Modern Data Ecosystem and the Role of Data Analytics

**Key players in the Data Ecosystem**:

**Data engineering** converts raw data into usable data. **Data analytics** uses this data to generate insights. **Data scientists** use data analytics and data engineering to predict the future using data from the past, **business analysts** and business intelligence analysts use these insights and predictions to drive decisions that benefit and grow their business.

Data Scientist:

Data scientists analyze data for actionable insights and build machine learning or deep learning models that train on past data to create predictive models. Data scientists are people who answer questions such as, how many new social media followers am I likely to get next month, or what percentage of my customers am I likely to lose to competition in the next quarter, or is this financial transaction unusual for this customer? Data scientists require knowledge of mathematics, statistics, and a fair understanding of programming languages, databases, and building data models. They also need to have domain knowledge.

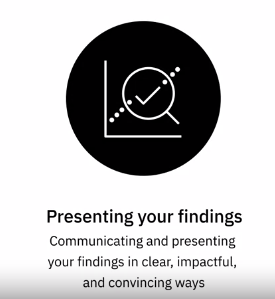
Data Analysis:

Data analysis is the process of gathering, cleaning, analyzing and mining data, interpreting results, and reporting the findings. With data analysis we find patterns within data and correlations between different data points. And it is through these patterns and correlations that insights are generated, and conclusions are drawn. Data analysis helps businesses understand their past performance and informs their decision-making for future actions. Using data analysis, businesses can validate a course of action before committing to it. Saving valuable time and resources and also ensuring greater success.

The Data Analysis Process:







Data Analytics:

Data analytics as a process or a phenomenon of taking information gathered from a relevant population, maybe our customers or our social audience, and breaking that information down into subsets, and using that data to make decisions about products or services that we want to offer, or in cases of the digital environment that we're in, making decisions about certain pieces of content that we want to publish so that it appeals to our target audience.

**Descriptive Analytics,** that helps decode “What happened.”

**Diagnostic Analytics**, that helps us understand “Why it happened.”

**Predictive Analytics**, that analyzes historical data and trends to suggest “What will happen next.”

**Prescriptive Analytics**, that prescribes “What should be done next.”

Responsibilities of a Data Analyst:

Acquiring data from primary and secondary data sources, Creating queries to extract required data from databases and other data collection systems, Filtering, cleaning, standardizing, and reorganizing data in preparation for data analysis, Using statistical tools to interpret data sets, Using statistical techniques to identify patterns and correlations in data, Analyzing patterns in complex data sets and interpreting trends, Preparing reports and charts that effectively communicate trends and patterns, Creating appropriate documentation to define and demonstrate the steps of the data analysis process.

Skills that a Data Analyst need:

**For technical skills**:

Expertise in using spreadsheets such as **Microsoft Excel or Google Sheets**

Proficiency in statistical analysis and visualization tools and software

**IBM Cognos, IBM SPSS, Oracle Visual Analyzer, Microsoft Power BI, SAS, and Tableau**

Proficiency in programming languages

**R, Python, and in some cases C++, Java, and MATLAB**

Databases

Good knowledge of **SQL**, and ability to work with **data in relational and NoSQL** databases

The ability to access and extract data from data repositories such as **data marts, data warehouses, data lakes, and data pipelines**

Familiarity with Big Data processing tools such as

**Hadoop, Hive, and Spark.**

**Functional Skills:**

Proficiency in Statistics

Analyze data, validate the analysis, identify fallacies and logical errors

Analytical skills

Research and interpret data, theorize, make forecasts

Problem-solving skills

Come up with possible solutions for a given problem

Probing skills

Identify and define the problem statement and desired outcome

Data Visualization skills

Create clear and compelling visualizations to present the analysis

Project management skills

Manage the process, people, dependencies and timelines

**Soft Skills:**

**Ability to**:

* Work collaboratively with business and cross-functional teams
* Communicate effectively to report and present findings
* Tell a compelling and convincing story
* Gather support and buy-in for work

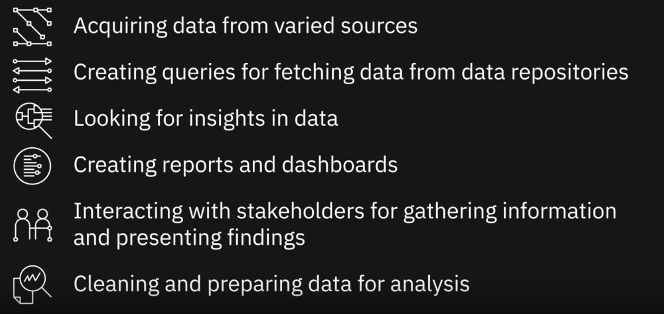
**Curiosity**

Allowing new questions to surface and challenging own assumptions and hypotheses

**Intuition**

Having a sense of the future based on pattern recognition and past experiences

A day in the life of a Data Analyst:



What are some of the applications of Data Analytics in todays’ world?

* Use of sentiment analysis and tweets and stories to inform investment decisions
* Use of satellite imagery data to track the development of industrial activities
* Use of geolocation data to track store traffic and predict sales volume

**The Data Ecosystem**:

We will learn about the different types of data structures, file formats, sources of data, and the languages data professionals use in their day-to-day tasks. We will gain an understanding of various types of data repositories such as Databases, Data Warehouses, Data Marts, Data Lakes, and Data Pipelines. In addition, we will learn about the Extract, Transform, and Load (ETL) Process, which is used to extract, transform, and load data into data repositories. We will gain a basic understanding of Big Data and Big Data processing tools such as Hadoop, Hadoop Distributed File System (HDFS), Hive, and Spark.

Learning Objectives

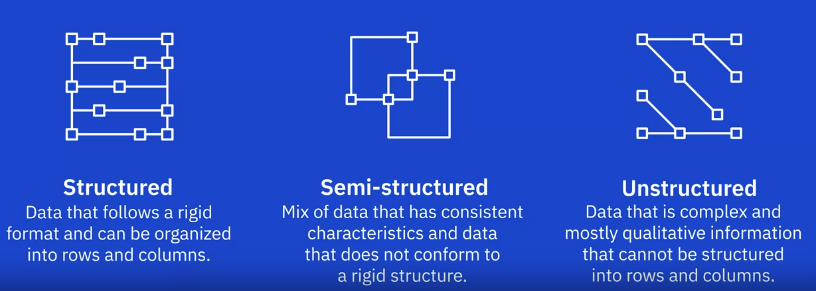
* Describe and differentiate between relational and non-relational database management systems.
* Classify data structures, file formats, and sources of data by their different types.
* Explain the features and use of the different languages used by data professionals.
* Describe how Data Warehouses, Data Marts, Data Lakes, and Data Pipelines work.
* Explain how the Extract, Transform, and Load process works to make raw data ready for analysis.
* Explain what Big Data is.
* Summarize the features and use of some of the Big Data processing tools.

Overview of the Data Analyst Ecosystem

A data analyst’s ecosystem includes the infrastructure, software, tools, frameworks, and processes used to

* Gather data
* Clean data
* Mine data
* Visualize data

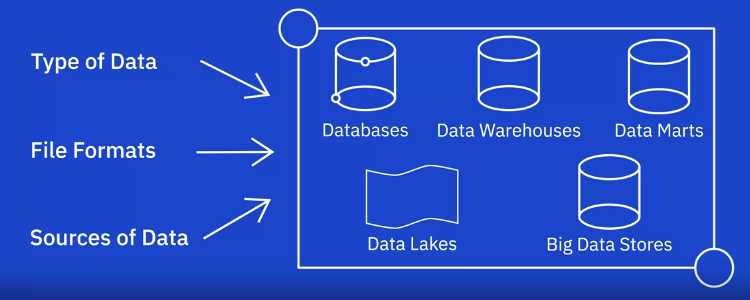
**3 types of Data:**



Data can come in a variety of file formats, such as

* Relational database
* Non-relational database
* APIs
* Web services
* Data streams
* Social platforms
* Sensor devices

**Data Repositories:**



**Languages:**

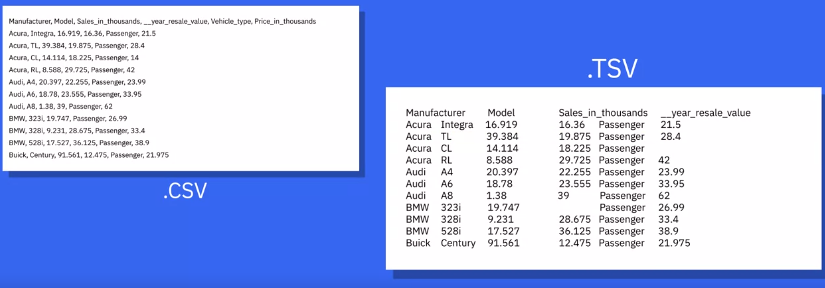
* Query languages such as SQL for querying and manipulating data
* Programming languages such as Python for developing data applications
* Shell and Scripting languages for repetitive operational tasks

Understanding Different types of file formats:

Standard file formats:

1. Delimited text file formats or .CSV
2. Microsoft Excel Open .XML Spreadsheet or .XLSX
3. Extensible Markup Language or .XML
4. Portable Document Format or .PDF
5. JavaScript Object Notation or. JSON

**Delimiter** – a sequence of one or more characters for specifying the boundary between independent entities or values. Comma, Tab, Colon, Vertical Bar, Space



**JSON** – is a text-based open standard designed for transmitting structured data over the web.

**Sources of Data:**

* Relational database (MSSQL Server, ORACLE, MySQL, IBM DB2)
* Flat files and XML Datasets (Spreadsheet, XML)
* APIs and Web services
* Web scraping (Beautiful Soup, Scrapy, Pandas, Selenium)
* Data streams and feeds (IoT devices, GPS data from cars, computer programs website and social media posts)

**APIs and Web Services**:

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.

**Popular examples of APIs** such as Twitter and Facebook APIs for customer sentiment analysis

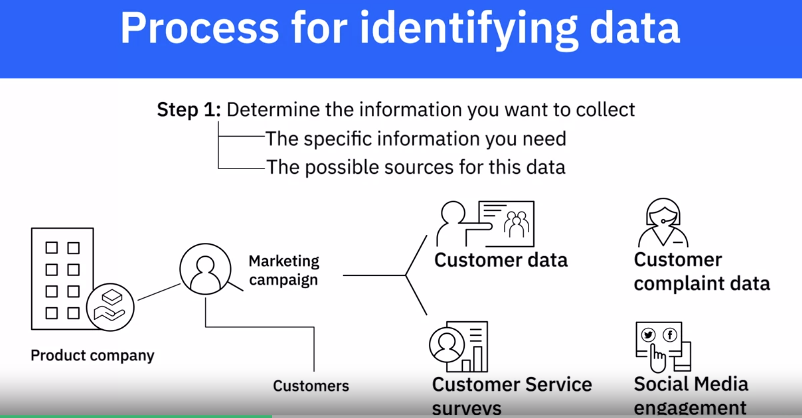
**Stock Market APIs** for trading and analysis

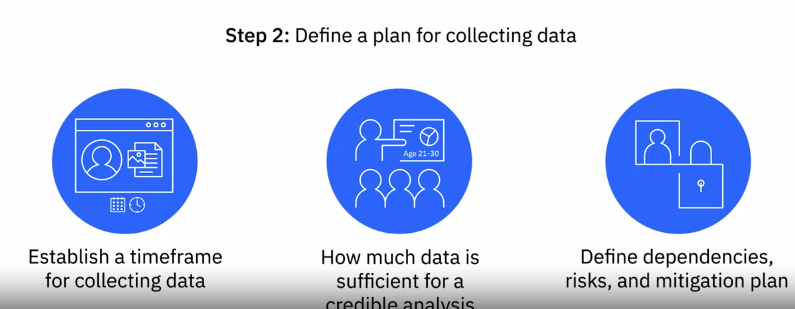
**Data Lookup and Validation APIs** for cleaning and co-relating data

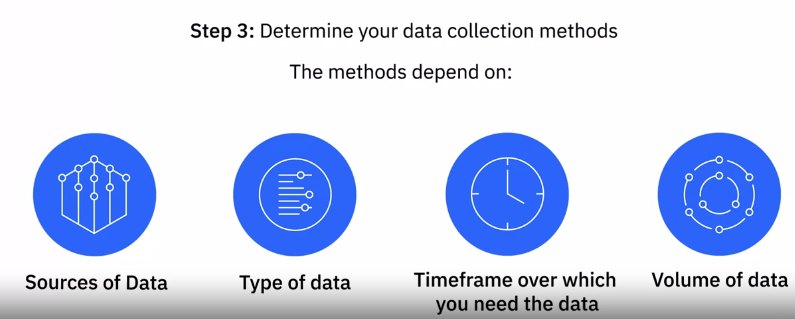
**Gathering and Wrangling Data**

We will learn about the process and steps involved in identifying, gathering, and importing data from disparate sources. We will learn about the tasks involved in wrangling and cleaning data in order to make it ready for analysis. In addition, we will gain an understanding of the different tools that can be used for gathering, importing, wrangling, and cleaning data, along with some of their characteristics, strengths, limitations, and applications.

**Identifying Data Analysis**:







**Summary:**

Having identified the data, your next step is to identify the sources from which you will extract the required data and define a plan for data collection. Decisions regarding the timeframe over which you need your data set, and how much data would suffice for arriving at a credible analysis also weigh in at this stage.

Data Sources can be internal or external to the organization, and they can be primary, secondary, or third-party, depending on whether you are obtaining the data directly from the original source, retrieving it from externally available data sources, or purchasing it from data aggregators.

Some of the data sources from which you could be gathering data include databases, the web, social media, interactive platforms, sensor devices, data exchanges, surveys and observation studies.

Data that has been identified and gathered from the various data sources is combined using a variety of tools and methods to provide a single interface using which data can be queried and manipulated.

The data you identify, the source of that data, and the practices you employ for gathering the data have implications for quality, security, and privacy, which need to be considered at this stage.

**How to Gather and Import Data**:

**Specific data repositories are optimized for certain types of data.**

**Structured data**

* Relational databases store structured data with a well-defined schema
* Sources include data from OLTP system, spreadsheets, online forms, sensors, network and web logs
* Can also be stored in NoSQL databases

**Semi-structured data**

* Sources include emails, XML, zipped files, binary executables, and TCP/IP protocols
* Can be stored in NoSQL clusters
* XML and JSON are commonly used for storing and exchanging semi-structured data

**Unstructured data:**

Unstructured data is data that does not have a structure and cannot be organized into a schema, such as data from web pages, social media feeds, images, videos, documents, media logs, and surveys. NoSQL databases and Data Lakes provide a good option to store and manipulate large volumes of unstructured data. Data lakes can accommodate all data types and schema. ETL tools and data pipelines provide automated functions that facilitate the process of importing data. Tools such as Talend and Informatica, and programming languages such as Python and R, and their libraries, are widely used for importing data.