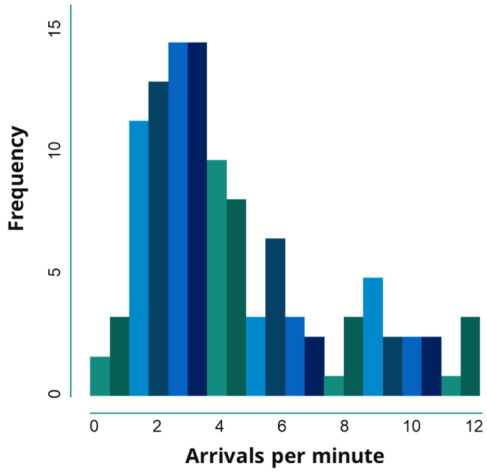

Figure 5.12: Example a of a histrogram

Scatter Plots

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables of a data set. If the points are color-coded, one additional variable can be displayed.

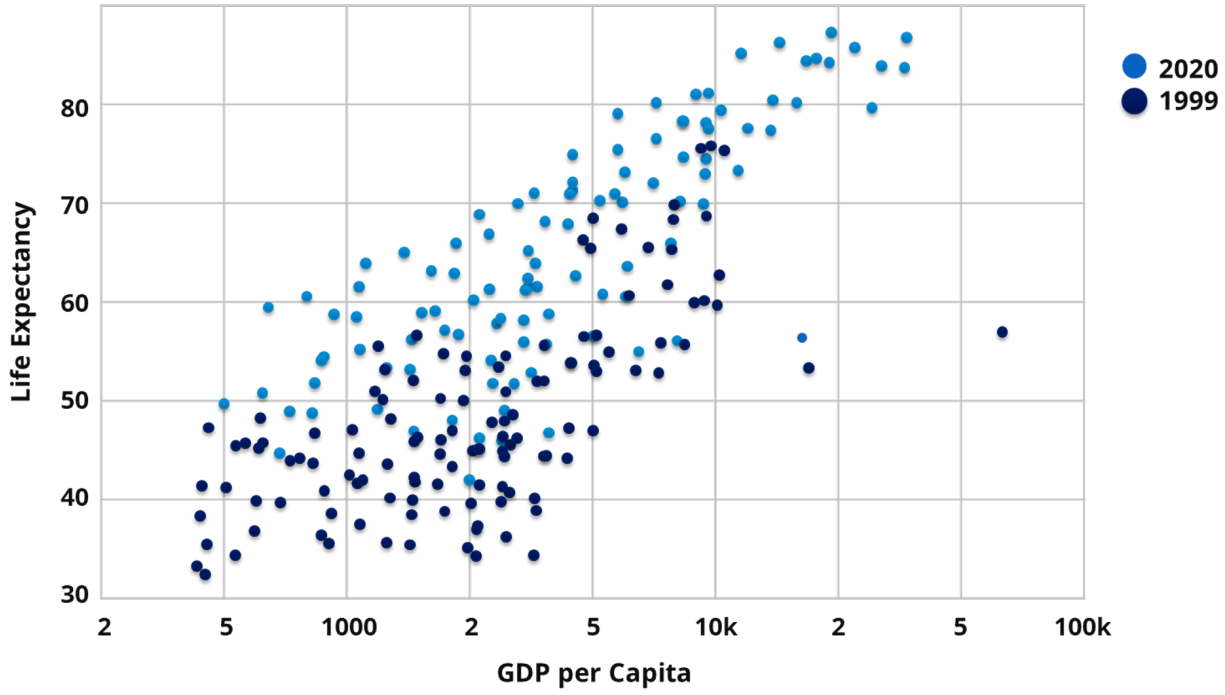

Figure 5.13: Example of a scatter plot

The data are displayed as a collection of points, each having the value of one variable determining

the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

Biplots

A Biplot is an enhanced scatterplot that makes use of both points and vectors to represent structure. A biplot uses points to represent the scores of the observations on the principal components, and it uses vectors to represent the coefficients of the variables on the principal components.

A biplot overlays a score plot and a loadings plot in a single graph. An example is shown at the right. Points are the projected observations; vectors are the projected variables. If the data are well-approximated by the first two principal components, a biplot enables you to visualize high-dimensional data by using a two-dimensional graph.

The advantage of the biplot is that the relative location of the points can be interpreted. Points that are close together correspond to observations that have similar scores on the components displayed in the plot. To the extent that these components fit the data well, the points also correspond to observations that have similar values on the variables.

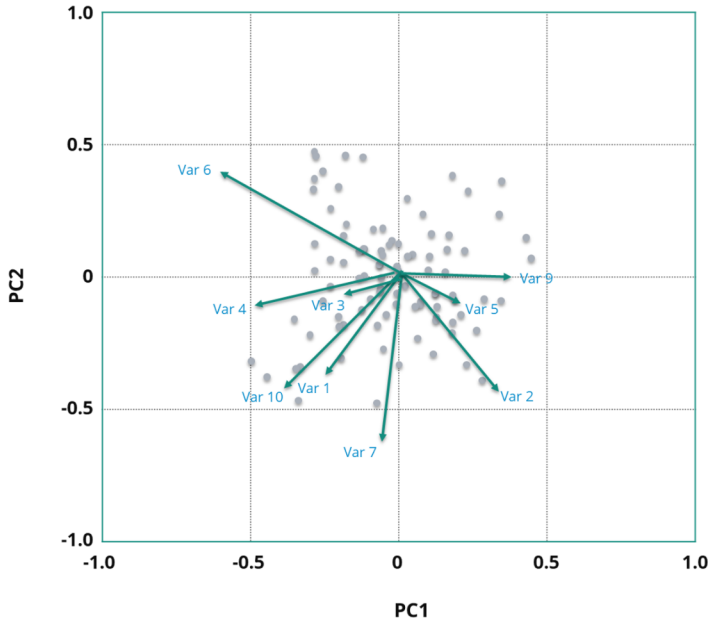

Figure 5.14: Example of a biplot

Box Plots

A box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points.

A boxplot is a standardized way of displaying the dataset based on the five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles.

A box-plot usually includes two parts, a box and a set of whiskers as shown in Figure 5.15. The box is drawn from Q1 to Q3 with a horizontal line drawn in the middle to denote the median. The whiskers can be defined in various ways. In the most straight-forward method, the boundary of the lower whisker is the minimum value of the data set, and the boundary of the upper whisker is the maximum value of the data set.

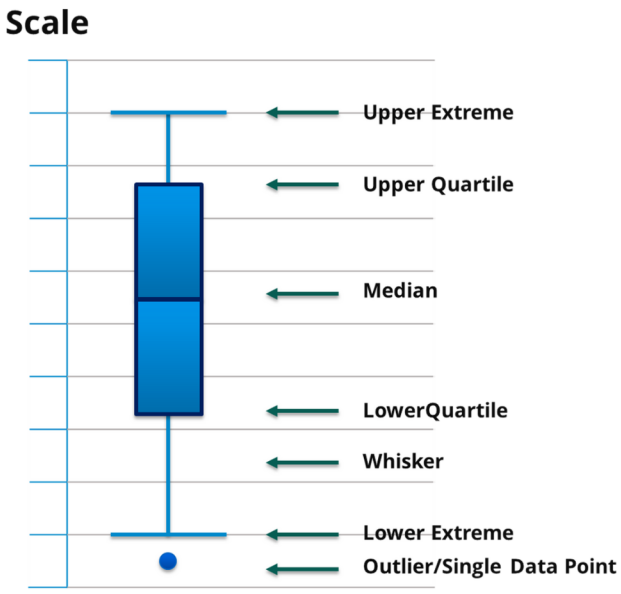

Figure 5.15: Example of a boxplot

Quantile-Quantile Plots

A Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

In data science, Q-Q plots are of great importance because they immediately show if one of the data sets that is analyzed has a greater variance than the other. If the points form a line that is flatter, the distribution plotted on the x-axis has a greater variance as compared to the distribution plotted on

the y-axis. However, if the points form a steeper line, then the distribution plotted on the y-axis has a greater variance as compared to the distribution plotted on the x-axis.

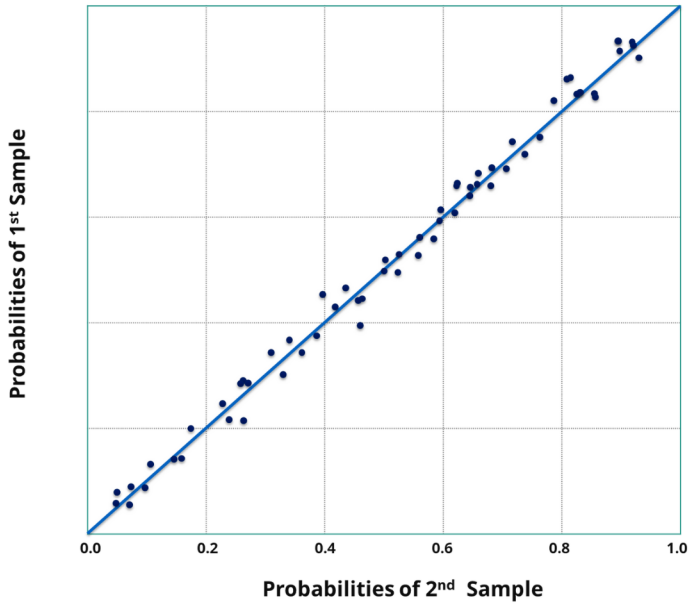

Figure 5.16: Example of a boxplot

Pie Charts

A pie chart (or a circle chart) is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In a pie-chart, the slices of the circle or pie are typically indicated with different colors. The slices are labeled and the numbers corresponding to each slice is also represented in the chart.

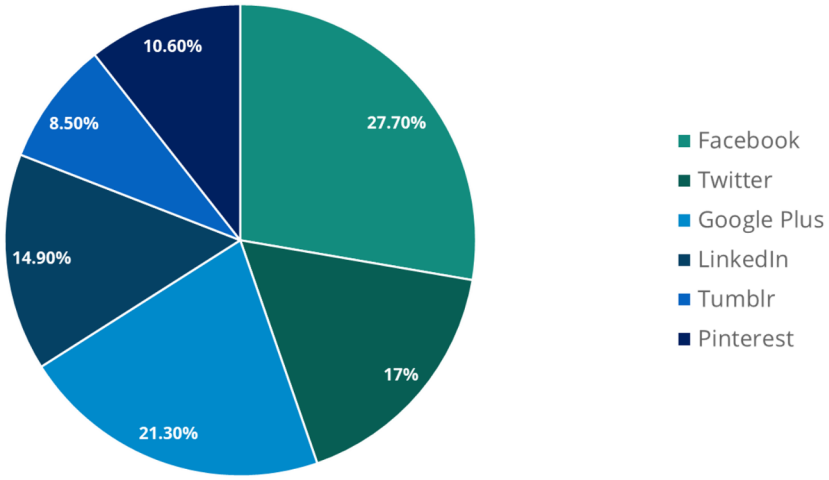

Figure 5.17: Example of a pie chart

In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented.

Radar Charts

A radar chart is a graphical method of displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point. The relative position and angle of the axes is typically uninformative. The radar chart is also known as spider chart due to the nature of its design.

Radar charts are a useful way to display multivariate observations with an arbitrary number of variables[49]. Each point represents a single observation. Typically, radar charts are generated in a multi-plot format with many stars on each page and each star representing one observation.

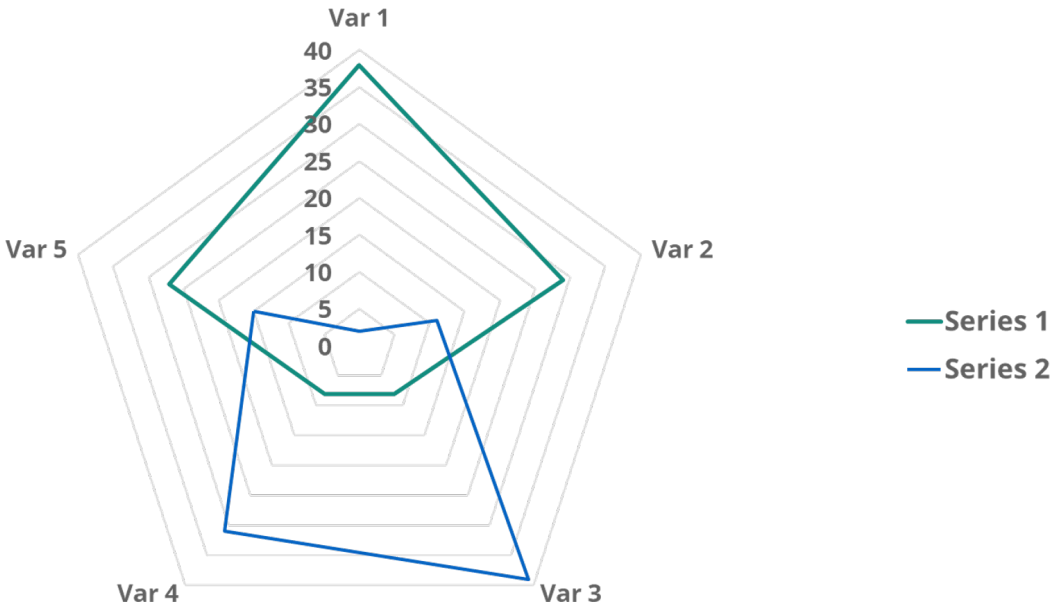

Figure 5.18: Example of a radar chart

Notes

42 Kurzweil, R., Richter, R., Kurzweil, R. and Schneider, M.L., 1990. The age of intelligent machines (Vol. 579). Cambridge: MIT press.

43 Turing, A.M., 2009. Computing machinery and intelligence. In Parsing the Turing Test (pp. 23-65). Springer, Dordrecht.

44 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice Hall.

45 Chui, M., 2017. Artificial intelligence the next digital frontier?. McKinsey and Company Global Institute, p.47.

46 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice hall.

47 Baral, C. and De Giacomo, G., 2015, January. Knowledge Representation and Reasoning: What's Hot. In AAAI (pp. 4316-4317).

48 Aron, A. and Aron, E.N., 1994. Statistics for psychology. Prentice-Hall, Inc.

49 Chambers, J.M., 2018. Graphical methods for data analysis. CRC Press.

6. Big Data Processes

6.1 Introduction to Big Data Processes

Executing Big Data projects that bring value to enterprises is difficult. Because of the volume, variety and velocity of available data sources, organizations can get lost and might see the forest for the trees. Combined with management agenda's, pressure to deliver results and complex algorithms, many organizations struggle to achieve the required return on investment from Big Data initiatives[50].

In order to avoid the potential pitfalls that Big Data brings, processes can help enterprises to focus their direction. Processes bring structure, measurable steps and can be effectively managed on a day-to-day basis. Additionally, processes embed Big Data expertise within the organization by following similar procedures and steps, embedding it as 'a practice' of the organization. Analysis becomes less dependent on individuals and therefore greatly enhances the chances of capturing value in the long term.

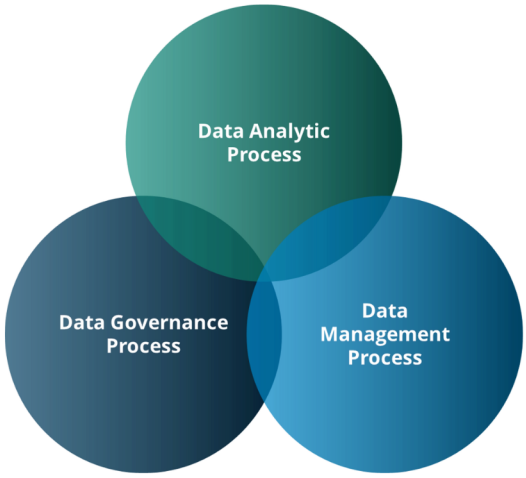

Figure 6.1: The three foundational Big Data processes

Setting up Big Data processes in the enterprise might be a time-consuming task at first, but definitely provides the benefits in the long run. In this section, we will discuss how Big Data processes can provide structure in the analysis of data. Big Data processes can be subdivided into three main sub-processes:

- Data analysis process (focus on control);
- Data governance process (focus on compliance);
- Data management process (focus on quality).

Although closely related and beneficial in any organization, each of the sub processes has a different focus and function: **control**, **compliance or quality**.

6.2 Data Analysis Process

The data analysis process contains sequential steps that the enterprises take in order to process Big Data. Ideally, the same process is used for every Big Data project to ensure consistency between projects and improve performance and efficiency across the enterprises.

As with any process, the data analysis process is sequential and has a clearly identified start (the trigger) and end result (the outcome). By managing the stages in the data analysis process, enterprises can better control the outcomes and results of their Big Data projects.

Figure 6.2: The eight steps of the Big Data analysis process

Step 1 – Determine the Business Objective

The first step of the data analysis process must happen before there is data — the goals and business objectives of the Big Data project need to be determined. Because of the sheer volume of data sets in the world, it is possible to lose focus quickly. The first step is therefore crucial in determining the scope of the project.

Within Big Data projects, the business objectives (and hence it's underlying problem) can frequently be subdivided into six types of problems.[51] Each of these types has its own way of dealing with the outcome of the problem and the way in which the final results need to be interpreted:

1. Descriptive business objective
2. Exploratory business objective
3. Inferential business objective
4. Predictive business objective
5. Causal business objective
6. Mechanistic business objective

A **descriptive business objective** aims to summarize the characteristics of different datasets, originating from either inside or outside the enterprise. The business objective is to collect and summarize data in order to make decisions. Examples include global sales figures at all gas stations for a specific company (internal), or calculating the market share for a company in a specific region.

An **exploratory business objective** aims to find a relation between two or more different variables data sets. The goal of this objective is to find a pattern or relationship in the data that can be used to optimize performance. Examples could include the identification of products that are bought together (market basket analysis) or the identification of a sales pattern based on the weather conditions.

An **inferential business objective** aims to find characteristics about a population by studying a sample of the data, as discussed in Chapter 5.3. Inferential business objectives are prevalent in targeting (potential) customers in marketing and sales organizations within the enterprise. Examples include finding new customers based on existing customer lists or determining in which regions to advertise based on previous purchase behaviors.

A **predictive business objective** aims to predict future behaviors by analyzing and extrapolating data sets, such as predicting which products customers are likely to buy (market basket analysis). With predictive objectives, the outcome is uncertain and Big Data is used in order to find the best possible answer. Examples include the determination of regions or properties for future investments.

A **causal business objective** aims to find the underlying relationship of a certain phenomenon (the cause). This type of objective aims to find the root cause of certain data patterns in order to better understand relationships. A causal business objective aims to learn why certain data were created. Examples include finding out why sales performance was higher in a certain month or what the root cause is of increased quality defects.

A **mechanistic business objective** aims to find how variables influence outcomes of data sets. It requires a deeper understanding of the underlying relationships and patterns within data sets. Examples include understanding how sales performance is influenced by the weather conditions or how the combinations of ingredients can increase revenues.

The first step of big data analysis — determining the business objective — is important because it specifies which algorithms and techniques (discussed in Chapter 5) should be used to solve the problem. A complete mapping of business objectives and algorithms is discussed in the Enterprise Big Data Analyst guide.

Step 2 – Data Identification

The second step in the data analysis process is to determine which sets of data need to be processed. Frequently, this is one of the most important and difficult steps. How to determine the data sets necessary to analyze the problem and provide an answer to the business objective?

Most data analysis starts with identification of the raw data. Raw data is data that has not been processed yet and that is coming directly from the source. Data sources can include:

- Binary files from measurement devices (sensors);
- CSV files that reside on a public website;
- Unformatted Excel sheet with multiple tabs of data;
- JSON data that is obtained from a Twitter API.

In order to identify the necessary data to meet the business objectives, a data identification graph can be drawn, that works backward from the processed data towards the raw (source) data. A data identification graph is depicted in figure 6.3.

State	City	Hospital	#Doctors	Success Rate	Legal Suits	Twitter Rank
AL	Centreville	Bibb Medical Center	37	94%	2	87%
AK	Anchorage	Alaska Regional Hospital	22	88%	15	67%
AZ	Phoenix	Mayo Clinic	65	99%	4	86%

Figure 6.3: Example of a data indentification graph

A data identification graph first identifies the desired (processed) data and then works backwards to identify where the raw data might be obtained. In the example above, the objective is to identify the best hospital per state, by combining information about doctors, malpractice lawsuits and Twitter chatter. The data identification graph above shows where the processed data originates from.

Step 3 – Data collection and sourcing

After it has been identified which data is necessary to achieve the required business outcome, the next step is to ensure that data is obtained for processing. Although this sounds as a relatively straightforward step, in practice it is frequently a difficult step.

Data collection

Within most enterprises, (internal) data is stored at various physical locations or data centers across the globe. In order to make use of this data, the data analyst or data scientist should obtain the appropriate access rights and collaborate with the data management team (see next section). Measures should be set up to ensure data integrity and data confidentiality, safeguarding the data will not fall into the wrong hands. Additionally, most countries have privacy and regulatory requirements that personal data may not cross the border. This could cause significant issues if Big Data teams for example aim to compare customer behaviors from Singapore (located in a data center in Singapore) with customer behavior from the USA (located in a data center in Boston).

Besides the security and privacy concerns, a second consideration in the data collection process is to deal with the volume, variety, and especially velocity aspects of the data sets. If data is renewed or refreshed on a daily basis (for example a Twitter feed), collection includes decisions regarding the frequency of data imports (real-time vs. batch imports) and how to deal with previous imports.

Data sourcing

To obtain value from Big Data, internal enterprise data sets are combined with external data sets (for example weather information or Twitter feeds). Some of these external data sets might be available for free, but most data sets will need to be acquired from external vendors.

Data acquisition and sourcing also requires the involvement from procurement and — according to sourcing procedures and regulations in most countries — an open and transparent bidding process. Besides the fact that these procedures take time, great care should be taken to ensure that the data vendor is not overselling the value of their data products, whilst major data processing costs are hidden in the data sourcing process.

Step 4 – Data review

After the required data sets have become available, the data review step starts. Data review is the process of exploring your data sets and typically includes examining the structure and variables of the various datasets. In this process, it is determined whether the data sets have been corrupted, whether there are missing values or if there are multiple (conflicting) sets of the same variables. It might, for example, be the case that sales data for a specific region from two different finance systems have different values.

Major objectives of the data review process include:

- To determine whether there are any problems or issues with the data sets;
- To determine the variables and distribution of data in the data sets;
- To determine if the data set contains missing values or corrupt data;
- To determine whether the business objectives (step 1) can be realized with these data sets.

The inclusion of missing, incorrect or outlier data in data sets can have a significant impact on the final results, as indicated by the example below:

```
# Import the statistics library
import statistics

# Consider the following three lists with data:
playerAge1 = [19, 19, 20, 24, 27, 28, 31, 34]
playerAge2 = [19, 19, 20, 24, 27, 28, 0, 0]
playerAge3 = [19, 19, 20, 24, 27, 28, NA, NA]

# The mean of each list will give the following results:
print(statistics.mean(playerAge1))    # 25.5
print(statistics.mean(playerAge2))    # 17.125
print(statistics.mean(playerAge3))    # Error
```

The reality in practice is that almost every data set, even if procured from expensive and trustworthy data providers, has incorrect or missing values that need to be accounted for. The data review process is therefore of paramount importance, because missing values of outliers can have a profound impact on the end result. In case there is flawed, incorrect or corrupt data, the data set needs to be cleansed in the next step.

Step 5 – Data Cleansing

Data cleansing is the process of amending or removing data in a database that is incorrect, incomplete, improperly formatted or duplicated. Data cleansing may be performed interactively through data cleansing tools, or as batch processing through scripting.[52] After cleansing, a data set should be consistent with other similar data sets in the system and ready to be used for data processing.

Enterprises in a data-intensive field like banking, insurance, retailing, telecommunications, or transportation might use data cleansing tools to systematically examine data for flaws by using rules, algorithms, and look-up tables. Typically, a data cleansing tool includes programs that are capable of correcting a number of specific type of mistakes, such as adding missing zip codes or finding duplicate records.

Step 6 – Model building

The next step in the data analysis process is the generation of a statistical model that can be used in finding the result to the business objective. Model building is the iterative process of defining and improving a statistical model that can be applied to the (cleansed) data set. In Chapter 5, we provided an introduction to commonly used models and algorithms that can be applied in this stage of the data analysis process.

Mathematical formulas or models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation. In general terms, models may be

developed to evaluate a particular variable in the data based on other variable in the data, with some residual error depending on model accuracy (i.e., Data = Model + Error).[53]

In the domain of politics, for example, a data analyst can use a sample of polls in order to predict the outcome of an election. In order to do this, the analyst would need to build a model that can be applied to the data. The process of building a model involves imposing a specific structure on the data and creating a summary of the data. The (statistical) model is one of the most valuable steps in the data analysis process, because the accuracy of the model determines the end result.

Step 7 – Data processing

The data processing step is dedicated to carrying out the actual analysis task, which typically involves running one or more (statistical) algorithms. This step can be iterative, especially if the data analysis is exploratory so that analysis is repeated until the appropriate pattern or correlation is uncovered.

Depending on the type of process required, the data processing step can be as simple as querying a dataset to averages, modes or median. On the other hand, it can be as complex as combining multiple complex algorithms to run algorithms for facial recognition, DNA sequencing, or financial market predictions. The duration of the data processing stage varies depending on the requirements,

As discussed in Chapter 4, most Big Data solutions will use some form of distributed processing (most commonly the Hadoop software framework) to reduce the necessary processing time.

Step 8 – Communicating the Results

The Big Data analysis process ends with communicating the end results. Although this is the logical last step of any analysis project, its importance cannot be underestimated. Communicating clearly is essential to sound data analysis.

Since Big Data and its underlying processing algorithms are sometimes difficult to explain to business leaders, good communication is essential to the success and value of Big Data. Communicating on a regular basis (e.g., interim reports) and in a structured fashion (every Friday) will provide teams and decision makers in enterprise organizations the required trust that structured procedures are followed.

One of the best ways to communicate the results of any Big Data project is to use the data visualization techniques that were discussed in Chapter 5.9. By summarizing data into graphs and figures, it becomes easier to understand. Data visualization technologies can be as powerful as they are easy to use, allowing data analysts to quickly and easily articulate and share the insights across the enterprise to others who are less comfortable with the nuances of data analysis.

6.3 Data Governance Process

The data governance process is a defined process that enterprises follow to ensure that they control their data throughout the complete lifecycle. Since “Big Data” is a strategic asset, most organizations

need to establish measures of control. The data governance process ensures that important data assets are formally managed throughout the enterprise, and the data can be trusted for decision-making. Often, the processes used in data governance include accountability for any adverse event that results from data quality.[54]

The focus on the data governance and its accompanying process has grown greatly over the last few years, especially due to increased data privacy- and data confidentiality requirements that have been set by the countries. The data governance process therefore not only needs to set the policies and assign responsibilities across the enterprise, it additionally needs to ensure that enterprises are compliant to (local) data laws and regulations.

There is a close relationship between the data governance process and the data management process, which will be discussed in the next section. Where the data governance process is concerned with setting the policies and responsibilities on a strategic level, the data management process executes and monitors these policies on an operational level. The synergy between the two processes is depicted in figure 6.4.

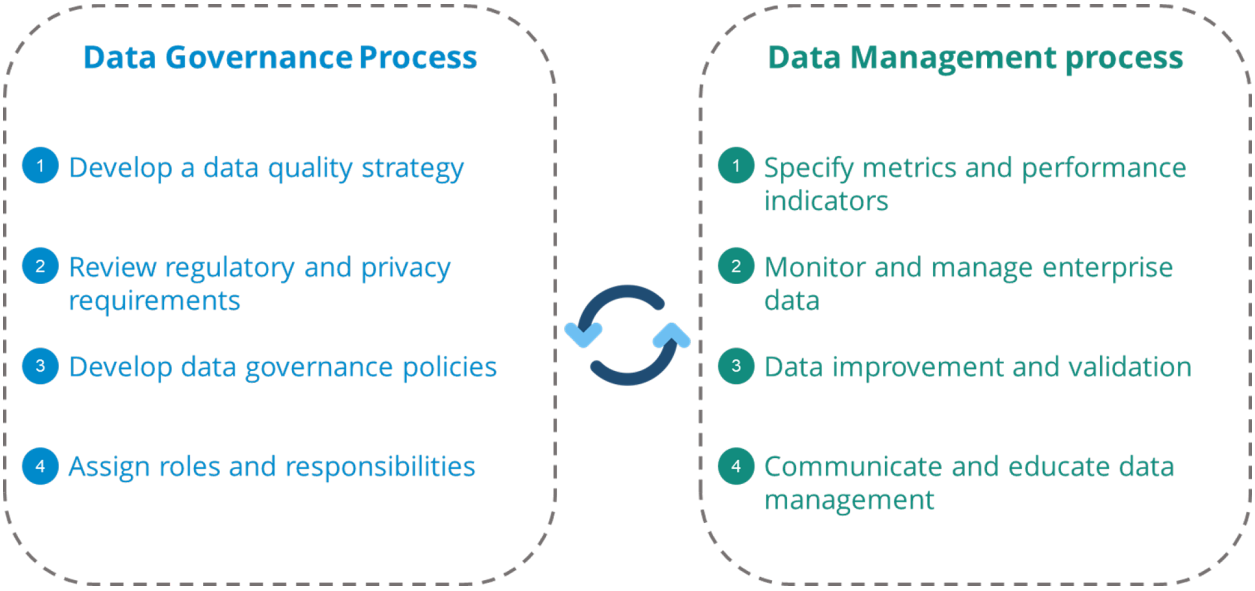

Figure 6.4: The synergy between the data governance and data management processes

Data Governance Process Activities

Data governance process encompasses the people, organizational structure and technology required to create a consistent and proper handling of an organization's data across the business enterprise. Although the goals may vary based on the nature of the enterprise, the required level of control and (local) regulatory requirements, a number of universal data governance activities are the same for every organization.[55]

Develop a Data Quality Strategy

A data quality strategy is needed to manage and direct all data quality activities in line with

the overall business strategy. The data quality strategy includes the strategic objectives which are pursued by data management process (discussed in Chapter 6.4), how it is aligned with the enterprise's strategic business goals and its overall functional scope. Moreover, it makes statements about the involvement of its stakeholders which means analyzing and comprehending the role of data within the enterprise. The data quality strategy should be reviewed and updated on a yearly basis at a minimum.

Review Regulatory and Privacy Requirements
Some countries (for example the U.S.A. or Singapore) have strict laws about the data privacy of their citizens and regulatory requirements for data transfer across borders. In most cases, these countries require (auditable) reports and logs to ensure organizations are compliant to these laws and regulations. As regulatory and fines can be steep (and reputations are damaged as a direct result), there is a strong imperative for enterprises to ensure they are compliant to these rules. The review of regulatory and privacy requirements is therefore an integral part of the data governance process that should be conducted on monthly basis to ensure compliance.

Develop Data Governance Policies
The data quality strategy and the regulatory requirements need to be translated into a number of data governance policies that are publicly available for everyone in the organization. These policy documents contain the decisions the enterprise has taken with regard to their data quality organization and explain the requirements for everyone. These policies are a useful input for sourcing decisions, external contracts with suppliers and can provide valuable input for other business processes.

Assign Roles and Responsibilities
The data governance process must define clear roles and responsibilities across divisional boundaries in the enterprise. The role assignment has to ensure accountability, authority and supervision as well as the involvement of senior executives and business management and encourage desirable behavior in the use of data.

6.4 Data Management Process

The data management process is a separate process that safeguards the quality of the data on a day-to-day operational level. This process executes upon the strategic directives of the data governance process (discussed in the previous section).

The primary objective of the data management process is to ensure data quality. The value that can be obtained by analyzing Big Data is highly dependent on the quality of the input data. Even with the most sophisticated Big Data solution the general 'Garbage-In-Garbage-Out' rule still applies. If data sets are corrupt or erroneous, data analysis might result in invalid results or conclusions.

Enterprises therefore need the data management process to continually verify, update and clean the enterprise data. The data management process outlined in this chapter provides a structured and practical approach to implement the following ideas:

- Enterprises need a way to formalize their expectations for measuring the conformance of data quality to these expectations;
- Enterprises must be able to baseline the level of data quality in order to identify problems and analyze root causes of data failure;
- Enterprises need to be able to communicate their level of confidence in the quality of their data.

Data Management Process Activities

The data management process is a practical and operational process (in line with the strategic directives of the data governance process) that monitors the data quality on daily basis. The process consists of the following activities:

Specify Metrics and Performance Indicators
To measure and assure data quality throughout the whole data life cycle, companies specify metrics and performance indicators based on the data quality dimensions that fit the enterprise's information needs. These metrics ought to be linked to the company's general goals and objectives as determined in the data governance policies.

Metric	Description
Uniqueness	Uniqueness refers to requirements that data within the enterprise is captured and represented uniquely within the relevant application architectures. Asserting uniqueness of the entities within a data set implies that no entity exists more than once within the data set.
Accuracy	Accuracy is the extent to which the data that is reflected in the data set corresponds with the truth. It refers to whether the data values stored for an object are the correct values. To be correct, a data value must be the right value and must be represented in a consistent and unambiguous form.
Consistency	In its most basic form, consistency refers to data values in one data set being consistent with values in another data set. A strict definition fo consistency specifies that two values drawn from seperate data sets must not conflict with each other, although consistency does not necessarily imply correctness.
Completeness	An expectation of completeness indicates that certain attributes shoudl be assigned values in a data set. Completeness rules can be assigned to data to validate that all attributes are present in the data set.
Timeliness	Timeliness refers to the time expectation for accessibility of information. Timeliness can be measured as the time between when information is expected, and when it is readily available for use.
Currency	Currency refers to the degree to which information is current with the world that it models. Currency can measure how "up-to-date" information is, and wheterh it is correct despisble possible time-related changes

Metric	Description
Conformance	This dimension refers to whether instances of data are either stored, exchanged or presented in a foramat that is consisten with the domain of values, as well as consisten with other similar attribute values.
Integrity	Data integrity is the maintenance and assurance of teh accuracy and consistency of data over its entire life-cycle and is a critical aspect to the design, implementation and useage of any system which stores, processes or retrieves data. Integrity means that the data has not been altered.

The metrics and performance indicators can be captured in a balanced data quality scorecard. The creation of such a scorecard provides an efficient means to continuously monitor and manage data based on key performance indicators. Common metrics and performance indicators that are included in the data quality scorecards are depicted in the table above.

Monitor and Manage Enterprise Data

Based on the metrics and performance indicators that have been specified in the previous activity, the enterprise data needs to be monitored. With (automated) tools, data sets can be monitored and indexed to measure the quality of the data against the specified performance indicators. The results can again be depicted in data quality scorecards.

One of the important elements of this process activity is the generation (and subsequent follow up) of alerts. Alerts need to be generated if it is detected that data has been corrupted or changed.

Data Improvement and Validation

The next activity of the data management process is to improve the enterprise data sets. The balanced data score card from the previous activity might, for example, indicate that there are many duplicate records in data sets. The data improvement and validation activity concerns itself with ‘cleaning’ up the datasets in order to improve the metrics and performance indicators.

Using validation rules and transformation rules, the quality of data can be improved as depicted in figure 6.5.

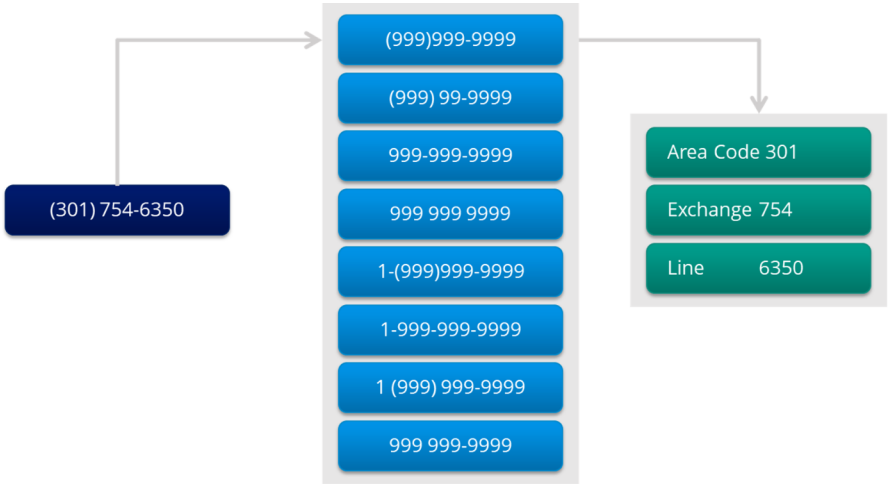

Figure 6.5: Validation rules can be used to improve data quality

Communicate and Educate on Data Management

The last activity of the data management process is to communicate and educate enterprise stakeholders to actively participate in data management initiatives. By ensuring that the data governance procedures are followed and systems are used correctly, the data quality of the enterprise can be significantly improved. In many cases, employees are unaware of data structures and are not aware of the value of data for the organization.

In order to improve this knowledge, training programs can reduce user errors, increase productivity and increase compliance with key controls. Education addresses core data principles and data quality practices complemented by role-specific training. In particular, data collectors have to understand why and how consumers use data.

Notes

50 Kurzweil, R., Richter, R., Kurzweil, R. and Schneider, M.L., 1990. The age of intelligent machines (Vol. 579). Cambridge: MIT press.

51 Turing, A.M., 2009. Computing machinery and intelligence. In Parsing the Turing Test (pp. 23-65). Springer, Dordrecht.

52 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice Hall.

53 Chui, M., 2017. Artificial intelligence the next digital frontier?. McKinsey and Company Global Institute, p.47.

54 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice hall.

55 Baral, C. and De Giacomo, G., 2015, January. Knowledge Representation and Reasoning: What's Hot. In AAAI (pp. 4316-4317).

7. Big Data Functions

7.1 Introduction to Big Data Functions

This chapter discusses the organizational aspects of setting up a Big Data practice in enterprises. Embedding a Big Data practice is more than sourcing some tools, finding data sets and hiring people with the right skills. In order to obtain long lasting value out of (significant) Big Data investments, the organizational aspect is at least as important.

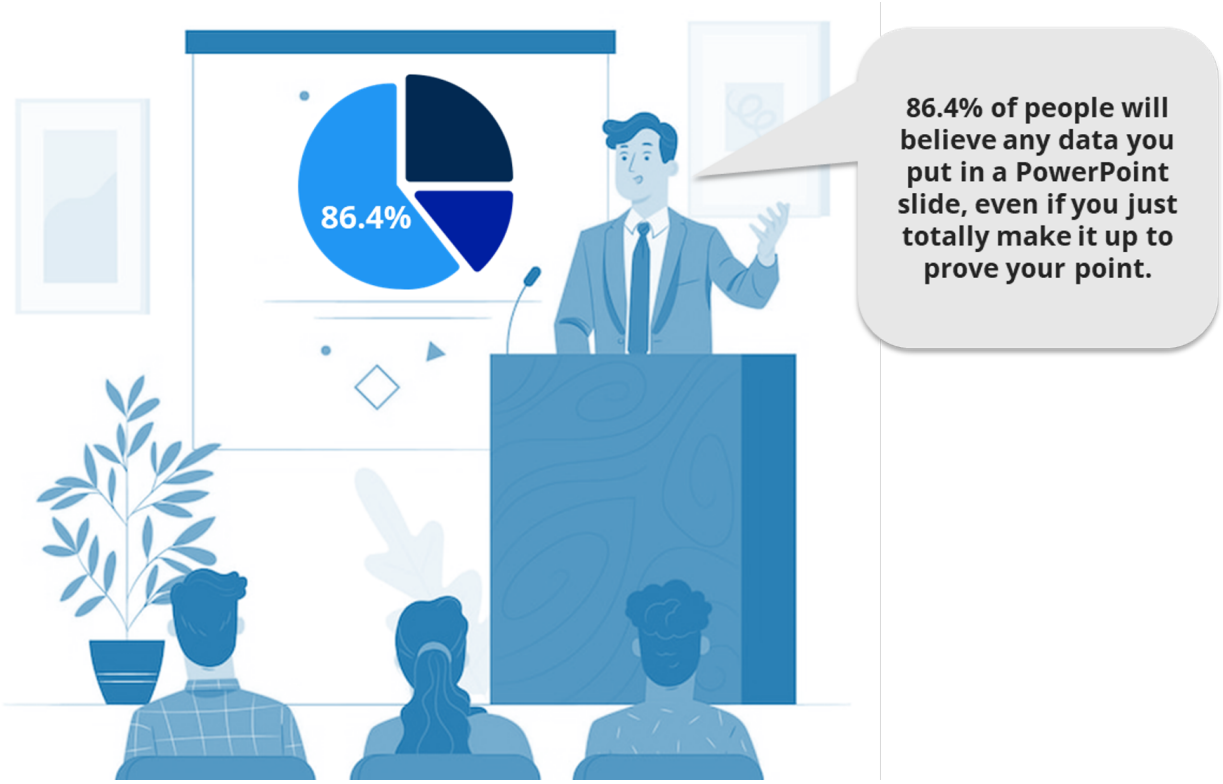

Figure 7.1: Becoming an data-driven organization requires change management

Old ways of working are deeply ingrained, especially if there is an underlying distrust of Big Data and analytics. Setting up a Big Data organization is therefore just as much change management, as it is sourcing the right skills, processes and technology. The benefits of Big Data can only be reaped if the hearts and minds of the people in the organization are aligned with the Big Data strategy. Not only will people need to start working in a different way, they will also need to make decisions differently. Insights that are deduced with Big Data analysis should be integrated in the daily decision-making process in order to become a data-driven enterprise (as discussed in Chapter 2).

Embedding Big Data in Enterprises is as much about change management as it is about Big Data. People buy into change when they understand it and feel part of it. The design of a Big Data organization (digital transformation) therefore needs to be user led and have participation from all levels in the enterprise from the start. It is essential to have a number of champions that truly understand the value of Big Data and can help with the development of Use Cases and the formulation of a Big Data strategy.

Organizational culture, organizational structures, and job roles have a large impact on the success of Big Data initiatives. In this chapter, we will therefore review some 'best practices' on how to establish a data-driven organization.

7.2 Designing a Big Data Organization

In most enterprises, Big Data projects get started when an executive becomes convinced that the company is missing out on opportunities in data.[56] For example, the CIO might have heard about the benefits a competitor has achieved in a Big Data project, and is eager to obtain similar results. But where to start?

In many organizations, there is a (large) gap between the first launch of a Big Data project (initiated by an enthusiastic sponsor) and scaling-up the benefits of a Big Data project across the enterprise. In order to obtain long-term value from big data and become a truly 'data driven' organization, it is crucial to set up a Big Data Centre of Excellence.

Big Data Centre of Excellence

A Big Data Centre of Excellence (BDCoE) is an enterprise function that takes an organization from zero knowledge to having a fully functional practice of Big Data technologies and processes to deliver robust business results. A BDCoE is where the organization identifies new technologies, learns about new skills and develops appropriate processes that are subsequently deployed throughout the other business units of the organization.[57]

A BDCoE is essential to accelerating Big Data adoption by the enterprise in a fast and structured manner. It reduces the implementation times drastically and therefore the time-to-market to deploy new data-driven products and services. More importantly, it ensures that the best practices and methodologies are shared through different teams in the organization. A BDCoE should be a live and evolving organizational function that expands and grows as the organization's needs evolve.

A centralized BDCoE can be the foundation for establishing a data driven enterprise that values data as its strategic asset. The BDCoE can partner with the business to identify which projects should be prioritized and what data is of strategic importance. As such, it operates as the strategic counterpart of the business to translate current and actual business requirement into live and actionable Big Data projects. Big Data's strategic importance is the value it represents for the business, but success with Big Data is not just about data. The people and the organization also play a vital role in that success.

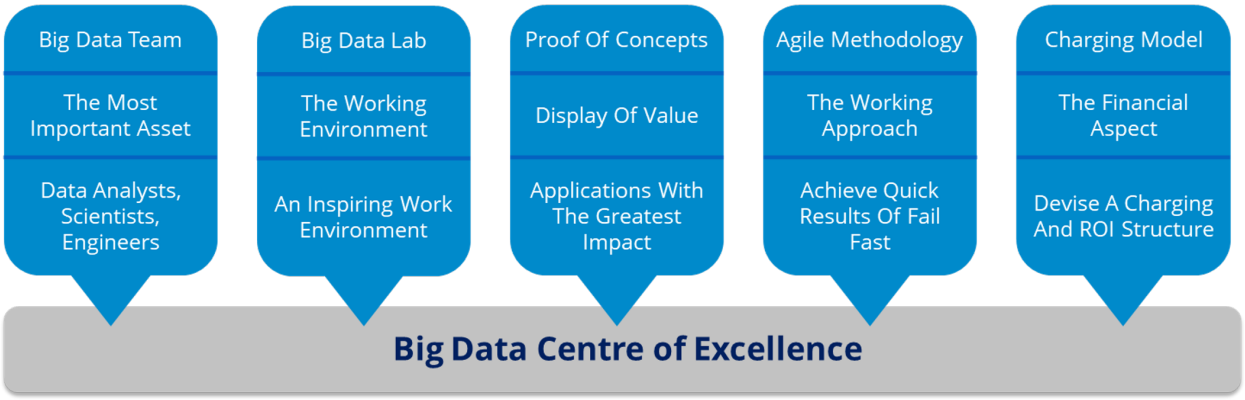

Figure 7.2: Structure of a Big Data Center of Excellence

An effective BDCoE consists of five major pillars that together form the structure for obtaining value from the centralized function.

Big Data Teams

The most important element, the quality of the Big Data analysts, Big Data scientists and Big Data engineers is paramount for creating success with Big Data. In the end, Big Data is knowledge domain, and that knowledge will come from the people. The Big Data professionals needs to be certified and experienced practitioners that have a track record of working with data.

Since the topic of Big Data team is so important of achieving success, a more detailed description of Big Data roles and responsibilities is discussed in Chapter 7.3. Chapter 7.4 discusses the required skills of Big Data professionals.

Big Data Labs

Big Data labs refer to the working environment of the BDCoE. The obvious link between the data 'science' and a lab is on purpose since the environment should be a creative space to experiment and run test data analyses in order to achieve the desired results.

A well-designed Big Data lab contains open work-spaces that allow for communication and collaboration as well as isolated work possibilities where data analysts can 'crunch the number' without distractions.

A second important requirement for the Big Data labs is to have the hardware compatible for Big Data processing. In general, Big Data labs require hardware with sufficiently larger RAM than usual for Big Data processing.

Big Data Proof-of-Concepts

Proof-of-Concepts (POC) are showcase solutions that can be provided to internal business units as well as external clients. The POCs should demonstrate a clear return on investment and clearly showcase the capabilities of the BDCoE in achieving the results.

Proof-of-Concepts are usually requested by internal business units or customers based upon specific Big Data questions (discussed in Chapter 6.2, the first step of the Bid Data analysis process). By demonstrating clear Proof-of-Concepts for potential use cases, the BDCoE can showcase its knowledge in the enterprise.

Agile Methodology

Agility and the ability to fail fast or achieve quick results are essential to reaching the potential of Big Data. An Agile working methodology provides the tools to deliver outcomes quickly and transparently, typically within two- to three-week sprints. The ability to fail fast is a key Big Data opportunity — business and technical roadmaps for delivering value need to change more often than in a traditional waterfall environment.

Charging Models

At the core of the BDCoE are the charging models to justify the (sometimes large) investments in the people, processes and technology of the Centre. In order to display value, a clear approach needs to be devised to charge other business units or external clients for services rendered.

Charging models can be devised based on the number or users, data processed, frequency of reports or subscription based. A sound and unambiguous charging model will greatly help to showcase the value of BDCoE to the enterprise.

7.3 Roles and Responsibilities in Big Data Teams

The most important aspect of Big Data are the people involved. While many organizations have plans to turn their data into value, they sometimes spend too much time on the ‘data’ and not enough time on the ‘people’ side of the equation. In the short period of time that data science is now part of professional enterprises, a number of new roles have formed that are essential to the success of Big Data. Each of these roles contributes to the Big Data team of the BDCoE that was explained in the previous section.

Big Data Analyst

The Big Data analyst is a role that involves acquiring, processing and summarizing the information from Big Data sets in order to discover business value. Unlike data scientists, data analysts are more generalists.

Big Data analysts are expected to know R, Python, HTML, SQL, C++, and Javascript. They need to be more than a little familiar with data retrieval and storing systems, data visualization and data warehousing using ETL tools, Hadoop-based analytics, and Business Intelligence concepts. These

persistent and passionate data miners usually have a strong background in math, statistics, machine learning, and programming.

Big Data analysts are involved in data crunching and data visualization. If there are requests for data insights from stakeholders, data analysts have to query databases. They are in charge of data that is scraped, assuring the quality and managing it. They have to interpret data and effectively communicate the findings.

Big Data Scientist

The Big Data scientist is a role that involves the development and deployment of algorithms and statistical models in order to predict future outcomes that provide business value based on Big Data sets. In recent years, the data scientist role has grown tremendously in popularity and there is significant demand for this job role.

The Big Data scientist job role is a senior role that requires deep understanding of algorithms and data processing operations. People in this role are expected to be experts in R, SAS, Python, SQL, MatLab, Hive, Pig, and Spark. Data scientists typically hold higher degrees in quantitative subjects such as statistics and mathematics and are proficient in Big Data technologies and analytical tools.

The role of a data scientist is not only about data crunching. It's about understanding business challenges, creating some valuable actionable insights to the data, and communicating their findings to the business. Additionally, the role of the data scientist requires creative thinking and problem solving skills that are necessary to design, develop, and deploy algorithms that can retrieve value from Big Data.

Big Data Engineer

The Big Data engineer is a role that designs, builds and manages the underlying IT infrastructure that is required to obtain value from Big Data sets. Data engineers ensure that an enterprise's Big Data ecosystem is running without glitches for data analysts and data scientists to carry out the analysis.

Big Data engineers are computer engineers who must know Pig, Hadoop, MapReduce, Hive, MySQL, Cassandra, MongoDB, NoSQL, SQL, Data streaming, and programming. Data engineers have to be proficient in R, Python, Ruby, C++, Perl, Java, SAS, SPSS, and Matlab. Other must-have skills include knowledge of ETL tools, data APIs, data modeling, and data warehousing solutions. They are typically not expected to know analytics or machine learning.

Big Data engineers develop, construct, test, and maintain highly scalable data management systems. Unlike data scientists who seek an exploratory and iterative path to arrive at a solution, data engineers look for the linear path. Data engineers will improve existing systems by integrating newer data management technologies. They will develop custom analytics applications and software components. Data engineers collect and store data, do real-time or batch processing, and serve it for analysis to data scientists via an API.

The Big Data engineer is frequently also referred to as a 'Big Data architect'. Since both job roles are very similar in nature (e.g., managing Big Data Infrastructure), this guide will use the term Big Data engineer.

Other Big Data roles

Since the domain of Big Data is rapidly growing, many more Big Data roles exist. Examples include Machine Learning Engineer, MIS Reporting Executive, Big Data solutions specialist, etc. Most of these roles require expertise of a specific Big Data platform or tool. The most essential roles to operate any BDCoE can however be summarized by the three roles that were discussed in the section above.

Big Data Roles in the Big Data Analysis Process

In Chapter 6.2, we discussed the Big Data analysis process and the steps that are involved in this process. In Figure 7.3, the job roles (data analyst, data scientist, data engineer) have been placed in the context of the data analysis process to illustrate the focus of every role in more detail. The Big Data Analyst focuses on the movement and interpretation of data, typically with a focus on the past and present.

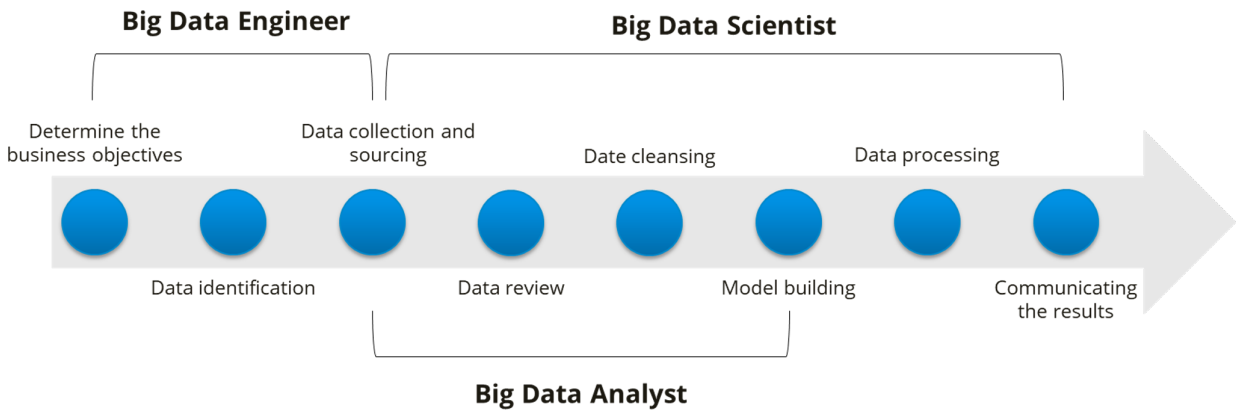

Figure 7.3: Big Data job roles in the data analysis process

Alternatively, the Data Scientist may be primarily responsible for summarizing data in such a way as to provide forecasting, or an insight into future based on the patterns identified from past and current data. The Big Data Engineer, lastly, is more concerned with making sure the underlying Big Data infrastructure is available, before the processing begins.

7.4 Big Data Skills

According to KPMG's 2021 CIO Survey, Big Data analytics is the most in-demand technology skill for the second year running, but nearly 40% of IT leaders say they suffer from shortfalls in skills

in this critical area.[58] What's more, less than 25% of organizations feel that their data and analytics maturity has reached a level where it has actually optimized business outcomes because of a lack of skills.

What are the skills that people who work in Big Data need to have? People working in Big Data are required to have six core skills for success, as indicated in Figure 7.4:[59]

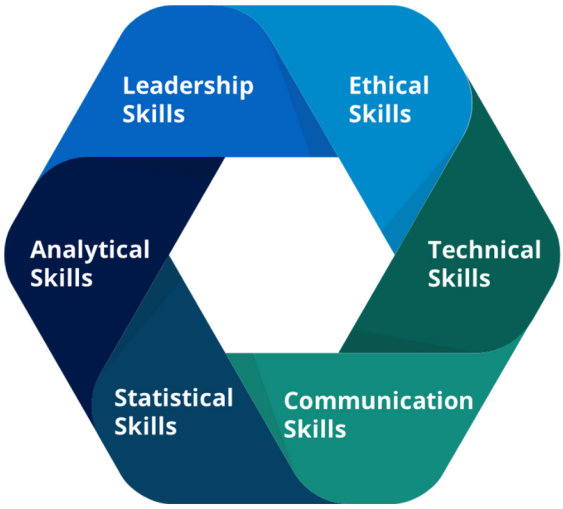

Figure 7.4: Critical skills for Big Data professionals

Leadership Skills

The domain of Big Data is located between the business domain and IT domain and, at times, will be under pressure from both sides. People involved in a Big Data team require strong leadership skills, i.e., the ability to guide or lead other individuals in the organization. Especially when establishing a Big Data organization (when building the BDCoE), many conflicting situations will arise. Strong leadership skills ensure that the individuals in the Big Data keep their focus on the end results in order to prove their value to the business.

Technical Skills

Big Data is deeply rooted in the domain of Information Technology. Without the technology, it would not be possible to achieve valuable results. Individuals involved in Big Data should therefore be strongly interested in technology and should understand the underlying concepts of Big Data solutions and processing technologies. Most of the roles discussed in the previous section outlined key skills and requirements that are requested most throughout the job market. Basic skills in technologies such as R, SAS, Python, SQL, MatLab, Hive, Pig, and Spark are highly recommended to succeed in the domain of Big Data.

Analytical Skills

Without the analysis of data and insights, a business wouldn't be able to function effectively. As a Big Data professional, it is important to have a solid understanding of the business environment and domain the business you work for operates in. The ability to visualize and interpret data is also an essential big data skill that combines both creativity and science. Being able to visualize and analyze data requires a lot of precise hard science and mathematics but it also calls for creativity, imagination, and curiosity.

Statistical Skills

Big Data and the analysis of data in general is always based on the scientific domain of statistics. Data processing and algorithms perform statistical operation on large data sets. Everyone working in the domain of Big Data therefore needs to have a firm understanding of statistical core concepts (such as standard deviations, standardization, etc.) and its underlying operations. An introduction to the required statistical skills for Big Data has been discussed in Chapter 4.

Communication Skills

A requirement to be successful in almost any profession, communication skills are essential for Big Data professionals. Big Data analysis, operations and algorithms are frequently complex and require a deeper subject matter expertise. In order to explain concepts or provide progress updates to business leaders, strong communication skills are required. The ability to translate complex processes and calculations into easy-to-understand summaries and advice is one of the most essential elements of success in Big Data projects.

Ethical Hacking Skills

The analysis of Big Data more often than not will result in problems that do not have a predetermined solution. These could be problems with the data, problems with to processing or simply unexplainable results. In order to overcome these obstacles, Big Data professionals are required to have ethical hacking skills. Individuals with hacking skills keep trying to find different solutions to problems, search for documentation in all different kinds of places or consult peers to seek guidance. They won't stop trying until they have found a way to solve the problem.

7.5 Organizational Success Factors for Big Data

Big Data projects have now been initiated for more than a decade. However, that does not mean that all Big Data initiatives have been highly successful. Various studies show that although investments in Big Data have been increasing in recent years, many organizations still struggle to show their

return on investment due to poor implementations.[60] In order to learn from previous mistakes, a number of success factors have been identified that can give enterprises a head start with launching their big data initiatives:

1. **Establish a vision on how to create value**: The first milestone is to gain a clear view of what your organization is trying to accomplish with Big Data. The fact that your organization captures terabytes of data on a daily basis is meaningless if there is not a clear view with a plan of action as to what the organization wants to accomplish with that data.
2. **To succeed with Big Data, start small**: Building Big Data capabilities take time. A one-time large investment in a Big Data team is not going to produce immediate results. Therefore, a small start with controlled growth is recommended. First, define a few relatively simple Big Data projects that won't take much time or data to run. For example, an online retailer might start by identifying what products each customer viewed so that the company can send a follow-up offer if they don't purchase. A few intuitive examples like this allow the organization to see what the data can do. More importantly, this approach yields results that are easy to test to see what type of returns Big Data provides.
3. **Establish Big Data processes from the start**: Make it clear from the very beginning who is responsible for what. Design effective data governance and data management processes, specifying who is responsible for data definition, creation, verification, curation, and validation—the business, IT, or the BDCoE. Section 6.1 discusses Big Data processes in more detail.
4. **Establish a Big Data Center of Excellence**: A centralized BDCoE provides a uniform point where expertise about Big Data practices and technologies is combined. The BDCoE can partner with the business to identify which projects should be prioritized and what data is of strategic importance. As such, it operates as the strategic counterpart of the business to translate current and actual business requirement into live an actionable Big Data projects. Section 7.2 discussed the BDCoE in more detail.
5. **Assess your readiness for Big Data**: In order to determine where potential gaps and risks might arise, conduct a Big Data readiness assessment. This is the assessment of the readiness of your IT environment and in-house skill sets to implement your organizations' Big Data project and empower members of your existing team as citizen data scientists throughout your organization to put the power of Big Data to work to drive your business forward.
6. **Set up an on-going Big Data training program**: Knowledge and skills are the most important key to success, yet one of the most difficult elements to obtain. Skilled Big Data professionals are not easy to find and even when a team is established, they required continual updates on their knowledge in order to grow further. Setting up an on-going Big Data education program will increase the competency of the organization and embeds a culture of continuous learning

Notes

56 Kurzweil, R., Richter, R., Kurzweil, R. and Schneider, M.L., 1990. The age of intelligent machines (Vol. 579). Cambridge: MIT press.

57 Turing, A.M., 2009. Computing machinery and intelligence. In Parsing the Turing Test (pp. 23-65). Springer, Dordrecht.

58 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice Hall.

59 Chui, M., 2017. Artificial intelligence the next digital frontier?. McKinsey and Company Global Institute, p.47.

60 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice hall.

8. Artificial Intelligence

8.1 Introduction to Artificial Intelligence

In May 1997, the IBM-built DEEP BLUE chess-playing computer famously won the decisive game against Garry Kasparov, the world chess world champion at the time. The event is significant, because it is one of the first moments in history that a human-build machine beat humankind at its own game. Ever since, research and interest in the domain of Artificial Intelligence (AI) has skyrocketed.

Artificial Intelligence is the art of creating machines that perform functions that require intelligence when performed by people.[61] Although there are many different theories about what 'intelligence' actually means, this guide will focus on the operational definition of intelligence by the mathematician and computer scientist Alan Turing. Alan Turing was a computer scientist who developed one of the first theories on Artificial Intelligence. A computer possesses intelligence if a human interrogator is given the task to determine which player – A or B – is a computer and which is a human, but the interrogator is unable to determine the difference. The interrogator is limited to using written questions. This operational definition of intelligence is now famously known as the Turing test.[62]

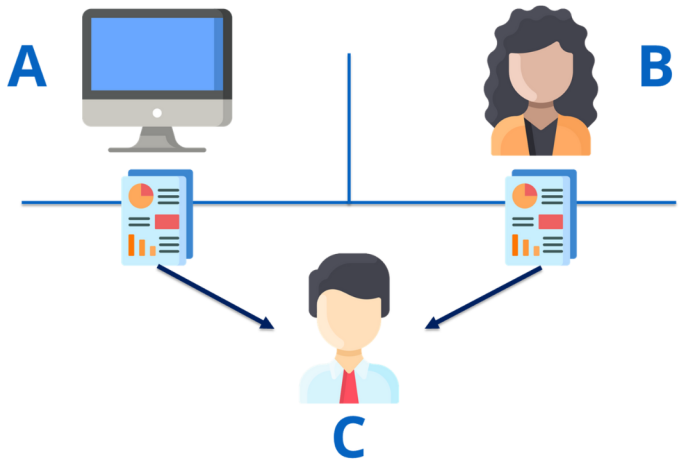

Figure 8.1: Graphical illustration of the Turing test

How is it possible for a computer to pass the Turing test? If we examine closer which underlying techniques a machine should have in order to pass, the computer would require at least the following capabilities:[63]

1. **Natural language processing** — the computer needs to be able to translate English in order to communicate effectively.

2. **Knowledge representation** — the computer needs to store input data and retrieve that same data at a later time.
3. **Automated reasoning** — the computer needs to be able to use the stored information to answer questions and draw conclusions. In order to achieve this, the computer would need to apply an algorithm.
4. **Machine learning** — the computer needs to adapt its response to previous input data in order to formulate new responses.

Each of these four disciplines is integral to the domain of Artificial Intelligence, and it is therefore easy to determine the link between Big Data and Artificial intelligence. The same statistical techniques and algorithms (discussed in chapter 5) that are applied for Big Data analysis are used in the study of Artificial Intelligence.

The main difference between Big Data and Artificial Intelligence is that, where Big Data analysis and analytics mostly stop at predictive and prescriptive analytics, Artificial Intelligence goes one step further. Artificial Intelligence aims to include cognitive science techniques in order to remodel the human brain. However, there is strong overlap between Big Data and AI and both domains continue to improve each other.

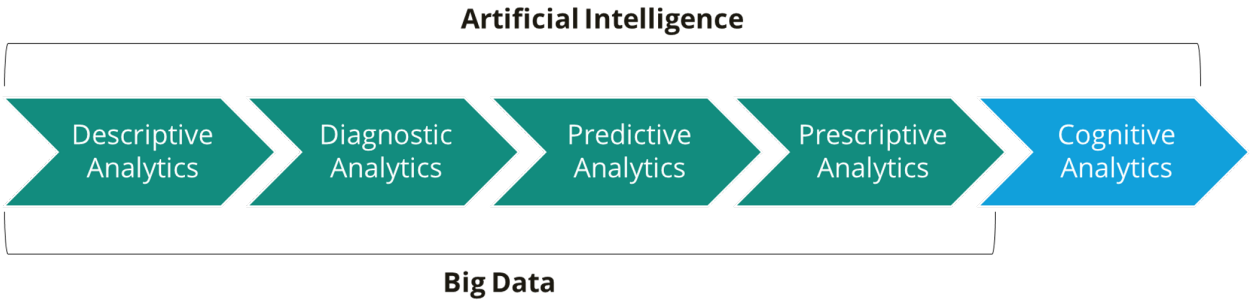

Figure 8.2: Cognitive analytics as an extension to the analytics domain

8.2 Artificial Intelligence in the Enterprise

The study of Artificial Intelligence is complex, highly scientific and requires years of experience. In this guide we will therefore limit the scope of AI to its practical application in enterprises to realize business benefits. The Big Data Framework therefore looks at AI as an extension of Big Data that includes cognitive analytics (as shown in figure 68). The outer ring of the Big Data Framework represents the next practical step an enterprise can take, provided it already has set up an effective Big Data practice.

Investments in Artificial Intelligence is growing fast, predominantly in the tech sector with companies such as Google and Baidu leading the way. The McKinsey Global Institute estimates that

between $20 billion to $30 billion was spent on AI research and development in 2016.[64] However, outside of the tech sector, AI is frequently at early and experimental stages. Most organizations concentrate their main efforts on machine learning — one of the enabling capabilities of AI (further discussed in Chapter 8.3).

So how can business benefit from Artificial Intelligence in a practical way, with clearly defined business objectives and return in investment? In recent years, a number of Use Cases have shown that AI can bring long-term business value.

- Highly autonomous (self-driving) cars are developed at most car productions companies and are forecasted to make up 10 to 15% of global car sales by 2030. The concept of self-driving cars can only be realized with AI technology, because the car needs to make decisions in the same way a human does.
- The use of virtual personal assistants has grown rapidly in recent years and are now included in almost every smartphone. Apple has the famous personal assistant Siri, Microsoft has Cortana and Amazon launched its own virtual assistant Alexa in 2014. Virtual personal assistants need to be able to translate (spoken) speech and translate that into answers, which can only be realized with Artificial Intelligence techniques.
- Call centre solutions based on AI provide real-time input to service desk agents about the human emotions of the persons on the phone. The technology detects whether customers on the other side of the line are happy, angry or afraid and adjust the scripts and advices accordingly. The detection of emotions and translation into the correct solution is a clear application of Artificial Intelligence.
- Smart thermostats in homes can now adjust the temperatures based on the individual characteristics of its users. These thermostats 'learn' the behavioral patterns of the persons that are living in the home and autonomously adjusts it decisions based on this. The underlying machine learning algorithms, together with the automated reasoning and decision-making makes it a powerful application of Artificial Intelligence.

These examples showcase the business opportunities that can be realized by applying Artificial Intelligence are numerous and can cover many different business domains. AI has great potential in almost every industry by facilitating better decision making — almost to the level of human decisions.

8.3 Cognitive Analytics

In order to keep a practical view on Artificial Intelligence in an enterprise context, this guide will work with the operational definition of AI and will focus on cognitive analytics as an extension of the Big Data analytics techniques discussed earlier throughout the guide.

Cognitive analytics is the design and development of algorithms that are able to reflect human decision making, based on the **perceived environment** and **personalized characteristics**. Cognitive analytics differentiates from other forms of analytics because of two main reasons:

1. Cognitive analytics makes decisions based on the perceived environment. The environment can be different at any given time of the day (as is clear from the example of self-driving cars) and needs to be processed based on the specific situation. In order to detect the perceived environment, input data needs to be captured through sensors.
2. Cognitive analytics makes decision based on personalized characteristics. The algorithms learn from its specific user in order to adjust its decision-making to that specific individual. In the example of learning thermostats, the temperature in two different homes will have different heating patterns based on the characteristics of its individual users.

In order to achieve these two key characteristics of Artificial Intelligence, cognitive analytics concerns itself with the development of rational agents. An agent is just something that acts (from the Latin "agere," which means "to do"). A rational agent is one that acts so as to achieve the best outcome or, where there is uncertainty, the best expected outcome.[65] A rational agent therefore tries to mimic the rational decisions that are made by humans. Cognitive analytics therefore concentrates on the design and development of rational agents.

As can be seen in figure 8.3, an agent percepts data from a specific environment (traffic in self driving cars or speech in the case of Siri) through one or more sensors. The agent subsequently processes this data (with some kind of algorithm) and subsequently takes a specific course of action. The decision is autonomous, and similar to the decision a human would take in similar circumstances.

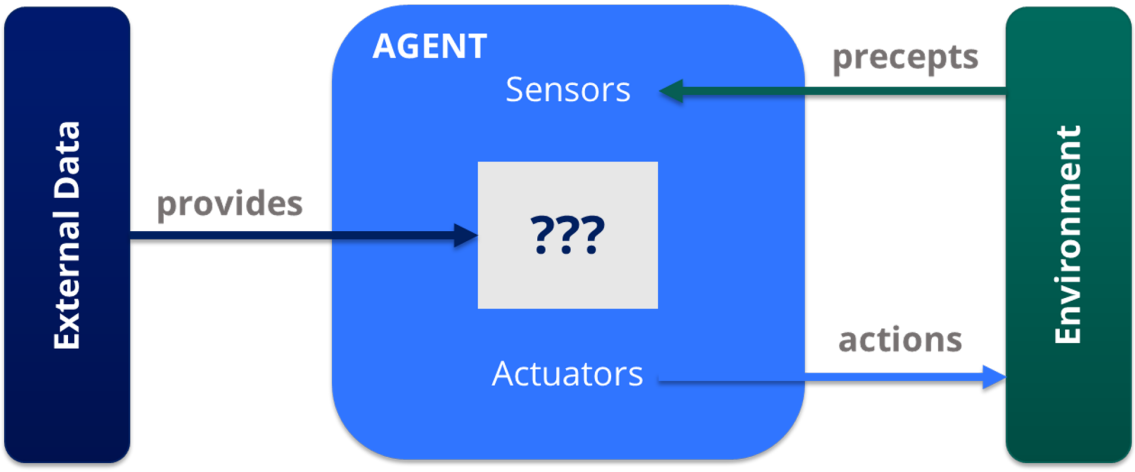

Figure 8.3: High level overview of a rational agent

External data source (Big Data) provides input or reference data to the rational agent in order to make its calculations. Since these external data sources can also be updated real-time (weather forecasts for example), decisions might vary per individual person.

Let's consider the rational intelligent agent for a self-driving car for example. The environment provides thousands of input signals to the rational agent. These input signals could be traffic light colors, distance to the object in front, speed of the car behind, etc. The agent combines this input data with external data from other sources. This external data could the vehicle's driving history (personalized data) or data from external data bases, such as the weather forecast (if rain of snow is

expected, the car speed is reduced). Based on the input data of the sensors and the external data, the rational agent makes a decision. In this example, that means the car increases or decreased speed, brakes, changes gears or changes into a different direction. The decision of the rational agent will have a direct impact on the environment because it needs to ensure everyone can drive safely.

8.4 Capabilities in Artificial Intelligence

Artificial Intelligence combines the capabilities of a great number of capabilities through design and implementation of rational agents. As discussed in Chapter 8.1, a system is deemed intelligent if it has the capabilities to pass the Turing test, which combines four essential capabilities:

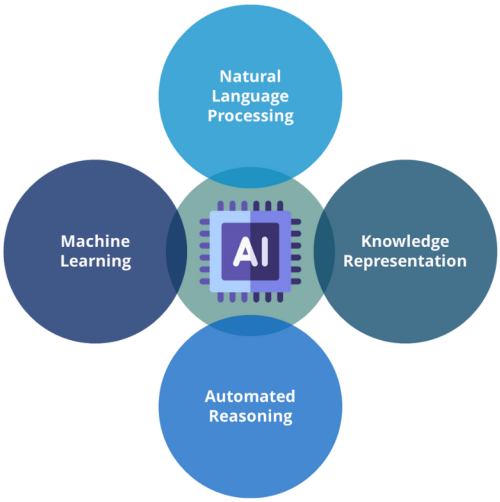

Figure 8.4: The four esential capabilities in Artificial Intelligence

Natural Language Processing

Natural Language Processing (NLP) is the domain that defines the interactions between computers and (natural) human languages, so that people can interact with the computer. Key challenges in natural language processing involve speech recognition, the understanding of meaning in sentences and language or dialect barriers.

Most Artificial Intelligence solutions will need to use some form of NLP in order to facilitate the transfer of data from the environment to the rational agent (as depicted in figure 70). In the example of speech recognition in call centre operations, NLP needs to detect the language of the person, detect the sequences of word and potentially detect emotions in the way the message is communicated.

Because of the size of different combinations that are possible in NLP, the development of NLP applications relies heavily on Big Data environments. The English dictionary consists of approximately 170,000 words and the number of Chinese words is approximately 370,000. The number of combinations that can be made with these are almost infinite. The amount of data necessary to store and process these combinations consist of multiple zettabytes.

Key challenges in NLP have to do with syntax and semantics. Syntax is the way sentences are constructed and how combinations of words give meaning to a sentence. Semantics on the other hand deals with the fact that (combinations of) words can have a different meaning when they are used in different contexts. A "half glass full" can have the literal meaning that the glass is filled with a liquid but can also mean an optimistic attitude of a person.

Knowledge Representation

Knowledge representation is the field of Artificial Intelligence dedicated to representing information about the world in a form that a computer system can utilize to solve complex tasks. Knowledge representation incorporates findings from psychology about how humans solve problems and represent knowledge in order to design logical statements that make complex systems easier to design and build. As such, it heavily relies on the application of logic in order to model reasoning.

Since most data sets are heterogeneous in terms of their type, structure and accessibility, they pose challenges for computer systems to interpret them in a systematic manner. Knowledge representation helps to identify where data is stored and how it can be retrieved at a later stage when it requires processing. In particular, it aims at building systems that know about their world and are able to act in an informed way in it, as humans do. A crucial part of these systems is that knowledge is represented symbolically, and that reasoning procedures are able to extract consequences of such knowledge as new symbolic representations.[66]

Automated Reasoning

Automated reasoning in Artificial Intelligence is the knowledge capability that concerns itself with understanding reasoning capabilities in computer systems. The goal of automated reasoning is to design computer systems that can reason completely automatically (without human involvement). Automated reasoning is necessary in the design of any Artificial Intelligence system in order to mimic the process that happens in the human brain. Given the conditions that can be observed or sensed, the computer system needs to arrive at the best possible conclusion by following an (automated) thought process.

Machine Learning

Machine Learning, as introduced in Chapter 1.7, is one of the fundamental capabilities required for both Big Data analytics as well as Artificial Intelligence. The objective of machine learning is to design a system that improves and gets better over time. Just like humans memorize information or relationships when they are presented to them, so can computer systems learn from previous interactions. Common machine learning algorithms (classification, regression and clustering) were discussed in Chapter 5. A more in-depth review will be discussed in the Enterprise Big Data Scientist guide.

8.5 Deep Learning

Deep Learning is an area of interest that has grown very rapidly in the last decade because of the increased interest in Artificial Intelligence. In short, Deep Learning is an advanced form of machine learning that utilizes learning data representations (as opposed to algorithms) that is specifically used for Artificial Intelligence. As discussed in Chapter 1.8, Deep Learning is a subset of Artificial Intelligence that has evolved from beginning of the 21st century onwards.

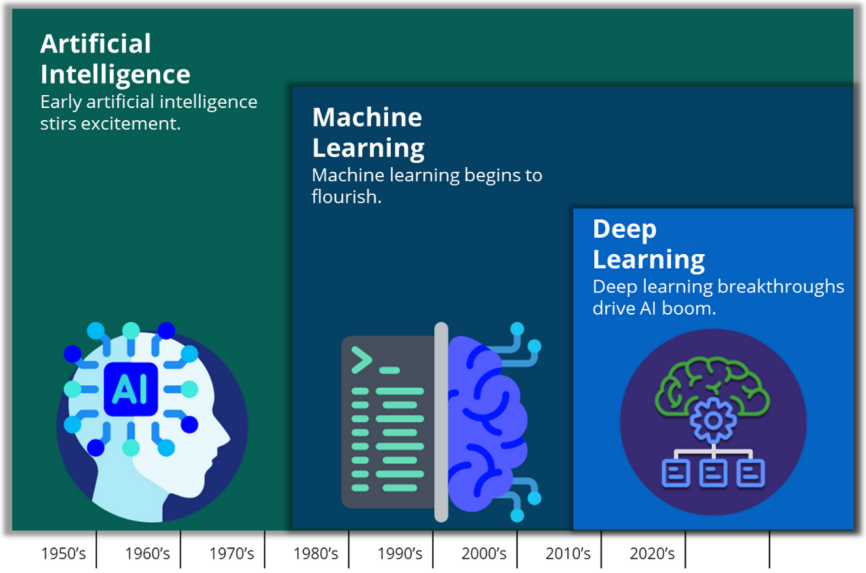

Figure 8.5: The evolution of AI, ML and Deep Learning

Deep learning is a type of machine learning that can process a wider range of data resources, requires less data pre-processing by humans, and can often produce more accurate results than traditional machine learning approaches (although it requires a larger amount of data to do so).

In deep learning, interconnected layers of software-based calculator, known as "neurons," form a neural network. The network can ingest vast amounts of input data and process them through multiple layers that learn increasingly complex features of the data at each layer. The network can then make a determination about the data, learn if its determination is correct, and use what it has learned to make determinations about new data. For example, once it learns what an object looks like, it can recognize the object in a new image.

Conventional machine-learning techniques that are used in the analysis of Big Data are limited in their ability to process data in their raw form. In the example of facial recognition systems, raw data (i.e., photos of individuals) needs to be transformed in feature vectors that can be compared with other vectors in the system. If a match between two feature factors is found, the person has been 'recognized.' In order to translate this 'raw data' into a useful vector, careful engineering and extensive domain expertise is required.

Deep Learning solves this problem using learning data presentation. This allows a machine to be fed with data and automatically discover the representations needed for detection or classification. In order to achieve this, Deep Learning breaks down raw data into a number of layers (using the back propagation algorithm), and subsequently compares these layers with each other. Using this technique, it becomes more efficient to break down large data sets into structured pieces of information that can be analyzed. Deep Learning is predominantly used in processing images, video, speech and audio.

8.6 Next Steps

In this final section of the Enterprise Big Data Professional guide, we have provided a high level overview of the relationship between the knowledge domain of Big Data and the knowledge domain of Artificial Intelligence. Both domains are rooted on the same foundation, which requires the capability to analyse massive quantities of data in order to make split-second decisions. Yet the underlying techniques and algorithms that analyse 'Big Data' vary based on the business objectives.

A more in-dept review about Big Data analysis techniques and Big Data algorithms is further discussed in the Enterprise Big Data Analyst and Enterprise Big Data Scientist professional guides. We highly encourage students and data enthusiasts to download copies of these guides in order to further study the world of Big Data.

Notes

61 Kurzweil, R., Richter, R., Kurzweil, R. and Schneider, M.L., 1990. The age of intelligent machines (Vol. 579). Cambridge: MIT press.

62 Turing, A.M., 2009. Computing machinery and intelligence. In Parsing the Turing Test (pp. 23-65). Springer, Dordrecht.

63 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice Hall.

64 Chui, M., 2017. Artificial intelligence the next digital frontier?. McKinsey and Company Global Institute, p.47.

65 Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice hall.

66 Baral, C. and De Giacomo, G., 2015, January. Knowledge Representation and Reasoning: What's Hot. In AAAI (pp. 4316-4317).

Glossary

The following is list of common Big Data terms and definitions, as used throughout this publication:

- **Algorithm**. A procedure or formula for solving a problem based on conducting a sequence of specified actions. In the context of Big Data, algorithm refers to a mathematical formula embedded in software to perform an analysis on a set of data.
- **Analytics**. Analytics is the systematic processing and manipulation of data to uncover patterns, relationships between data, historical trends and attempts at predictions of future behaviours and events.
- **Artificial Intelligence**. The simulation of human intelligence processes by machines, especially computer systems. These machines can perceive the environment and take corresponding required actions and even learn from those actions.
- **Big Data**. Big Data is the knowledge domain that explores the techniques, skills and technology to deduce valuable insights out of massive quantities of data.
- **Big Data Engineering**. Big Data Engineering is the discipline for engineering scalable systems for data-intensive processing.
- **Data Analysis Process**. The Data Analysis Process is the set of processes that is guided by the organizational need to transform raw data into actionable knowledge, which includes data collection, preparation, analytics, visualization, and access.
- **Data Governance**. Data Governance refers to a system, including policies, people, practices, and technologies, necessary to ensure data management within an organization.
- **Data Science**. Data Science is the methodology for the synthesis of useful knowledge directly from data through a process of discovery or of hypothesis formulation and hypothesis testing.
- **Data Scientist**. A Data Scientist is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the Data Analysis Process.
- **Data Set**. A collection of related, discrete items of data that may be accessed individually or collectively, or managed as a single, holistic entity. Data sets are generally organized into some formal structure, often in a tabular format.
- **Distributed Computing**. Distributed Computing is a computing system in which components located on networked computers communicate and coordinate their actions by passing messages.
- **Hadoop**. An open-source distributed processing framework that manages data processing and storage for Big Data applications. It provides a reliable means for managing pools of Big Data and supporting related analytics applications.
- **Hadoop Distributed File System**. The primary data storage system used by Hadoop HDFS employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters.

- **Machine Learning**. A type of artificial intelligence that improves software applications' ability to predict accurate outcomes without being explicitly programmed to do so. Common use cases for machine learning include recommendation engines, fraud and malware threat detection, business process automation and predictive maintenance.
- **MapReduce**. MapReduce is the technique to allow queries to run across distributed data nodes. One of the core components of the Apache Hadoop software framework.
- **Metadata**. Metadata is data employed to annotate other data with descriptive information, possibly including their 245 data descriptions, data about data ownership, access paths, access rights, and data volatility.
- **Natural Language Processing (NLP)**. A computer program's ability to understand both written and spoken human language. A component of artificial intelligence, NLP has existed for over five decades and has roots in the field of linguistics.
- **No-SQL**. An approach to database design that can accommodate a wide variety of data models, including key-value, document, columnar and graph formats. NoSQL, which stands for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schemais carefully designed before the database is built.
- **SQL**. SQL is the Structured Query Language (SQL) standard used to query r elational databases.
- **Structured Data**. Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis.
- **Unstructured Data**. Everything that can't be organized in the manner of structured data. Unstructured data includes emails and social media posts, blogs, and messages, transcripts of audio recordings of people's speech, images and video files, and machine data, such as log files from websites, servers, networks and applications.