Data Science: Data science is the field of exploring, manipulating, and analysing data, and using data to answer questions or make recommendations.

Data science can help organisations

1. understand their environments
2. analyse existing issues
3. reveal previously hidden opportunities.

The three important qualities of a Data Scientist are being curious, judgemental, and argumentative.

Term	Definition
Algorithms	A set of step-by-step instructions to solve a problem or complete a task.
Model	A representation of the relationships and patterns found in data to make predictions or analyse complex systems retaining essential elements needed for analysis.
Outliers	When a data point or points occur significantly outside of most of the other data in a data set, potentially indicating anomalies, errors, or unique phenomena that could impact statistical analysis or modelling.
Quantitative analysis	A systematic approach using mathematical and statistical analysis is used to interpret numerical data.
Structured data	Data is organised and formatted into a predictable schema, usually related tables with rows and columns.
Unstructured data	Unorganised data that lacks a predefined data model or organisation makes it harder to analyse using traditional methods. This data type often includes text, images, videos, and other content that doesn't fit neatly into rows and columns like structured data.

Some of the data types used are:

Delimited text file formats,
Microsoft Excel Open XML Spreadsheet, or XLSX
Extensible Markup Language, or XML,
Portable Document Format, or PDF,
JavaScript Object Notation, or JSON.

Using complicated machine learning algorithms does **not** always guarantee achieving a better performance. Occasionally, a simple algorithm such as *k*-nearest neighbour can yield a satisfactory performance comparable to the one achieved using a complicated algorithm. It all depends on the data.

The availability of vast amounts of data, and the competitive advantage that analysing it brings, has triggered digital transformations throughout many industries. Netflix moved from being a postal DVD lending system to one of the world's foremost video streaming providers, the Houston Rockets NBA team used data gathered by overhead cameras to analyse the most productive plays, and Lufthansa analysed customer data to improve its service.

Cloud Computing:

Cloud computing, also referred to as the cloud, is the delivery of on-demand computing resources such as networks, servers, storage, applications, services, and data centers over the Internet on a pay-for-use basis.
The term "cloud computing" can be used to describe applications and data that users access over the Internet rather than on their local computer. Examples of cloud computing include users using online web apps, employees using secure online business applications to conduct their work, and users storing personal files on cloud-based storage platforms such as Google Drive, OneDrive, and Dropbox.

Cloud computing is composed of five essential characteristics, three deployment models, and three service models.

five essential characteristics of the cloud:

1. On-demand self-service(Processing power, Storage, Network)
2. Broad Network access(can be accessed via mobile,desktops tablets)
3. Resource Pooling
4. Rapid elasticity
5. Measured Service (pay what u use)

Cloud Deployment models:

1. Public
2. Private (On premise for organisation)
3. Hybrid

Cloud Service Models:

1. IaaS (Servers, Storage, Network)
2. Paas (Software and Hardware tools)
3. Saas (Application as hosted)

Cloud allows you to bypass the physical limitations of the computers and the systems you're using and it allows you to deploy the analytics and storage capacities of advanced machines that do not necessarily have to be your machine. Cloud allows you not just to store large amounts of data on servers somewhere, but it also allows you to deploy very advanced computing algorithms and the ability to do high-performance computing using machines that are not yours.

Big Data:

“Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.”

The V’s of Big Data

1. Velocity (Data generating in rapid pace)
2. Volume (More and more data is generating)
3. Variety (Structured and unstructured)
4. Veracity (Quality and Origin of data)
5. Value (Ability and Need of data to turn use)

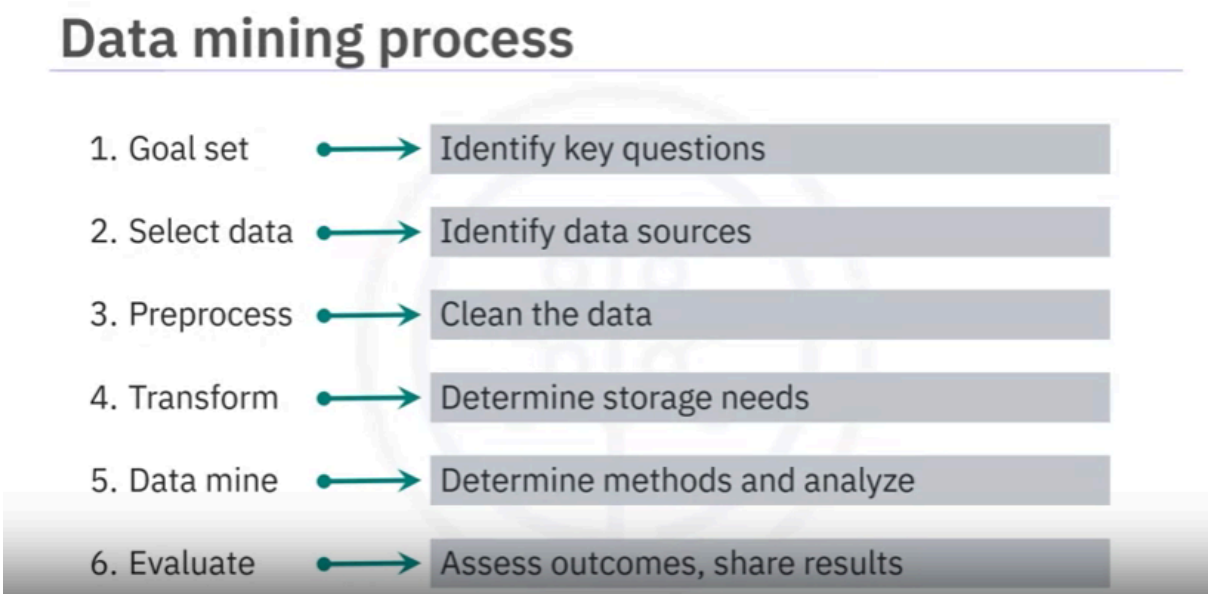
Big Data Processing Tools:

Big Data processing technologies provide ways to work with large sets of structured, semi-structured, and unstructured data so that value can be derived from big data.

Apache Hadoop
Hadoop is a collection of tools that provides distributed storage and processing of big data.

Apache Hive
Hive is a data warehouse for data query and analysis built on top of Hadoop.

Apache Spark
Spark is a distributed data analytics framework designed to perform complex data analytics in real-time.



Establishing Data Mining Goals:

The first step in data mining requires you to set up goals for the exercise. Obviously, you must identify the key questions that need to be answered. However, going beyond identifying the key questions are the concerns about the costs and benefits of the exercise. Furthermore, you must determine, in advance, the expected level of accuracy and usefulness of the results obtained from data mining. If money were no object, you could throw as many funds as necessary to get the answers required. However, the cost-benefit trade-off is always instrumental in determining the goals and scope of the data mining exercise. The level of accuracy expected from the results also influences the costs. High levels of accuracy from data mining would cost more and vice versa. Furthermore, beyond a certain level of accuracy, you do not gain much from the exercise, given the diminishing returns. Thus, the cost-benefit trade-offs for the desired level of accuracy are important considerations for data mining goals.

Selecting Data

The output of a data-mining exercise largely depends upon the quality of data being used. At times, data is readily available for further processing. For instance, retailers often possess large databases of customer purchases and demographics. On the other hand, data may not be readily available for data mining. In such cases, you must identify other sources of data or even plan new data collection initiatives, including surveys. The type of data, its size, and frequency of collection have a direct bearing on the cost of data mining exercise. Therefore, identifying the right kind of data needed for data mining that could answer the questions at reasonable costs is critical.

Preprocessing Data

Preprocessing data is an important step in data mining. Often raw data are messy, containing erroneous or irrelevant data. In addition, even with relevant data, information is sometimes missing. In the preprocessing stage, you identify the irrelevant attributes of data and expunge such attributes from further consideration. At the same time, identifying the erroneous aspects of the data set and flagging them as such is necessary. For instance, human error might lead to inadvertent merging or incorrect parsing of information between columns. Data should be subject

to checks to ensure integrity. Lastly, you must develop a formal method of dealing with missing data and determine whether the data are missing randomly or systematically.

If the data were missing randomly, a simple set of solutions would suffice. However, when data are missing in a systematic way, you must determine the impact of missing data on the results. For instance, a particular subset of individuals in a large data set may have refused to disclose their income. Findings relying on an individual's income as input would exclude details of those individuals whose income was not reported. This would lead to systematic biases in the analysis. Therefore, you must consider in advance if observations or variables containing missing data be excluded from the entire analysis or parts of it.

Transforming Data

After the relevant attributes of data have been retained, the next step is to determine the appropriate format in which data must be stored. An important consideration in data mining is to reduce the number of attributes needed to explain the phenomena. This may require transforming data Data reduction algorithms, such as Principal Component Analysis (demonstrated and explained later in the chapter), can reduce the number of attributes without a significant loss in information. In addition, variables may need to be transformed to help explain the phenomenon being studied. For instance, an individual's income may be recorded in the data set as wage income; income from other sources, such as rental properties; support payments from the government, and the like. Aggregating income from all sources will develop a representative indicator for the individual income.

Often you need to transform variables from one type to another. It may be prudent to transform the continuous variable for income into a categorical variable where each record in the database is identified as low, medium, and high-income individual. This could help capture the non-linearities in the underlying behaviours.

Storing Data

The transformed data must be stored in a format that makes it conducive for data mining. The data must be stored in a format that gives unrestricted and immediate read/write privileges to the data scientist. During data mining, new variables are created, which are written back to the original database, which is why the data storage scheme should facilitate efficiently reading from

and writing to the database. It is also important to store data on servers or storage media that keeps the data secure and also prevents the data mining algorithm from unnecessarily searching for pieces of data scattered on different servers or storage media. Data safety and privacy should be a prime concern for storing data.

Mining Data

After data is appropriately processed, transformed, and stored, it is subject to data mining. This step covers data analysis methods, including parametric and non-parametric methods, and machine-learning algorithms. A good starting point for data mining is data visualisation. Multidimensional views of the data using the advanced graphing capabilities of data mining software are very helpful in developing a preliminary understanding of the trends hidden in the data set.

Later sections in this chapter detail data mining algorithms and methods.

Evaluating Mining Results

After results have been extracted from data mining, you do a formal evaluation of the results. Formal evaluation could include testing the predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data. This is known as an "in-sample forecast". In addition, the results are shared with the key stakeholders for feedback, which is then incorporated in the later iterations of data mining to improve the process.

Data mining and evaluating the results becomes an iterative process such that the analysts use better and improved algorithms to improve the quality of results generated in light of the feedback received from the key stakeholders.

\

Artificial Intelligence and Data Science:

Machine learning is a subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what it is learned without being explicitly programmed. Machine learning algorithms are trained with large sets of data and they learn from examples. They do not follow rules-based algorithms. Machine learning is what enables machines to solve problems on their own and make accurate predictions using the provided data.

Deep learning is a specialised subset of machine learning that uses layered neural networks to simulate human decision-making. Deep learning algorithms can label and categorize information and identify patterns. It is what enables AI systems to continuously learn on the job and improve the quality and accuracy of results by determining whether decisions were correct.

Artificial neural networks, often referred to simply as neural networks, take inspiration from biological neural networks, although they work quite a bit differently. A neural network in AI is a collection of small computing units called neurons that take incoming data and learn to make decisions over time. Neural networks are often layer-deep and are the reason deep learning algorithms become more efficient as the data sets increase in volume, as opposed to other machine learning algorithms that may plateau as data increases.

Generative AI:

Generative AI is a subset of artificial intelligence that focuses on producing new data rather than just analysing existing data. It allows machines to create content, including images, music, language, computer code, and more, mimicking creations by people.

How Generative AI Works:

Generative AI operates, with Deep learning models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are at the foundation of this technique. These models create new instances that replicate the underlying distribution of the original data by learning patterns from enormous volumes of data.

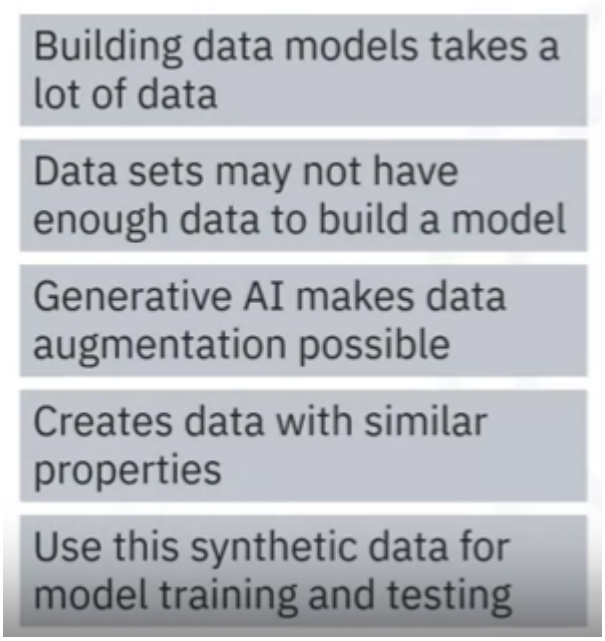
Applications of Generative AI:

Generative AI has found diverse applications across various industries. Let's look at some fascinating examples! Natural language processing like OpenAI's GPT-3 can generate human-like text, revolutionising content creation and chatbots. In healthcare, Generative AI can synthesise medical images, aiding in the training of medical professionals. Generative AI can create unique and visually stunning artworks, generating endless creative visual compositions. Game developers use Generative AI to generate realistic environments, characters, and game levels. Generative AI assists in fashion by designing new styles and creating personalized shopping recommendations.

Synthetic Data:

Artificially generated data with properties similar to real data, used by data scientists to augment their datasets and improve model training.

How Data Scientists use Generative AI:



Regression:

Regression identifies the strength and amount of the correlation between one or more inputs and an output. For instance, how much does the price of a house increase based on its square footage and number of bedrooms, and how confident can you be of this relationship?

Natural Language Processing (NLP):

A field of AI that enables machines to understand, generate, and interact with human language, revolutionising content creation and chatbots.

Generative AI helps bridge the gap by allowing the exploration and evaluation of multiple hypotheses from various data sources, enhancing decision-making.

Applications of Data Science:

1. Recommendation Engine
2. Personal AI assistant
3. Google Recommendations by data collecting
4. Business enhancement

Data Science can predict Natural Disasters

Used in Medical Science

- The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.
- The report should present a thorough analysis of the data and communicate the project findings.
- Companies should look for someone excited about working with the data in their particular industry. They should seek out someone curious who can ask interesting, meaningful questions about the types of data they intend to collect. They should hire people who love working with data, are fluent in statistics, and are competent in applying machine learning algorithms.
- A clearly organised and logical report should communicate the following to the reader:
    - What they gain by reading the report
    - Clearly defined goals
    - The significance of your contribution
    - Appropriate context by giving sufficient background
    - Why this work is practical and useful
    - Conjecture plausible future developments that might result from your work

Case Study: Lila's Journey to Becoming a Data Scientist: Her Working Approach on the First Task

This case study explores the data scientist's career path and key attributes, highlighting the skills, education, and experiences required to excel in this dynamic field. We'll follow the story of Lila, a fictional individual who aspires to become a successful data scientist.
There will be a quiz after this reading based on the contents of this case study.

1. Education and Skill Acquisition

With an economics undergraduate degree and a substantial data analysis background, Lila finds data science and its potential to drive meaningful change captivating. Inspired by her experiences, she makes a determined decision to transition her career and step into the role of a data scientist.
Lila realizes that to embark on her data science journey, she needs to enhance her skills and knowledge. She enrolled in the IBM Data Science Professional Certificate online program that covers key topics like statistics, machine learning, data analysis, and programming languages like Python and SQL. She diligently completes coursework and practises her coding skills on real datasets.

2. Building a Strong Foundation

As she progresses in her studies, Lila gains a deep understanding of data science fundamentals such as data manipulation and visualisation with Python libraries like NumPy, Pandas, and Matplotlib. This strong foundation equips her with essential skills for data analysis.

3. Visualisation for Storytelling

Lila knows she must communicate her findings effectively, so she learns which types of data visualisations will be most informative. She learns to create charts and graphs that visually represent data like sales trends, customer segmentation, and product popularity, allowing stakeholders to grasp the data's significance. These visualisations help in storytelling and decision-making.

4. Hands-On Experience

Lila understands that practical experience is invaluable in data science. She started participating in Kaggle competitions and working on personal data projects. These experiences expose her to real-world data problems and help her develop problem-solving skills. Furthermore, she created her GitHub account and uploaded her projects to build her profile.

5. Data Wrangling and Preprocessing

Lila learns that data scientists spend a significant portion of their time on data cleaning and preprocessing. She worked on various datasets, learned data preprocessing as she used sed NumPy and pandas Python libraries, and became skilled in handling missing data, outlier detection, and feature engineering to improve model performance.

6. Communication and Storytelling

Recognizing that data scientists must communicate their findings effectively, Lila honed her data storytelling skills. She learned various tools like matplotlib and plotly while she pursued her IBM Data Science Professional Certificate. She learned how to create compelling visualisations and present her insights in a clear and understandable manner.

7. Networking and Collaboration

Lila actively participates in data science communities and attends meetups and conferences. She collaborates on open-source projects, connects with fellow data scientists, and gains exposure to various industries when she attends the IBM TechXchange Conference.

8. Domain Expertise

Understanding that domain knowledge is crucial, Lila chooses a niche that aligns with her interests. She looks deeply into several domains, including e-commerce, healthcare, finance, and several other fields to which she could apply her data science skills effectively. Since her master's in economics, she chose e-commerce as her core domain to land herself a data science career.

9. Landing the First Job

After months of preparation, Lila started applying for data scientist positions. She tailors her resume to highlight her relevant skills and projects. Her online portfolio showcases her capabilities and demonstrates her commitment to the field.

10. Lila's Approach to Working on Her First Task as a Data Scientist

As a newly hired junior data scientist at a retail company, Lila uses data insights to improve customer service. Her first assignment involves diving into customer data to identify patterns and anomalies that could impact customer service. She uses data analysis to enhance the overall customer experience.

11. Dataset Selection and Sourcing

In the initial phase of her data science journey, Lila faced the challenge of selecting a suitable dataset and procuring it from different sources. Apart from the historical data available for the organisations for the past four years, she scoured various repositories, websites, and databases to find the right datasets for her project. Upon collecting data from diverse sources, Lila encountered another crucial decision point. She had to decide how to harmonise and integrate these disparate datasets into a cohesive whole. She reached out to product professionals, data engineers, and domain specialists, seeking their input and expertise in merging datasets.

12. Data Understanding and Cleaning

Lila begins by importing the dataset into her data analysis environment using Python and SQL. She loads the data and examines the first few rows to understand its structure and contents. Upon acquiring the dataset, Lila encounters her first challenge: data cleaning. Lila checks for missing values, duplicates, and outliers in the dataset. She addresses missing data by imputing or removing rows or columns with missing values. Outliers are identified and treated appropriately based on their impact on the analysis.

13. Exploratory Data Analysis (EDA)

As she delves into exploratory data analysis, Lila faces numerous choices. She must determine which summary statistics, visualisations, and distribution analyses will best reveal insights into customer behaviour and sales trends. Each choice she makes during EDA influences the story the data will tell. Lila conducts EDA to gain insights into the dataset. She generates summary statistics and visualisations (histograms, scatter plots) and explores the distribution of variables. EDA helps her understand customer behaviour, popular products, and sales trends.

14. Feature Engineering

Lila recognizes the potential for feature engineering to enhance her analysis. She assesses whether creating new features, such as calculating total purchase amounts, will improve the dataset's utility for her project.

15. Statistical Analysis, Machine Learning

Lila evaluates whether statistical tests or machine learning algorithms are necessary. She employs regression analysis to understand relationships between variables and explore machine learning models for demand forecasting or customer segmentation tasks. Lila also performs statistical tests to uncover patterns in the data. She uses regression analysis to understand the impact of unit price on sales.

16. Presentation and Reporting

At the culmination of her analysis, Lila faces the challenge of presenting her findings. Lila compiles her analysis and findings using a Jupyter Notebook into a comprehensive report and presentation. She highlights actionable insights and recommendations for the e-commerce platform's stakeholders.

17. Continuous Learning

After completing her first project, Lila continues to refine her skills, explores more complex datasets, and tackles increasingly challenging data science tasks.

18. Machine Learning Skills

Although Lila took an introductory course on Machine Learning as part of the IBM Data Science Professional Certificate, the field intrigues her, and she wants to develop her skills further by taking the IBM Machine Learning Professional Certificate. She identified Machine Learning Repository datasets in the course and experimented with various algorithms. Lila dives into machine learning to excel as a data scientist, wherein she studies various algorithms, such as linear regression, decision trees, and deep learning models. She continues to gain expertise in selecting and fine-tuning algorithms based on specific data problems.

Types of Data:

1. Structured
2. Semi-Structured
3. Unstructured

1. Structured Data:

structured data is data that is well organised in formats that can be stored in databases and lends itself to standard data analysis methods and tools.
Some of the sources of structured data could include: SQL Databases and Online Transaction Processing (or OLTP) Systems that focus on business transactions, Spreadsheets such as Excel and Google Spreadsheets, Online forms, Sensors such as Global Positioning Systems (or GPS) and Radio Frequency Identification (or RFID) tags; and Network and Web server logs. You can typically store structured data in relational or SQL databases. You can also easily examine structured data with standard data analysis methods and tools.

2. Semi-Structured Data:

Semi-structured data is data that is somewhat organised and relies on meta tags for grouping and hierarchy
Some of the sources of semi-structured data could include: E-mails, XML, and other markup languages, Binary executables, TCP/IP packets, Zipped files, Integration of data from different sources. XML and JSON allow users to define tags and attributes to store data in a hierarchical form and are used widely to store and exchange semi-structured data.

3. **Unstructured Data:**

Unstructured data is data that is not conventionally organised in the form of rows and columns in a particular format.

Some of the sources of unstructured data could include: Web pages, Social media feeds, Images in varied file formats (such as JPEG, GIF, and PNG), video and audio files, documents and PDF files, PowerPoint presentations, media logs; and surveys. Unstructured data can be stored in files and documents (such as a Word doc) for manual analysis or in NoSQL databases that have their own analysis tools for examining this type of data.

Data Sources:

1. **Relational Databases:**

organisations have internal applications to support them in managing their day to day business activities, customer transactions, human resource activities, and their workflows. These systems use relational databases such as SQL Server, Oracle, MySQL, and IBM DB2, to store data in a structured way. Data stored in databases and data warehouses can be used as a source for analysis. For example, data from a retail transactions system can be used to analyze sales in different regions, and data from a customer relationship management system can be used for making sales projections.

2. **Flat File and XML Datasets:**

External to the organisation, there are other publicly and privately available datasets. For example, government organisations releasing demographic and economic datasets on an ongoing basis. Then there are companies that sell specific data, for example, Point-of-Sale data or Financial data, or Weather data, which businesses can use to define strategy, predict demand, and make decisions related to distribution or marketing promotions, among other things. Such data sets are typically made available as flat files, spreadsheet files, or XML documents. Flat files store data in plain text format, with one record or row per line, and each value separated by delimiters such as commas, semi-colons, or tabs. Data in a flat file maps to a single table, unlike relational databases that contain multiple tables. One of the most common flat-file formats is CSV in which values are separated by commas. Spreadsheet files are a special type of flat files, that also organise data in a tabular format–rows and columns. But a spreadsheet can contain multiple worksheets, and each worksheet can map to a different table. Although data in spreadsheets is in plain text, the files can be stored in custom formats and include additional information such as formatting, formulas, etc. Microsoft Excel, which stores data in .XLS or .XLSX format is probably the most common spreadsheet. Others include Google sheets, Apple Numbers, and LibreOffice. XML files contain data values that are identified or marked up using tags. While data in flat files is “flat” or maps to a single table, XML files can support more complex data structures, such as hierarchical. Some common uses of XML include data from online surveys, bank statements, and other unstructured data sets.

3. **APIs and Web Services:**

Many data providers and websites provide APIs, or Application Program Interfaces, and Web Services, which multiple users or applications can interact with and obtain data for processing or analysis. APIs and Web Services typically listen for incoming requests, which can be in the form of web requests from users or network requests from applications, and return data in plain text, XML, HTML, JSON, or media files. Let’s look at some popular examples of APIs being used as a data source for data analytics: The use of Twitter and Facebook APIs to source data from tweets and posts for performing tasks such as opinion mining or sentiment analysis—which is to summarise the amount of appreciation and criticism on a given subject, such as policies of a

government, a product, a service, or customer satisfaction in general. Stock Market APIs used for pulling data such as share and commodity prices, earnings per share, and historical prices, for trading and analysis. Data Lookup and Validation APIs, which can be very useful for Data Analysts for cleaning and preparing data, as well as for correlating data—for example, to check which city or state a postal or zip code belongs to. APIs are also used for pulling data from database sources, within and external to the organisation.

4. **Web Scraping:**

Web ScrapingWeb scraping is used to extract relevant data from unstructured sources. Also known as screen scraping, web harvesting, and web data extraction, web scraping makes it possible to download specific data from web pages based on defined parameters. Web scrapers can, among other things, extract text, contact information, images, videos, product items, and much more from a website. Some popular uses of web scraping include collecting product details from retailers, manufacturers, and eCommerce websites to provide price comparisons; generating sales leads through public data sources; extracting data from posts and authors on various forums and communities; and collecting training and testing datasets for machine learning models Some of the popular web scraping tools include BeautifulSoup, Scrapy, Pandas, and Selenium.

5. **Data Streams and feeds:**

Data streams are another widely used source for aggregating constant streams of data flowing from sources such as instruments, IoT devices, and applications, GPS data from cars, computer programs, websites, and social media posts. This data is generally timestamped and also geo-tagged for geographical identification. Some of the data streams and ways in which they can be leveraged include: stock and market tickers for financial trading; retail transaction streams for predicting demand and supply chain management; surveillance and video feeds for threat detection; social media feeds for sentiment analysis;sensor data feeds for monitoring industrial or farming machinery; web click feeds for monitoring web performance and improving design; and real-time flight events for rebooking and rescheduling. Some popular applications used to process data streams include Apache Kafka, Apache Spark Streaming, and Apache Storm. RSS (or Really Simple Syndication) feeds are another popular data source. These are typically used for capturing updated data from online forums and news sites where data is refreshed on an ongoing basis. Using a feed reader, which is an interface that converts RSS text files into a stream of updated data, updates are streamed to user devices.

Metadata and Metadata Management

What is metadata?

Metadata is data that provides information about other data.
This is a very broad definition. Here we will consider the concept of metadata within the context of databases, data warehousing, business intelligence systems, and all kinds of data repositories and platforms.
We'll consider the following three main types of metadata:

- Technical metadata
- Process metadata, and
- Business metadata

Technical metadata

Technical metadata is metadata which defines the data structures in data repositories or platforms, primarily from a technical perspective.
For example, technical metadata in a data warehouse includes assets such as:

- Tables that record information about the tables stored in a database, like:
    - each table's name
    - the number of columns and rows each table has
- A data catalog, which is an inventory of tables that contain information, like:
    - the name of each database in the enteprise data warehouse
    - the name of each column present in each database
    - the names of every table that each column is contained in
    - the type of data that each column contains

The technical metadata for relational databases is typically stored in specialised tables in the database called the System Catalog.

Process metadata

Process metadata describes the processes that operate behind business systems such as data warehouses, accounting systems, or customer relationship management tools.
Many important enterprise systems are responsible for collecting and processing data from various sources. Such critical systems need to be monitored for failures and any performance anomalies that arise. Process metadata for such systems includes tracking things like:

- process start and end times
- disk usage
- where data was moved from and to, and
- how many users access the system at any given time

This sort of data is invaluable for troubleshooting and optimising workflows and ad hoc queries.

Business metadata

Users who want to explore and analyse data within and outside the enterprise are typically interested in data discovery. They need to be able to find data which is meaningful and valuable to them and know where that data can be accessed from. These business-minded users are thus interested in business metadata, which is information about the data described in readily interpretable ways, such as:

- how the data is acquired
- what the data is measuring or describing
- the connection between the data and other data sources

Business metadata also serves as documentation for the entire data warehouse system.

Managing metadata

Managing metadata includes developing and administering policies and processes to ensure information can be accessed and integrated from various sources and appropriately shared across the entire enterprise.
Creation of a reliable, user-friendly data catalogue is a primary objective of a metadata management model. The data catalogue is a core component of a modern metadata management system, serving as the main asset around which metadata management is administered. It serves as the basis by which companies can inventory and efficiently organise their data systems. A modern metadata management model will include a web-based user interface that enables engineers and business users to easily search for and find information on key attributes such as CustomerName or ProductType. This kind of model is central to any Data Governance initiative.

Why is metadata management important?

Good metadata management has many valuable benefits. Having access to a well implemented data catalogue greatly enhances data discovery, repeatability, governance, and can also facilitate access to data.
Well managed metadata helps you to understand both the business context associated with the enterprise data and the data lineage, which helps to improve data governance. Data lineage provides information about the origin of the data and how it gets transformed and moved, and thus it facilitates tracing of data errors back to their root cause. Data governance is a data management concept concerning the capability that enables an organisation to ensure that high data quality exists throughout the complete lifecycle of the data, and data controls are implemented that support business objectives.
The key focus areas of data governance include availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be used by the entire organisation.

Popular tools for metadata management

Popular metadata management tools include:

- IBM InfoSphere Information Server
- CA Erwin Data Modeler
- Oracle Warehouse Builder
- SAS Data Integration Server
- Talend Data Fabric
- Alation Data Catalog
- SAP Information Steward
- Microsoft Azure Data Catalog
- IBM Watson Knowledge Catalog
- Oracle Enterprise Metadata Management (OEMM)
- Adaptive Metadata Manager

- Unifi Data Catalog
- data.world
- Informatica Enterprise Data Catalog

Term	Definition
ACID-compliance	Ensuring data accuracy and consistency through Atomicity, Consistency, Isolation, and Durability (ACID) in database transactions.
Cloud-based Integration Platform as a Service (iPaaS)	Cloud-hosted integration platforms that offer integration services through virtual private clouds or hybrid cloud models, providing scalability and flexibility.
Column-based Database	A type of NoSQL database that organises data in cells grouped as columns, often used for systems requiring high write request volume and storage of time-series or IoT data.
Data at rest	Data that is stored and not actively in motion, typically residing in a database or storage system for various purposes, including backup.
Data integration	A discipline involving practices, architectural techniques, and tools that enable organisations to ingest, transform, combine, and provision data across various data types, used for purposes such as data consistency, master data management, data sharing, and data migration.
Data Lake	A data repository for storing large volumes of structured, semi-structured, and unstructured data in its native format, facilitating agile data exploration and analysis.
Data mart	A subset of a data warehouse designed for specific business functions or user communities, providing isolated security and performance for focused analytics.
Data pipeline	A comprehensive data movement process that covers the entire journey of data from source systems to destination systems, which includes data integration as a key component.
Data repository	A general term referring to data that has been collected, organised, and isolated for business operations or data analysis. It can include databases, data warehouses, and big data stores.
Data warehouse	A central repository that consolidates data from various sources through the Extract, Transform, and Load (ETL) process, making it accessible for analytics and business intelligence.

Document-based Database	A type of NoSQL database that stores each record and its associated data within a single document, allowing flexible indexing, ad hoc queries, and analytics over collections of documents.
ETL process	The Extract, Transform, and Load process for data integration involves extracting data from various sources, transforming it into a usable format, and loading it into a repository.
Graph-based Database	A type of NoSQL database that uses a graphical model to represent and store data, ideal for visualising, analysing, and discovering connections between interconnected data points.
Key-value store	A type of NoSQL database where data is stored as key-value pairs, with the key serving as a unique identifier and the value containing data, which can be simple or complex.
Portability	The capability of data integration tools to be used in various environments, including single-cloud, multi-cloud, or hybrid-cloud scenarios, provides flexibility in deployment options.
Pre-built connectors	Catalogued connectors and adapters that simplify connecting and building integration flows with diverse data sources like databases, flat files, social media, APIs, CRM, and ERP applications.
Relational databases (RDBMSes)	Databases that organise data into a tabular format with rows and columns, following a well-defined structure and schema.
Scalability	The ability of a data repository to grow and expand its capacity to handle increasing data volumes and workload demands over time.
Schema	The predefined structure that describes the organisation and format of data within a database, indicating the types of data allowed and their relationships.
Streaming data	Data that is continuously generated and transmitted in real-time requires specialised handling and processing to capture and analyse.
Use cases for relational databases	Applications such as Online Transaction Processing (OLTP), Data Warehouses (OLAP), and IoT solutions where relational databases excel.