Data Analytics Basics for Everyone

General Information

Course Overview

In this course, you will learn about the various components of a modern data ecosystem and the role Data Analysts, Data Scientists, and Data Engineers play in this ecosystem. You will gain an understanding of data structures, file formats, sources of data, and data repositories. You will understand what Big Data is and the features and uses of some of the Big Data processing tools.

This course will introduce you to the key tasks a Data Analyst performs in a typical day, and includes how a Data Analyst identifies, gathers, wrangles, mines and analyzes data, and finally communicates their findings to different stakeholders in an impactful way. Throughout the course you will be introduced to some of the tools Data Analysts use for each of these tasks.

By the end of this course you will know about the various career opportunities available in the field of Data Analytics, and the different learning paths you can consider to gain entry into this field.

Who Should Take This Course

This introductory Data Analytics course is designed for anyone who wants to learn about Data Analytics. If you are beginning your Data Analytics journey, this is a great place to start and this course will give you a dynamic skill set.

Pre-requisite

This program does not require any pre-requisites, and is suitable for learners with or without college degrees. All you need to get started is basic computer literacy, a comfort working with numbers, a willingness to self-learn online, and the desire to enrich your profile with valuable skills.

Course Syllabus

Module 1: Modern Data Ecosystem and the Role of Data Analytics

- Modern Data Ecosystem
- Key Players in the Data Ecosystem
- Defining Data Analysis
- Data Analytics vs. Data Analysis

Module 2: The Data Analyst Role

- Responsibilities of a Data Analyst
- A Day in the Life of a Data Analyst
- Introduction to Kubernetes Objects

Module 3: The Data Ecosystem and Languages for Data Professionals

- Overview of the Data Analyst Ecosystem
- Types of Data
- Understanding Different Types of File Formats
- Sources of Data Using Service Bindings
- Languages for Data Professionals

Module 4 - Understanding Data Repositories and Big Data Platforms

- RDBMS
- NoSQL
- Data Marts, Data Lakes, ETL, and Data Pipelines
- Foundations of Big Data
- Big Data Processing Tools

Module 5: Gathering Data

- Identifying Data for Analysis
- Data Sources
- How to Gather and Import Data

Module 6: Wrangling Data

- What is Data Wrangling?
- Tools for Data Wrangling
- Data Cleaning

Module 7: Analyzing and Mining Data

- Overview of Statistical Analysis

- What is Data Mining?
- Tools for Data Mining

Module 8: Communicating Data Analysis Findings

- Overview of Communicating and Sharing Data Analysis Findings
- Introduction to Data Visualization
- Introduction to Visualization and Dashboarding Software

Module 9: Opportunities and Learning Paths

- Career Opportunities in Data Analysis
- The Many Paths to Data Analysis

Final Assignments

GRADING SCHEME

This section contains information for those earning a certificate. Those auditing the course can skip this section and click next.

1. This course contains 9 Graded Quizzes, 1 Final Quiz and 1 Final Assignment. There is 1 Graded Quiz per module. Your total grade at 100% is weighted as follows:
   - Each of the 9 Graded Quizzes carries an equal weight totaling 80% of your total grade.
   - Final Quiz carries a weight of 10% of your total grade.
   - Final Assignment carries a weight of 10% of your total grade.
2. The minimum passing mark for the **course** is 70%.
3. Permitted attempts are per **question**:
   - One attempt - For True/False questions
   - Two attempts - For any question other than True/False
4. There are no penalties for incorrect attempts.
5. Clicking the "**Final Check**" button when it appears, means your submission is **FINAL**. You will **NOT** be able to resubmit your answer for that question again.
6. Check your grades in the course at any time by clicking on the "Progress" tab.

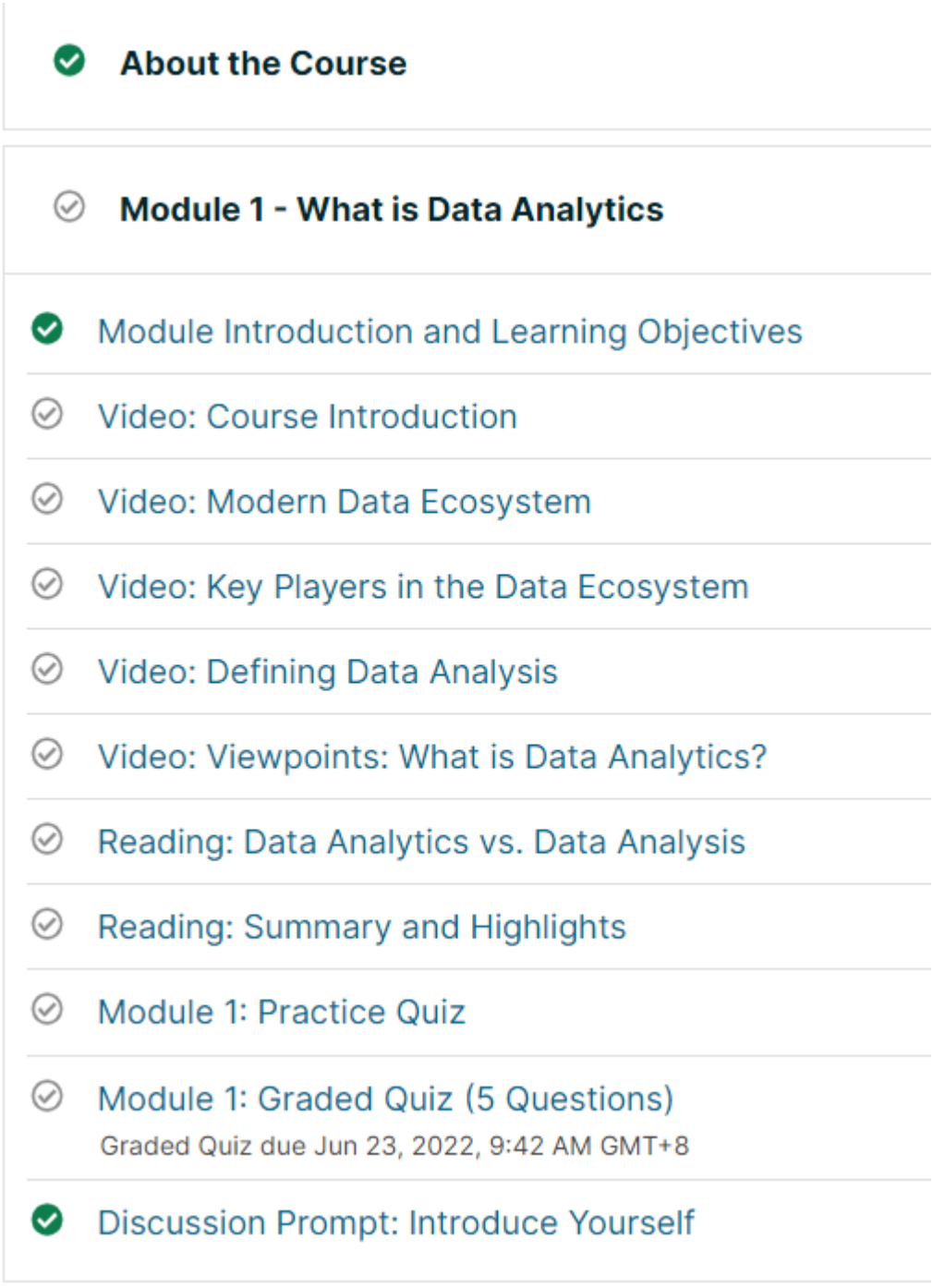
Module Introduction

In this module, you will gain an understanding of the different types of data analytics and the key steps in the overall data analytics process. You will learn about the different components of a modern data ecosystem, and the role Data Engineers, Data Analysts, Data Scientists, Business Analysts, and Business Intelligence Analysts play in this ecosystem.

Learning Objectives

After completing this module, you will be able to:

- Explain the different components of a modern data ecosystem.
- Describe and differentiate between the role different data professionals play in this ecosystem.
- Explain what data analytics is, the different types of data analytics, and the key steps in the data analytics process.



A Career in Data Analytics

"Businesses today recognize the untapped value in data and data analytics as a crucial factor for business competitiveness. To drive their data and analytics initiatives,companies are hiring and upskilling people. They're expanding their teams and creating centers of excellence to set up a multipronged data and analytics practice in their organizations."

-The Power Of Data To Transform Your Business, a Forrester report

Combined to this is the significant supply and demand mismatch in skilled data analysts, making it a highly sought after and well-paid profession.

Mastering data analytics as a career path

Branching into other data professions

Data Science

Business Analytics

Data Engineering

Business Intelligence Analytics

This job is for you if:

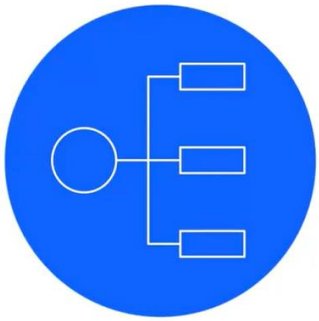
Fresh Graduate

Working Professional considering a mid-career transition

Data-driven Decision-maker

Analytics-enabled job role

The course introduces you to the core concepts, processes, and tools you need to



- Gain entry into data analytics
- Strengthen your current role as a data-driven decision-maker

The course will equip you with an understanding of

- Data ecosystem
- Fundamentals of data analysis, such as data gathering, wrangling, mining, analysis, and data visualization
- A day in the life of a Data Analyst

Practicing data analyst share

- Their experience in gaining entry into this field
- Career options and learning paths you can consider
- What employers look for in a data analyst
- Their knowledge and best practices about the data analysis process

Modern Data Ecosystem

To quote a Forbes 2020 report on data in the coming decade, "The constant increase in data processing speeds and bandwidth, the non stop invention of new tools for creating, sharing and consuming data, and the steady addition of new data creators and consumers around the world ensure that data growth continues unabated. Data begets more data in a constant virtuous cycle."

A modern data ecosystem includes a whole network of interconnected, independent, and continually evolving entities. It includes data that has to be integrated from disparate sources, different types of analysis and skills to generate insights, active stakeholders to collaborate and act on insights generated, and tools, applications, and infrastructure to store, process, and disseminate data as required.

Let's start with the data sources. Data is available in a variety of structured and unstructured datasets residing in text, images, videos, clickstreams, user conversations, social media platforms, the Internet of things, or IoT devices, real time events that stream data, legacy databases and data sourced from professional data providers and agencies.

The sources have never before been so diverse and dynamic. When you're working with so many different sources of data, the first step is to pull a copy of the data from the original sources into a data repository. At this stage, you're only looking at acquiring the data you need, working with data formats, sources, and interfaces through which this data can be pulled in.

Reliability, security and integrity of the data being acquired are some of the challenges you work through at this stage. Once the raw data is in a common place, it needs to get organized, cleaned up, and optimized for access by end users. The data will also need to conform to Compliances and Standards enforced in the Organization.

For example, conforming to guidelines that regulate the storage and use of personal data, such as health, biometrics or household data in the case of IoT devices. Adhering to master data tables within the organization to ensure standardization of master data across all applications and systems of an organization is another example.

The key challenges at this stage could involve data management and working with data repositories that provide high availability, flexibility, accessibility, and security.

Finally, we have our business stakeholders, applications, programmers, analysts, and data science use cases all pulling this data from the enterprise data repository. The key challenges at this stage could include the interfaces, APIs, and applications that can get this data to the end users in line with their specific needs.

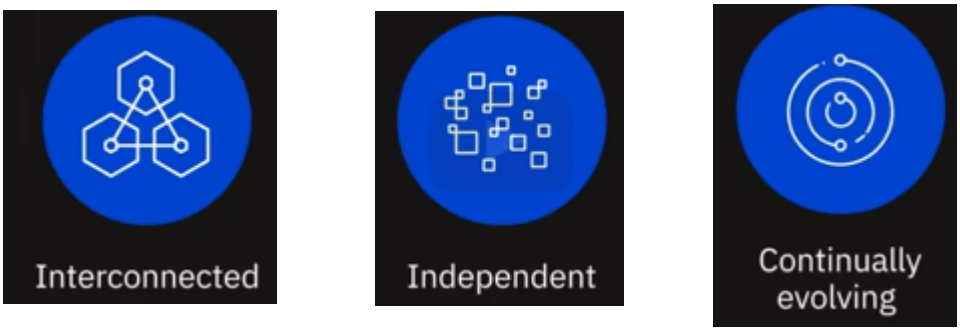
For example, data analysts may need the raw data to work with business. Stakeholders may need reports and dashboards. Applications may need custom API's to pull this data. It's important to note the influence of some of the new and emerging technologies that are shaping today's data ecosystem and its possibilities.

For example, cloud computing, machine learning, and big data to name a few. Thanks to cloud technologies, every enterprise today has access to limitless storage, high performance computing, open source technologies, machine learning technologies, and the latest

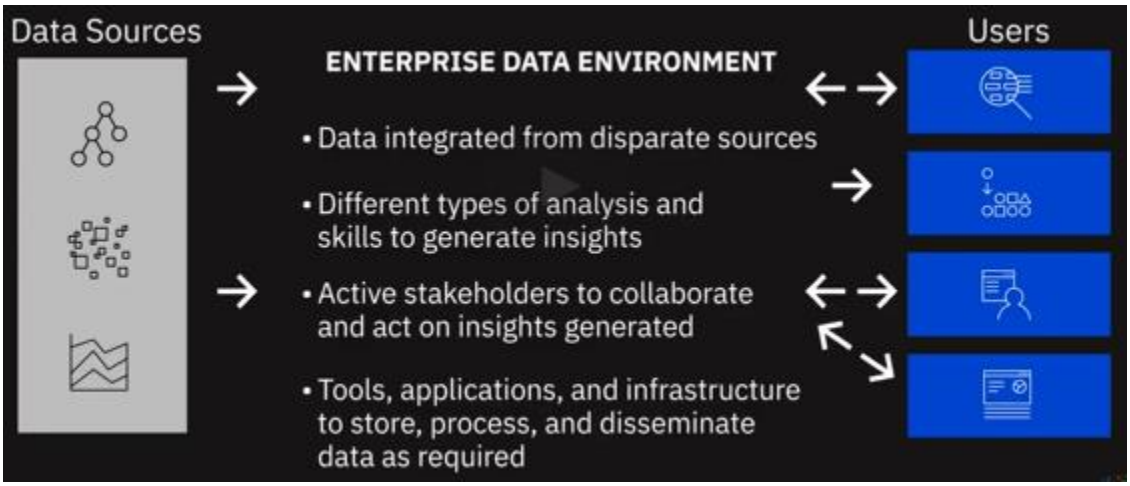
tools and libraries. Data scientists are creating predictive models by training machine learning algorithms on past data. Also, big data. Today, we're dealing with datasets that are so massive and so varied that traditional tools and analysis methods are no longer adequate, paving the way for new tools and techniques, and also new knowledge and insights.

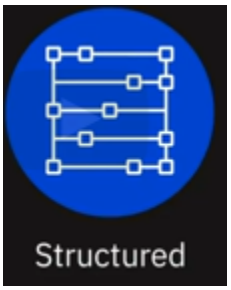
Modern Data Ecosystem

To quote a Forbes 2020 report on data in the coming decade, "The constant increase in data processing speeds and bandwidth, the non stop invention of new tools for creating, sharing and consuming data, and the steady addition of new data creators and consumers around the world ensure that data growth continues unabated. Data begets more data in a constant virtuous cycle."



A modern data ecosystem includes a whole network of interconnected, independent, and continually evolving entities.





Data is available in a variety of structured and unstructured datasets residing in text, images, videos, clickstreams, user conversations, social media platforms, the Internet of things, or IoT devices, real time events that stream data, legacy databases and data sourced from professional data providers and agencies.

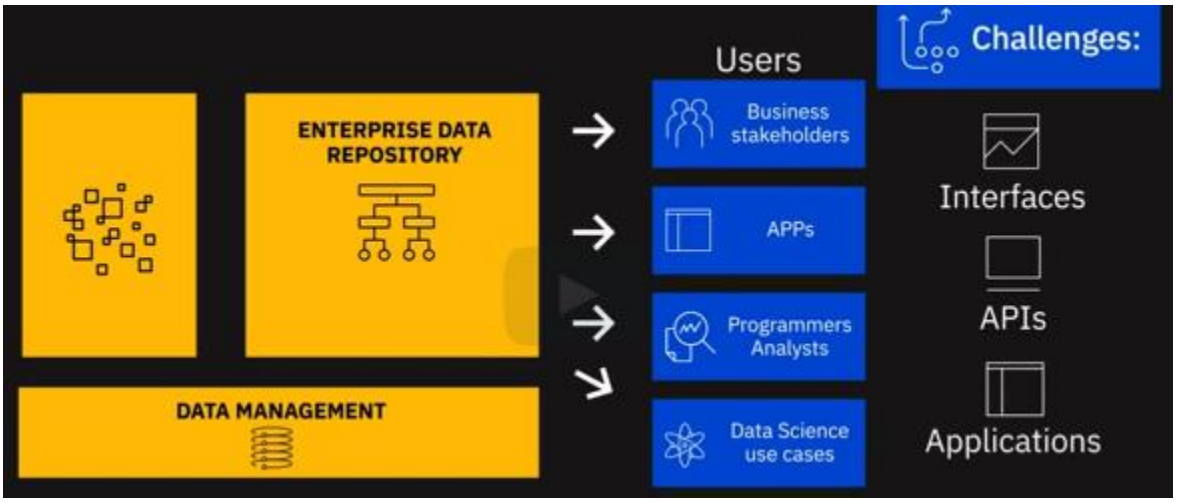
The first step is to pull a copy of the data from the original sources into a data repository

Acquiring the data you need, working with data formats, sources, and interfaces through which this data can be pulled.

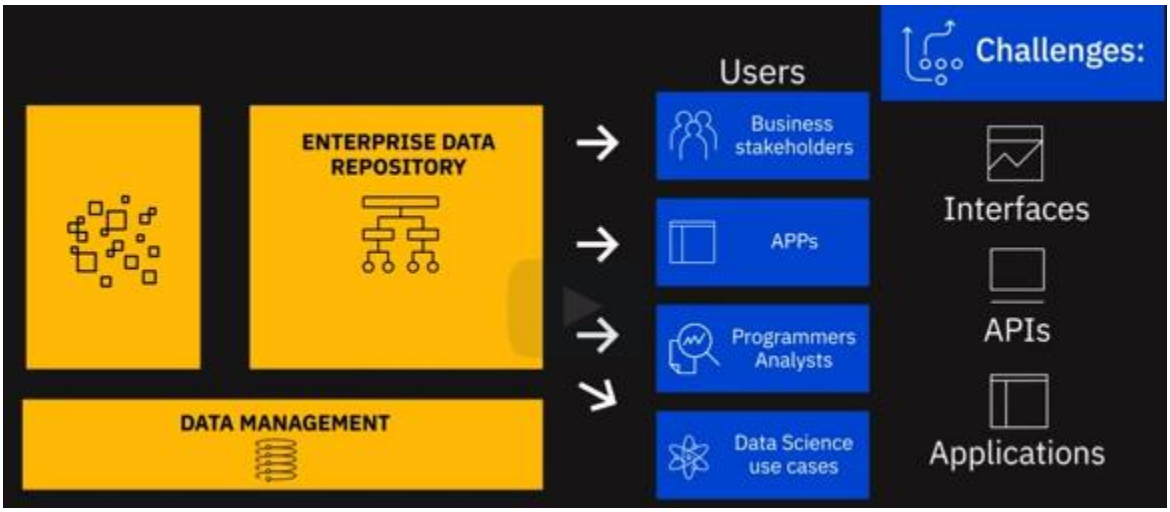
When you're working with so many different sources of data, the first step is to pull a copy of the data from the original sources into a data repository. At this stage, you're only looking at acquiring the data you need, working with data formats, sources, and interfaces through which this data can be pulled in. Reliability, security and integrity of the data being acquired are some of the challenges you work through at this stage.

At this stage, you're only looking at acquiring the data you need, working with data formats, sources, and interfaces through which this data can be pulled in. Reliability, security and integrity of the data being acquired are some of the challenges you work through at this stage. Once the raw data is in a common place, it needs to get organized, cleaned up, and optimized for access by end users. The data will also need to conform to Compliances and Standards enforced in the Organization.

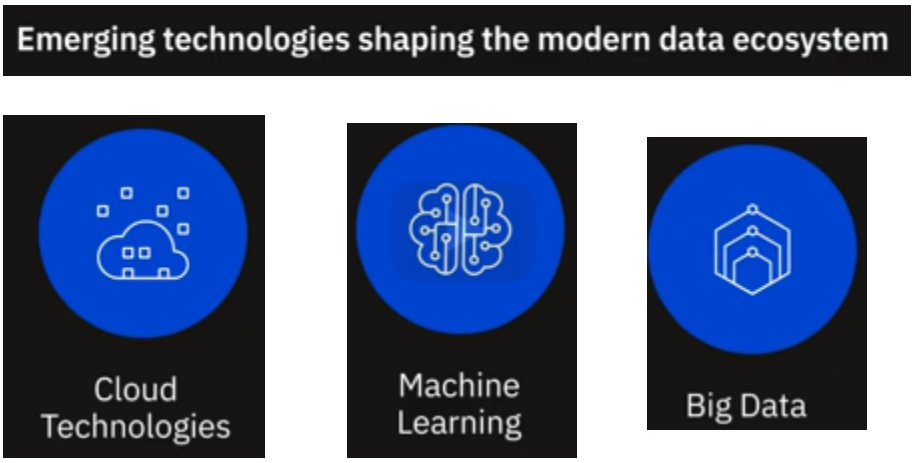
For example, conforming to guidelines that regulate the storage and use of personal data, such as health, biometrics or household data in the case of IoT devices. Adhering to master data tables within the organization to ensure standardization of master data across all applications and systems of an organization is another example. The key challenges at this stage could involve data management and working with data repositories that provide high availability, flexibility, accessibility, and security.



Finally, we have our business stakeholders, applications, programmers, analysts, and data science use cases all pulling this data from the enterprise data repository. The key challenges at this stage could include the interfaces, APIs, and applications that can get this data to the end users in line with their specific needs.



For example, data analysts may need the raw data to work with business. Stakeholders may need reports and dashboards. Applications may need custom API's to pull this data. The key challenges at this stage could include the interfaces, APIs, and applications that can get this data to the end users in line with their specific needs.



It's important to note the influence of some of the new and emerging technologies that are shaping today's data ecosystem and its possibilities. For example, cloud computing, machine learning, and big data to name a few.

- Every enterprise today has access to limitless storage, high performance computing, open source technologies, machine learning technologies, and the latest tools and libraries.
- Data scientists are creating predictive models by training machine learning algorithms on past data.
- Big data is paving the way for new tools and techniques and also new knowledge and insights.

Key Players in the Data Ecosystem

Today, organizations that are using data to uncover opportunities and are applying that knowledge to differentiate themselves are the ones leading into the future. Whether looking for patterns in financial transactions to detect fraud, using recommendation engines to drive conversion, mining, social media posts for customer voice or brands personalizing their offers based on customer behavior analysis, business leaders realized that data holds the key to competitive advantage.

To get value from data, you need a vast number of skill sets and people playing different roles. In this video, we're going to look at the role data engineers, data analysts, data scientists, business analysts, and business intelligence or BI analysts play in helping organizations tap into vast amounts of data and turn them into actionable insights.

It all starts with a data engineer.

- Data engineers are people who develop and maintain data architectures and make data available for business operations and analysis.
- Data engineers work within the data ecosystem to extract, integrate, and organize data from disparate sources. Clean transform and prepare data design, store and manage data in data repositories.
- They enabled data to be accessible in formats and systems that the various business applications as well as stakeholders like data analysts and data scientists can utilize.
- A data engineer must have good knowledge of programming, sound knowledge of systems and technology architectures, and in depth understanding of relational databases and non-relational data stores.

Now let's look at the role of a data analyst. In short,

- a data analyst translates data and numbers into plain language, so organizations can make decisions, data analysts inspect and clean data for deriving insights, identify correlations, find patterns, and apply statistical methods to analyze and mined data and visualize data to interpret and present the findings of data analysis.
- Analysts are the people who answer questions such as, Are the users search experiences generally good or bad with the search functionality on our site? or What is the popular perception of people regarding our rebranding initiatives? Or is there a correlation between sales, and one product and another?
- Data analysts require good knowledge of spreadsheets, writing queries, and using statistical tools to create charts and dashboards.
- Modern data analysts also need to have some programming skills. They also need strong analytical and storytelling skills. And now let's look at the role data scientists play in this ecosystem.

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.

Then we also have business analysts and BI analysts.

- Business analysts leverage the work of data analysts and data scientists to look at possible implications for their business and the actions they need to take or recommend.
- BI analysts do the same except. Their focus is on the market forces and external influences that shape their business.
- They provide business intelligent solutions by organizing and monitoring data on different business functions and exploring that data to extract insights and actionables that improve business performance.

To summarize, in simple terms,

- 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.

Interestingly, it's not uncommon for data professionals to start their career in one of the data roles and transition to another role within the data ecosystem by supplementing their skills.

Key Players in the Data Ecosystem

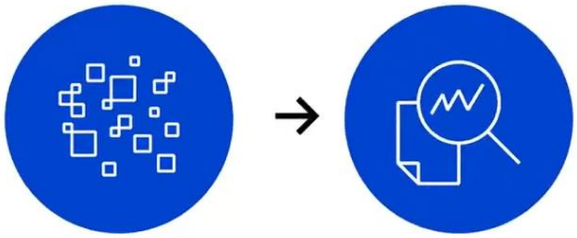
Overview

Organizations that are using data to uncover opportunities and are applying that knowledge to differentiate themselves are the ones leading into the future.

- Looking for patterns in financial transactions to detect fraud
- Using recommendation engines to drive conversion
- Mining, social media posts for customer voice
- Analyzing customer behavior for personalizing offers

Data Professionals:

- Data Engineers
- Data Analysts
- Data Scientists
- Business Analysts
- Business Intelligence Analysts



Data Engineers

Data Engineers work within the data ecosystem to:

- Extract, integrate, and organize data from disparate sources
- Clean, transform, and prepare data
- Design, store, and manage data in data repositories

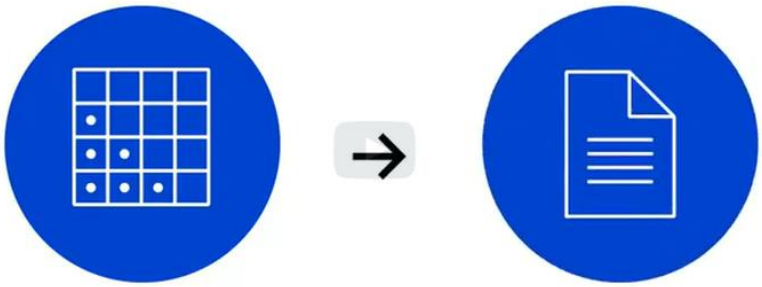
They enabled data to be accessible in formats and systems that the various business applications as well as stakeholders like data analysts and data scientists can utilize.

Skills:

- Good knowledge of programming
- Sound knowledge of systems and technology architectures
- In-depth understanding of relational databases and non-relational data stores

Data Analyst

In short, a data analyst translates data and numbers into plain language, so organizations can make decisions.



Responsibilities of a Data Analyst:

- Inspect and clean data for deriving insights
- Identify correlations, find patterns, and apply statistical methods to analyze and mine data
- Visualize data to interpret and present the findings of data analysis

People who answer questions such as:

"Are the users' search experiences generally good or bad with the search functionality on our site?"

"What is the popular perception of people regarding our rebranding initiatives?"

"Is there a co-relation between sales of one product and another?"

Skills:

- Good knowledge of spreadsheets, writing queries, and using statistical tools to create charts and dashboards
- Programming skills
- Strong analytical and story-telling skills

Data Scientist

Responsibilities of a Data Scientist:

- Analyze data for actionable insights
- Create predictive models using Machine Learning and Deep Learning

People who answer questions such as:

"How many new social media followers am I likely to get next month?"

"What percentage of my customers am I likely to lose to competition in the next quarter?"

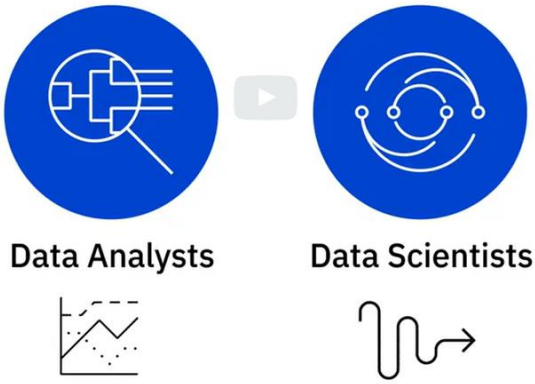
"Is this financial transaction unusual for this customer?"

Skills:

- Knowledge of Mathematics and Statistics
- Understanding of programming languages, databases, and building data models
- Domain knowledge

Business Analyst and BI Analyst

Business Analysts leverage the work of Data Analysts and Data Scientists to look at possible implications for their business and the actions they need to take or recommend.



BI Analysts

- Focus on market forces and external influences that shape their business
- Organize and monitor data on different business functions
- Explore data to extract insights and actionables that improve business performance

To summarize:

- 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.

Defining 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. We will explore four primary types of data analysis, each with a different goal and place in the data analysis process.

Descriptive Analytics helps answer questions about what happened over a given period of time by summarizing past data and presenting the findings to stakeholders. It helps provide essential insights into past events. For example, tracking past performance based on the organization's key performance indicators or cash flow analysis.

Diagnostic analytics helps answer the question. Why did it happen? It takes the insights from descriptive analytics to dig deeper to find the cause of the outcome. For example, a sudden change in traffic to a website without an obvious cause or an increase in sales in a region where there has been no change in marketing.

Predictive analytics helps answer the question, What will happen next? Historical data and trends are used to predict future outcomes. Some of the areas in which businesses apply predictive analysis are risk assessment and sales forecasts.

It's important to note that the purpose of predictive analytics is not. to say what will happen in the future, it's objective is to forecast what might happen in the future. All predictions are probabilistic in nature.

Prescriptive Analytics helps answer the question, What should be done about it? By analyzing past decisions and events, the likelihood of different outcomes. Is estimated on the basis of which a course of action is decided. Self-driving cars are a good example of Prescriptive Analytics. They analyze the environment to make decisions regarding speed, changing lanes, which route to take, etc. Or airlines automatically adjusting ticket prices based on customer demand. Gas prices, the weather or traffic on connecting routes.

Now let's look at some of the **Key steps in any data analysis process**.

1. **Understanding the problem and desired result**. Data analysis begins with understanding the problem that needs to be solved and the desired outcome that needs to be achieved. Where you are and where you want to be needs to be clearly defined before the analysis process can begin.

Setting a clear metric. This stage of the process includes deciding what will be measured. For example, number of product X sold in a region and how it will be measured, for example. In a quarter or during a festival season, gathering data once you know what you're going to measure and how you're going to measure it, you identify the data you require, the data sources you need to pull this data from, and the best tools for the job.

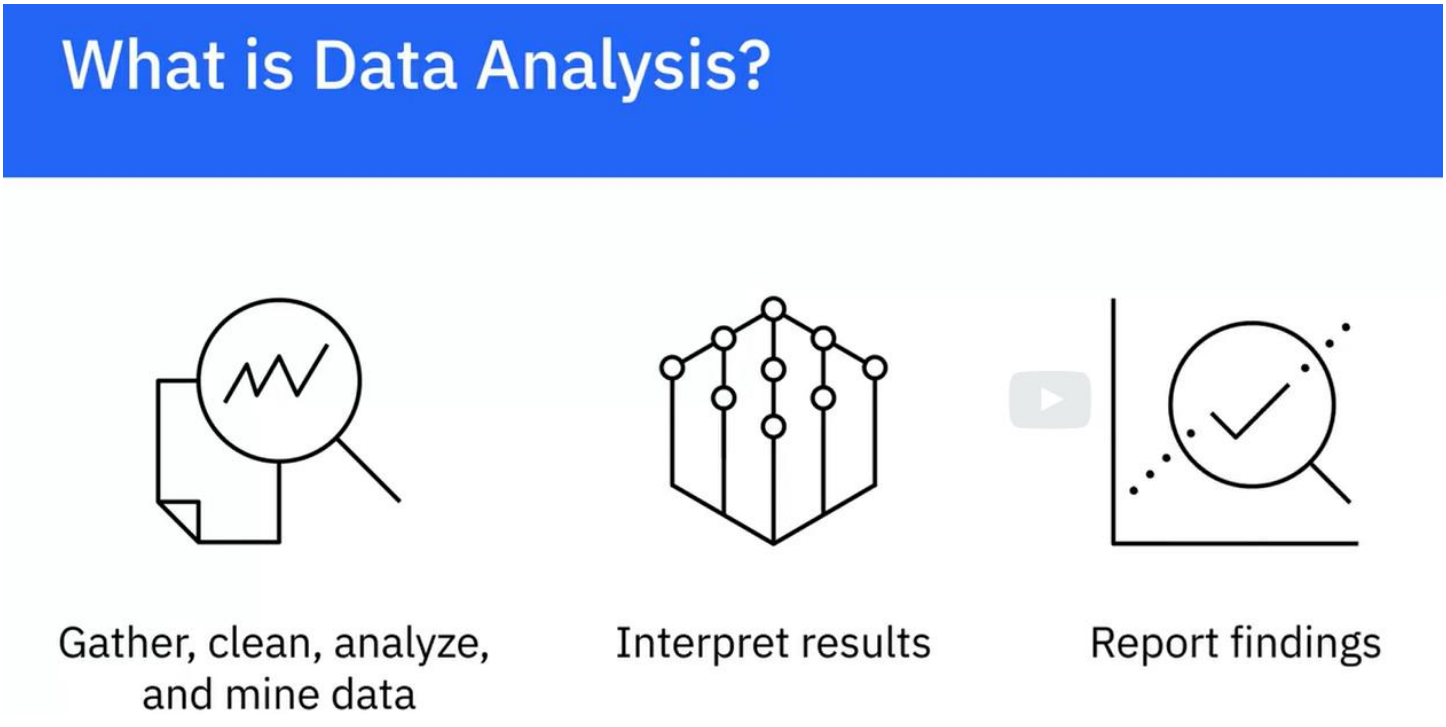
2.Cleaning data. Having gathered the data, the next step is to fix quality issues in the data that could affect the accuracy of the analysis. This is a critical step because the accuracy of the analysis can only be ensured if the data is clean. You will clean the data for missing or incomplete values and outliers. For example, a customer demographics data in which the age field has a value of 150 is an outlier. You will also standardize the data coming in from multiple sources.

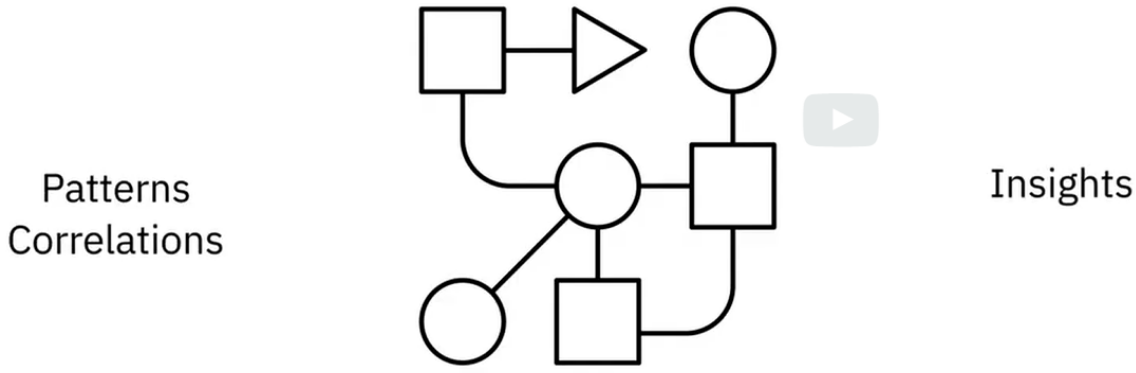
3.Analyzing and mining data. Once the data is clean, you will extract and analyze the data from different perspectives. You may need to manipulate your data in several different ways to understand the trends, identify correlations and find patterns and variations. Interpreting results.

After analyzing your data and possibly conducting further research, which can be an iterative loop, it's time to

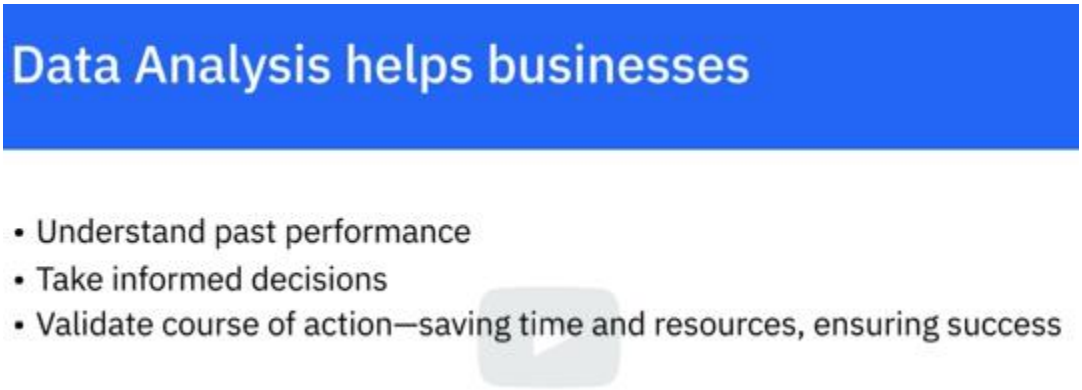
4.interpret your results. As you interpret your results, you need to evaluate if your analysis is defendable against objections, and if there are any limitations or circumstances under which your analysis may not hold true.

5.Presenting your findings. Ultimately, the goal of any analysis is to impact decision making. The ability to communicate and present your findings in clear and impactful ways is as important a part of the data analysis process as is the analysis itself. Reports, dashboards, charts, graphs, maps, case studies are just some of the ways in which you can present your data.





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.



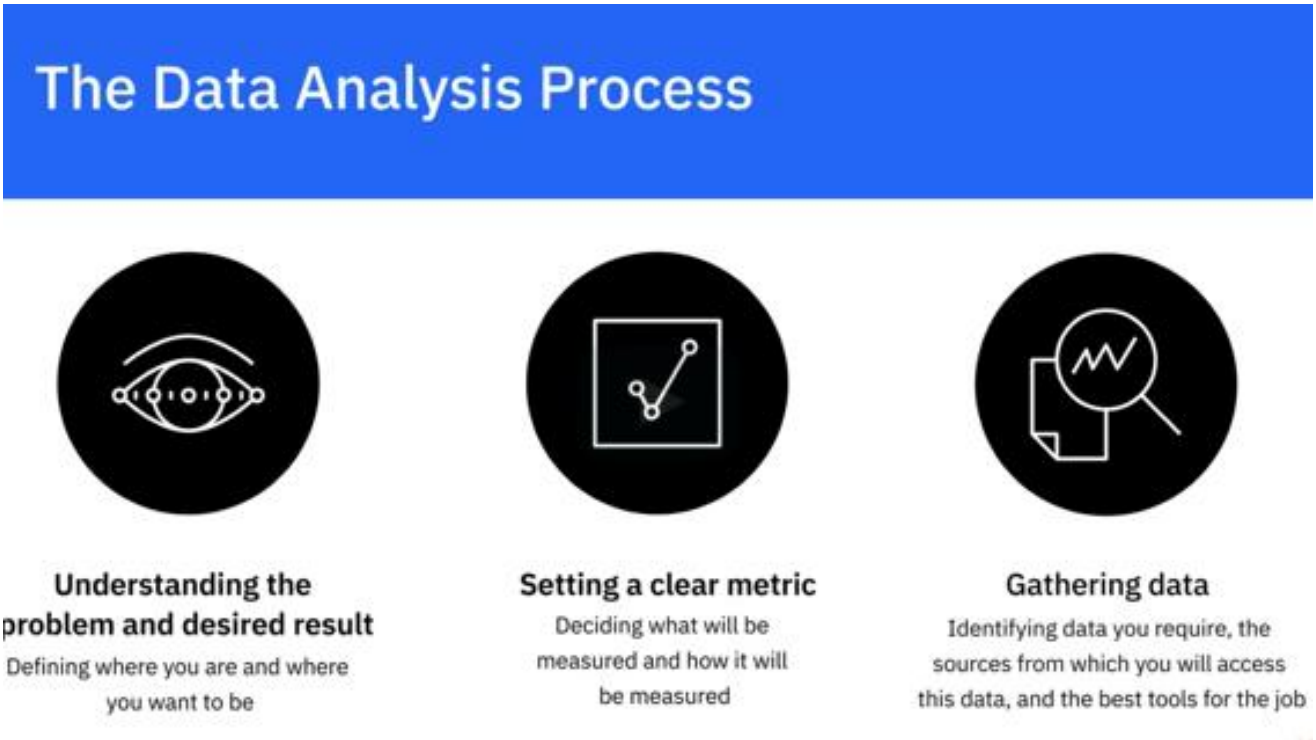
Different types of Data Analysis

Descriptive Analytics helps answer questions about what happened over a given period of time by summarizing past data and presenting the findings to stakeholders. It helps provide essential insights into past events. For example, tracking past performance based on the organization's key performance indicators or cash flow analysis.

Diagnostic analytics helps answer the question. Why did it happen? It takes the insights from descriptive analytics to dig deeper to find the cause of the outcome. For example, a sudden change in traffic to a website without an obvious cause or an increase in sales in a region where there has been no change in marketing.

Predictive analytics helps answer the question, What will happen next? Historical data and trends are used to predict future outcomes. Some of the areas in which businesses apply predictive analysis are risk assessment and sales forecasts. It's important to note that the purpose of predictive analytics is not. to say what will happen in the future, it's objective is to forecast what might happen in the future. All predictions are probabilistic in nature.

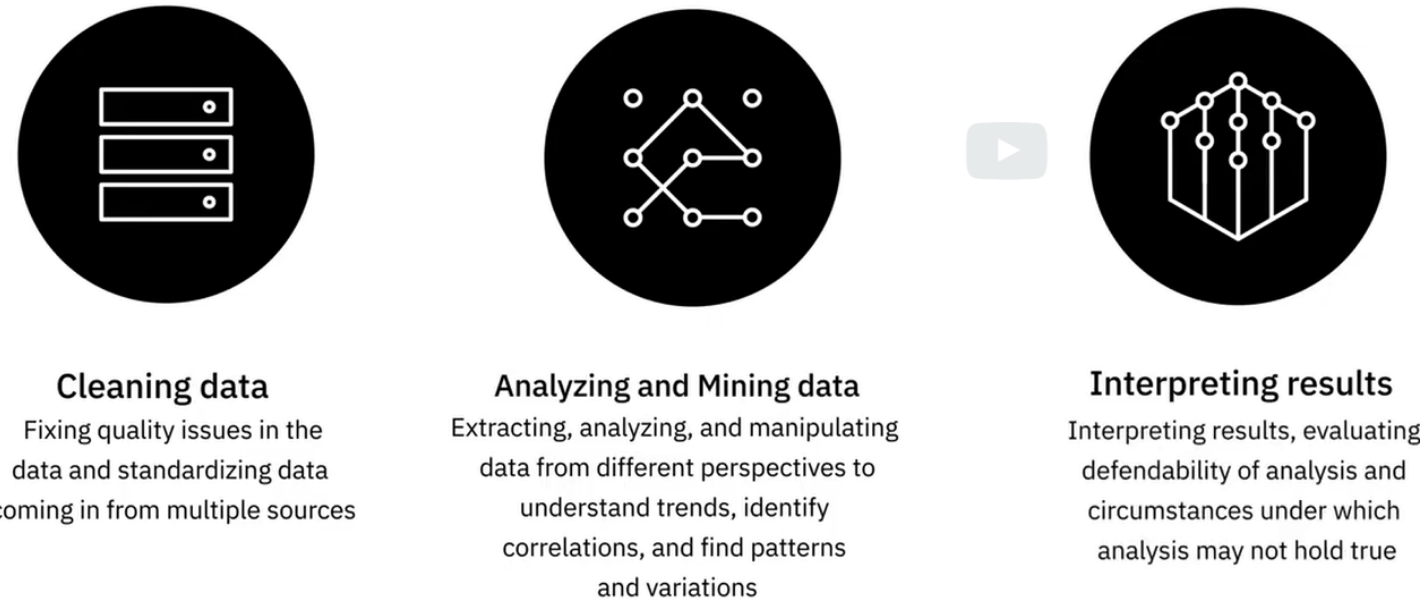
Prescriptive Analytics helps answer the question, What should be done about it? By analyzing past decisions and events, the likelihood of different outcomes. Is estimated on the basis of which a course of action is decided. Self-driving cars are a good example of Prescriptive Analytics. They analyze the environment to make decisions regarding speed, changing lanes, which route to take, etc. Or airlines automatically adjusting ticket prices based on customer demand. Gas prices, the weather or traffic on connecting routes.



1. **Understanding the problem and desired result**. Data analysis begins with understanding the problem that needs to be solved and the desired outcome that needs to be achieved. Where you are and where you want to be needs to be clearly defined before the analysis process can begin.

2. **Setting a clear metric**. This stage of the process includes deciding what will be measured. For example, number of product X sold in a region and how it will be measured, for example. In a quarter or during a festival season,

3. **gathering data** once you know what you're going to measure and how you're going to measure it, you identify the data you require, the data sources you need to pull this data from, and the best tools for the job.



4.**Cleaning data**. Having gathered the data, the next step is to fix quality issues in the data that could affect the accuracy of the analysis. This is a critical step because the accuracy of the analysis can only be ensured if the data is clean. You will clean the data for missing or incomplete values and outliers. For example, a customer demographics data in which the age field has a value of 150 is an outlier. You will also standardize the data coming in from multiple sources.

5.**Analyzing and mining data**. Once the data is clean, you will extract and analyze the data from different perspectives. You may need to manipulate your data in several different ways to understand the trends, identify correlations and find patterns and variations. Interpreting results.

After analyzing your data and possibly conducting further research, which can be an iterative loop, it's time to

6.**interpret your results**. As you interpret your results, you need to evaluate if your analysis is defendable

against objections, and if there are any limitations or circumstances under which your analysis may not hold true.

7.Presenting your findings. Ultimately, the goal of any analysis is to impact decision making. The ability to communicate and present your findings in clear and impactful ways is as important a part of the data analysis process as is the analysis itself. Reports, dashboards, charts, graphs, maps, case studies are just some of the ways in which you can present your data.

What is Data Analytics?

define data analytics as the process of collecting information and then analyzing that information to confirm various hypothesis. To me, data analytics also means storytelling with data. Using data to clearly and concisely convey the state of the world to the people around you.

Data analysis is the use of information around you to make decisions. Just like you get up every morning, you watch the news. The weather report will tell you the temperature for the day, whether it's going to rain. That may dictate what you're going to wear or what activities you can do. Data analysis isn't an abstract concept, it's something that we do naturally, but it has a technical name and now people are being paid to do it in a much larger or grander experience. But really, it's not that complicated.

The way I put it is that you've got a problem and you need to use facts to test a hypothesis, that's where data analytics comes into play. The process starts from defining the problem and then you need to create your own hypothesis. To test that, you need to collect data, clean data, analyze data, and then present it to the key stakeholders.

Data analytics is really any sets of data that you can use to review information, anything that's going to help you to understand what is going on. In my case as a CPA, I am always looking at financial state. I'm always analyzing data to predict where someone's been, where they are right now, and where they're headed. That data helps me to see further and almost predict the future of any company that I'm working with.

Data analytics is the collecting, cleansing, analyzing, presenting, and ultimately sharing of data and your analysis to be able to help communicate exactly what's going on with your business, what's going on in the data so that you can help make better decisions.

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

What is Data Analytics?

- Collecting nformation
- Analyzing Data
- Confirming Hypothesis
- Story telling with data
- Use of information to make decisions
- Defining the problem
- Create your own hypothesis.
- To test that, you need to
- Collect data, clean data,
- analyze data, and
- present it to the key stakeholders

- Analyzing sets of data to understand what's going on
- Understanding where business is coming from, it's present and future direction
- Analyzing and presenting and sharing your data
- Communicating business insights
- Helping make better decisions
- Process
- Taking information gathered from a population
- Breaking the information into subsets
- Using data to make decisions

Reading: Data Analytics vs. Data Analysis

The terms Data Analysis and Data Analytics are often used interchangeably, including in this course.

However it is important to note that there is a subtle difference between the terms and meaning of the words Analysis and Analytics. In fact some people go far as saying that these terms mean different things and should not be used interchangeably. Yes, there is a technical difference...

The dictionary meanings are:

Analysis - detailed examination of the elements or structure of something

Analytics - the systematic computational analysis of data or statistics

Analysis can be done without numbers or data, such as business analysis psycho analysis, etc. Whereas Analytics, even when used without the prefix "Data", almost invariably implies use of data for perfoming numerical manipulation and inference.

Some experts even say that Data Analysis is based on inferences based on historical data whereas Data Analytics is for predicting future performance. The design team of this course does not subscribe to this view, and you will see why later in the course as you become familiar with the terms like predictive analytics, prescriptive analytics, etc.

So in this course we take a more liberal view, and use the terms Data Analysis and Data Analytics to mean the same thing. For example, an earlier video is titled Defining Data Analysis, whereas the preceeding video with the viewpoints of several data professionals is titled What is Data Analytics. The difference in these titles is not intentional.

Summary and Highlights

In this lesson, you have learned the following information:

A modern data ecosystem includes a network of interconnected and continually evolving entities that include:

- Data that is available in a host of different formats, structure, and sources.

- Enterprise Data Environment in which raw data is staged so it can be organized, cleaned, and optimized for use by end-users.
- End-users such as business stakeholders, analysts, and programmers who consume data for various purposes.

Emerging technologies such as Cloud Computing, Machine Learning, and Big Data, are continually reshaping the data ecosystem and the possibilities it offers. Data Engineers, Data Analysts, Data Scientists, Business Analysts, and Business Intelligence Analysts, all play a vital role in the ecosystem for deriving insights and business results from data.

Based on the goals and outcomes that need to be achieved, there are four primary types of Data Analysis:

- 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.”

The Data Analysis process involves:

- Developing an understanding of the problem and the desired outcome.
- Setting a clear metric for evaluating outcomes.
- Gathering, cleaning, analyzing, and mining data to interpret results.
- Communicating the findings in ways that impact decision-making.

Module 1: Practice Quiz

Bookmark this page

Question 1

1/1 point (ungraded)

Which emerging technology has made it possible for every enterprise to have access to limitless storage and high-performance computing?

- ○ Internet of Things
- ○ Machine learning
- ○ Big Data
- ◉ Cloud computing

✔

Question 2

1/1 point (ungraded)

Which of the data roles is responsible for extracting, integrating, and organizing data into data repositories?

- ○ Business Intelligence Analyst
- ○ Data Analyst
- ○ Data Scientist
- ◉ Data Engineer

✔

Question 3

1/1 point (ungraded)

When you analyze historical data to predict future outcomes what type of Data Analytics are you performing?

- ○ Prescriptive Analytics
- ○ Descriptive Analytics
- ○ Diagnostic Analytics
- ◉ Predictive Analytics

✔

Module 1: Graded Quiz

Bookmarked

Graded Quiz due Jun 23, 2022 09:42 +08

Question 1

1/1 point (graded)

A modern data ecosystem includes a network of continually evolving entities. It includes:

- ○ Social media sources, data repositories, and APIs
- ○ Data sources, databases, and programming languages
- ◉ Data sources, enterprise data repository, business stakeholders, and tools, applications, and infrastructure to manage data
- ○ Data providers, databases, and programming languages

✔

Data Analysts work within the data ecosystem to:

- ◉ Gather, clean, mine, and analyze data for deriving insights
- ○ Develop and maintain data architectures
- ○ Provide business intelligence solutions by monitoring data on different business functions
- ○ Build Machine Learning or Deep Learning models

✔

Question 3

1/1 point (graded)

When we analyze data in order to understand why an event took place, which of the four types of data analytics are we performing?

- (●) Diagnostic Analysis
- ( ) Predictive Analysis
- ( ) Prescriptive Analysis
- ( ) Descriptive Analysis

✔

Question 4

1/1 point (graded)

The first step in the data analysis process is to gain an in-depth understanding of the problem and the desired outcome. What are you seeking answers to at this stage of the data analysis process?

- ( ) The data you need
- ( ) What will be measured and how it will be measured
- (●) Where you are and where you need to be
- ( ) The best tools for sourcing data

✔

Question 5

1/1 point (graded)

From the provided list, select the three emerging technologies that are shaping today's data ecosystem.

- ○ Cloud Computing, Internet of Things, and Dashboarding
- ◉ Cloud Computing, Machine Learning, and Big Data
- ○ Machine Language, Cloud Computing, and Internet of Things
- ○ Big Data, Internet of Things, and Dashboarding

✔

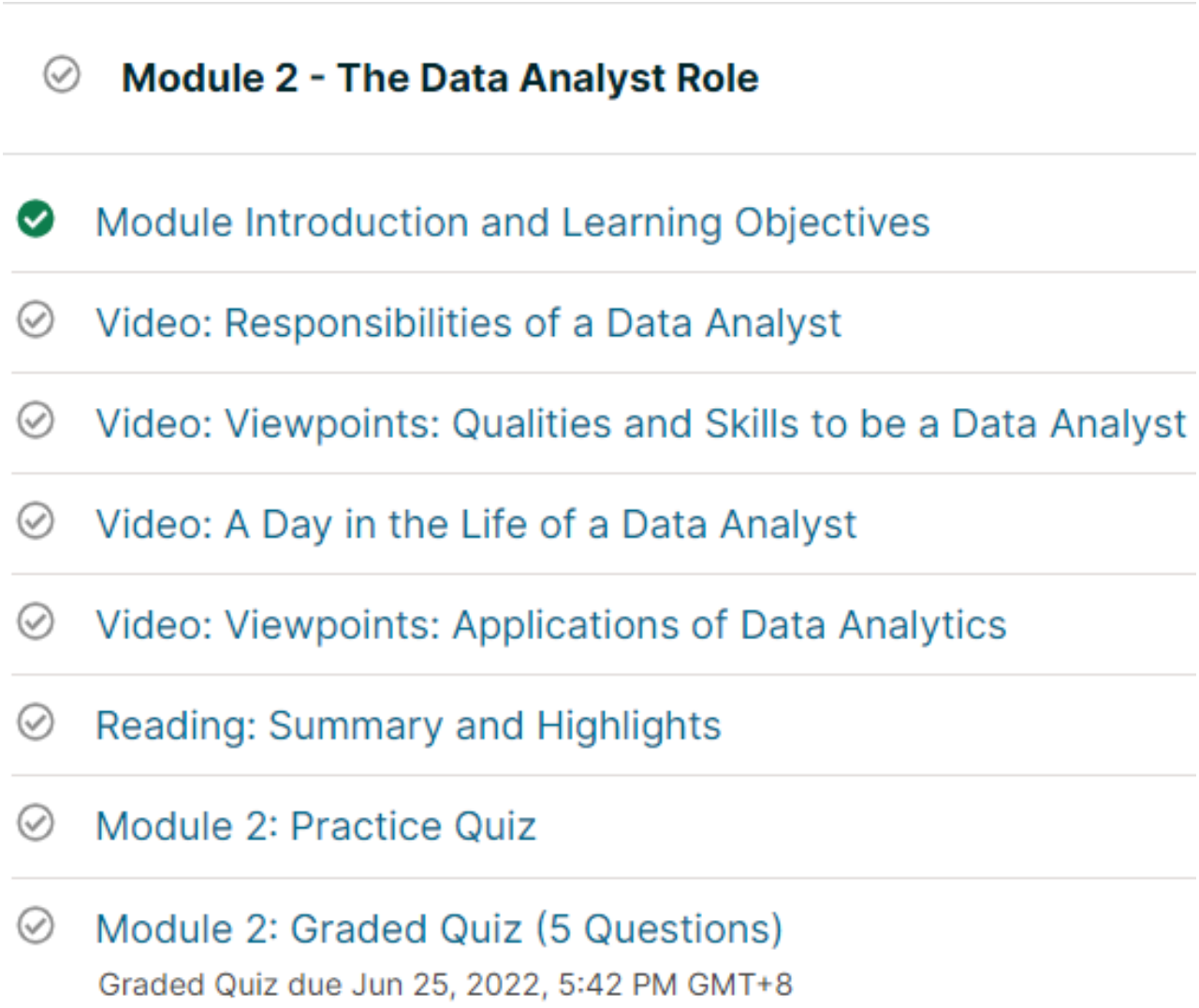
Module Introduction

In this module, you will learn about the role, responsibilities, and skillsets required to be a Data Analyst. You will gain an understanding of what a day in the life of a data analyst can look like. You will also hear from data professionals on some of the applications of data analytics in the real world.

Learning Objectives

After completing this module, you will be able to:

- Explain the responsibilities and skill sets of a Data Analyst.
- Describe what a day in life of a Data Analyst can look like.
- Describe some of the real-world applications of data analytics.



Responsibilities of a Data Analyst

While the role of a Data Analyst varies depending on the type of organization and the extent to which it has adopted data-driven practices, there are some responsibilities that are typical to a Data Analyst role in today's organizations. These include:

- 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.

Corresponding to these responsibilities, let's look at some of the skills that are valuable for a Data Analyst. The data analysis process requires a combination of technical, functional, and soft skills.

Let's first look at some of the technical skills that you need in your role as a Data Analyst.

These include:

- Expertise in using spreadsheets such as Microsoft Excel or Google Sheets,
- Proficiency in statistical analysis and visualization tools and software such as IBM Cognos, IBM SPSS, Oracle Visual Analyzer, Microsoft Power BI, SAS, and Tableau
- Proficiency in at least one of the programming languages such as R, Python, and in some cases C++, Java, and MATLAB,
- 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.

We will understand more about the features and use cases of some of these programming languages, databases, data repositories, and big data processing tools further along in the course.

Now we'll look at some of the functional skills that you require for the role of Data Analyst.

These include:

- Proficiency in Statistics to help you analyze your data, validate your analysis, and identify fallacies and logical errors.
- Analytical skills that help you research and interpret data, theorize, and make forecasts.
- Problem-solving skills, because ultimately, the end-goal of all data analysis is to solve problems.

- Probing skills that are essential for the discovery process, that is, for understanding a problem from the perspective of varied stakeholders and users—because the data analysis process really begins with a clear articulation of the problem statement and desired outcome.
- Data Visualization skills that help you decide on the techniques and tools that present your findings effectively based on your audience, type of data, context, and end-goal of your analysis.
- Project Management skills to manage the process, people, dependencies, and timelines of the initiative.

That brings us to your soft skills as a Data Analyst. Data Analysis is both a science and an art. You can ace the technical and functional expertise, but one of the key differentiators for your success is going to be soft skills.

This includes;

- your ability to work collaboratively with business and cross-functional teams;
- communicate effectively to report and present your findings;
- tell a compelling and convincing story; and
- gather support and buy-in for your work.
- Above all, being curious, is at the heart of data analysis.

In the course of your work, you will stumble upon patterns, phenomena, and anomalies that may show you a different path. The ability to allow new questions to surface and challenge your assumptions and hypotheses makes for a great analyst.

You will also hear data analysis practitioners talk about intuition as a must-have quality. It's essential to note that intuition, in this context, is the ability to have a sense of the future based on pattern recognition and past experiences.

In this video, we learned about the responsibilities and skillsets of a Data Analyst. In the next video, we will walk you through a day in the life of a Data Analyst.

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

Technical Skills:

Expertise in using spreadsheets
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, C++, Java, and MATLAB

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
Data Marts, Data Warehouses, Data Lakes, and Data Pipelines

Familiarity with Big Data processing tools
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:

Data Analysis

Science Art

You can ace the technical and functional expertise, but one of the key differentiators for your success is going to be your soft skills.

Your ability to:

- work collaboratively with business and cross-functional teams
- communicate effectively to report and present your findings
- tell a compelling and convincing story
- gather support and buy-in for your work

Curiosity
Allowing new questions to surface and challenging your own assumptions and hypotheses

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

Qualities and Skills to be a Data Analyst

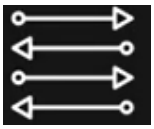
- Natural curiosity
- Attention to detail
- Enjoy working with computers
- Curiosity - Looking for answers even when there isn't a question
- Looking in areas that may not have been thought of before
- Attention to detail
    - Looking for patterns
    - Paying attention to close details
- Enjoying computers – developing skills to keep pace with technology
- Technical Skills and Soft Skills required
- Technical Skills
  Python, SQL, Tableau, and Power BI
- Soft Skills
  Decisions regarding the right data and the right tools to use
  Presenting data to key stakeholders
  Business acumen

Presentation skills

- Detail oriented
- Love numbers and information
- Not take things at face value
- An eye and mindset to keep to catch things that don’t look right
- Soft Skills
    - Curiosity
    - Thoughtfulness
    - Ability to listen carefully
    - Ability to understand both user and co-worker perspective
    - Willingness to learn
- SQL for extracting data
- Python and R programming langauges
- At least one data visualization tool
- Know what problem is to be solved
- Pull data in the required structure from data lake, using SQL
- Clean, Wrangle, Manipulate, and Mine data to glean insights
- Present insights clearly and using good visualizations and dashboard
- Tell a good story with that data
-

A Day in the life of a Data Analyst

Acquiring data from varied source



Creating queries for pulling data from data repositories,

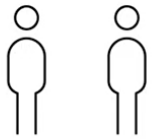
Looking for insights in data

Creating reports and dashboards

Interacting with stakeholders for gathering information

Cleaning and preparing data for analysis

Introduction



Fluentgrid is an IBM partner and the recipient of IBM Beacon awards for its solutions in the areas of smart energy and smart city industry segments.

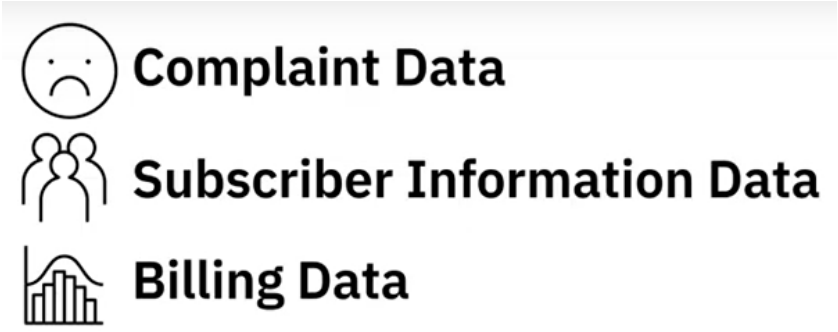
We offer integrated operations center solutions for power utilities and smart cities, leveraging our actionable intelligence platform known as Fluentgrid Actilligence.

Our client, a power utility company in South India, has been noticing a spike in complaints regarding overbilling.

The frequency of these complaints seems to suggest there's something more to it than random occurrences.

Working Hypothesis

Starting Point



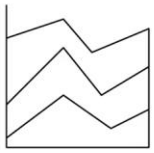
Initial Hypotheses:

1. Is there a consumption range for which overbilling is occurring more than others?
2. Are the complaints concentrated in specific localities within the city?
3. Are the same subscribers reporting overbilling repetitively?

Identifying Datasets

1.

Identify the datasets that I am going to isolate and analyze to validate or refute my hypotheses



The annual, quarterly, and monthly billing amounts of the complainants and look for a range in which the complaints are falling more than others

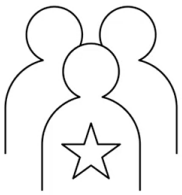
2.

Identify the datasets that I am going to isolate and analyze to validate or refute my hypotheses

Is there a connection between overbilling and zip codes?

3.

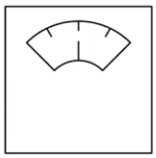
Identify the datasets that I am going to isolate and analyze to validate or refute my hypotheses



More than 95% of the complainants had been our subscribers for more than seven years.

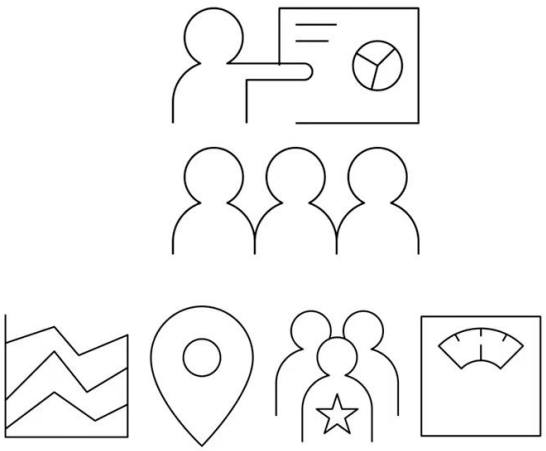
4.

Identify the datasets that I am going to isolate and analyze to validate or refute my hypotheses



The make and the serial number of the meters belong to a single supplier.

5. **Presentation to stakeholders**



A Day in the life of a Data Analyst

A day in the life of a Data Analyst can include a number of possibilities — from acquiring data from varied data sources to creating queries for pulling data from data repositories, foraging through rows of data to look for insights,creating reports and dashboards, and interacting with stakeholders for gathering information and presenting the findings, it's a spectrum.

And yes, the big one — cleaning and preparing the data so that the findings have a credible basis — which, by the way, is a large part of what any Data Analyst may find themselves doing in their jobs. But if I had to walk you through any one "type" of day, I'm going to pick one which has me foraging through data looking for insights. This is the part of my job that I am totally in awe of.

Hi. I'm Sivaram Jaladi. I work as a Data Analyst with Fluentgrid, a smart grid technology solutions company based in Vishakhapatnam in India.

Fluentgrid is an IBM partner and the recipient of IBM Beacon awards for its solutions in the areas of smart energy and smart city industry segments. We offer integrated operations center solutions for power utilities and smart cities, leveraging our actionable intelligence platform known as Fluentgrid Actilligence.

Our client, a power utility company in South India, has been noticing a spike in complaints regarding overbilling. And the frequency of these complaints seems to suggest there's something more to it than random occurrences. So, I'm asked to look at the complaints and the billing data and see if I can spot something.

I start by taking stock of what I have. Some of the obvious places that I know I'm going to be looking into is the complaint data, the subscriber information data, and the billing data. That's going to be my starting point. Before I dive into the specifics of the data, I'm going to make a list of questions, initial hypotheses, that I am going to start with. Such as the usage pattern of subscribers reporting this issue: Is there a consumption range for which overbilling is occurring more than others?

Area-wise concentration of complaints: Are the complaints concentrated in specific localities within the city? Frequency and occurrence of complaints based on individual subscribers: Are the same subscribers reporting overbilling repetitively? If yes, what is the frequency of occurrence in repeat cases? If a subscriber is overbilled once, does the overbilling occur every month from the first occurrence, or are repeat occurrences sporadic, or not at all?

As I get clear on my initial hypotheses and the set of questions I'm going to start with, I identify the datasets that I am going to isolate and analyze to validate or refute my hypotheses. I pull out the average annual, quarterly, and monthly billing amounts of the complainants and look for a range in which the complaints are falling more than others.

I then pull up the location data of the complainants to see if there is a connection between overbilling and zip codes. Here I see what seems to be a concentration of complaints in certain areas. This looked like it could add up to something. So instead of moving to the third hypothesis, I decide to get a little deeper into this data. Next, I pull out the date of connection data. More than 95% of the complainants had been our subscribers for more than seven years, though not all subscribers over the 7-year mark were facing this complaint.

So now, we see some area-wise concentration, and we see a significant concentration of complaints based on the date of connection.

Next, I pull out the make and the serial number of the meters. And there it is — the serial numbers belonged to the same batch of meters provided by the same supplier. The concentration of these meters, and therefore the complaints, was coming from areas in which these meters were installed.

At this stage, I feel confident in presenting these findings to the stakeholders. I'm also going to share the data sources and my process of arriving at this analysis — that always goes a long way in lending credibility to the findings. This could be the end of this project, or it may very well come back. Maybe the same complaints with different commonalities, or a completely different set of complaints for which we need to find answers.

Viewpoints: Applications of Data Analytics

In this video, practicing data professionals talk about some of the applications of data analytics in today's world. >> The applications of data analytics in the world today is everywhere.

Every commercial that you see, someone had to analyze and identify either from the consumer or for the company, what information they wanted to share. So you know four out of 10 dentists or you'll see information related to calorie counts or reactions to certain things, all of that required analysis.

This isn't something that should be thought of separate and apart from, it's what we do every day in our lives.

Even people monitoring their sugar level with diabetes, there's always analysis going on, so the applications are universal. >> So the great thing with analytics in this day and age is that it's very widely applicable.

Every industry, every vertical, every function within a given organization can benefit from data and analytics. Whether you're doing sales pipeline analysis, whether you're doing financials at the end of the month, creating predefined and standardized formatted reports.

Or if you're doing something like headcount planning or headcount review, all of these across every vertical, as I said, whether its airlines. Pharmaceuticals, banking all these and the functions within them can benefit from analytics. >> And in this climate that we're in right now with the pandemic, there are companies who are paying close attention to their customers buying habits.

Obviously they may have varied from what these companies expected these habits to be. And so now data analytics is more important because they need to make sure they can pivot and keep up with the demand. And really be able to cater to what their clients and their customers want.

>> I can talk about applications of data analytics in finance, this is we have seen more and more applications of alternative data analytics in the finance world. For example, we can use sentiment analysis of tweets and new stories to supplement traditional financial analysis and to inform better investment decisions. Besides the satellite imagery data can be used to track the development of industrial activities. And a geolocation data can be used to track the store traffic and to predict the sales volume.

What are some of the applications of Data aAnalytics in today's world?

- Application of data analytics is everywhere
- Used by companies to identify what information consumers want them to share
- Used by people with diabetes monitoring sugar levels
- Widely applicable across industries verticals, and functions within an organizations
- Sales pipeline analysis
- Financial reporting
- Headcount planning
- All verticals, such as airlines, pharmaceuticals and banking can benefit from it.
- Companies are using data analysis to pay attention to changes in their customer's buying habits
- Use of sentiment analysis of 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

Summary and Highlights

In this lesson, you have learned the following information:

The role of a Data Analyst spans across:

- Acquiring data that best serves the use case.
- Preparing and analyzing data to understand what it represents.
- Interpreting and effectively communicating the message to stakeholders who need to act on the findings.
- Ensuring that the process is documented for future reference and repeatability.

In order to play this role successfully, Data Analysts need a mix of technical, functional, and soft skills:

- Technical Skills include varying levels of proficiency in using spreadsheets, statistical tools, visualization tools, programming and querying languages, and the ability to work with different types of data repositories and big data platforms.

- An understanding of Statistics, Analytical techniques, problem-solving, the ability to probe a situation from multiple perspectives, data visualization, and project management skills – all of which come under Functional Skills a Data Analyst needs in order to play an effective role.

- Soft Skills include the ability to work collaboratively, communicate effectively, tell a compelling story with data, and garner support and buy-in from stakeholders. Curiosity to explore different pathways and intuition that helps to give a sense of the future based on past experiences are also essential skills for being a good Data Analyst.

Quiz: Practice Quiz

Bookmarked

Question 1

1/1 point (ungraded)

Which of these skills is essential to the role of a Data Analyst?

- ○ Machine learning
- ◉ Statistics
- ○ Big Data Engineering
- ○ Deep Learning models

✔

Submit

✔ Correct (1/1 point)

Question 2

1/1 point (ungraded)

What, according to Sivaram Jaladi, goes a long way in lending credibility to your data analysis findings?

- ○ Making sure the presentation looks good
- ◉ Sharing your process of arriving at the findings with your stakeholders
- ○ Writing good queries
- ○ Networking with your stakeholders

✔

Submit

Shov

✔ Correct (1/1 point)

Quiz: Graded Quiz

Bookmarked

Graded Quiz due Jun 25, 2022 17:42 +08

Question 1

1/1 point (graded)

Why is proficiency in Statistics an important skill for a Data Analyst?

- ○ For acquiring data from multiple sources
- ○ For creating project documentation
- ◉ For identifying patterns and correlations in data
- ○ For creating queries to extract required data

✔

Question 2

1/1 point (graded)

Which of these is one of the soft skills required to be a successful Data Analyst?

- ◉ Work collaboratively with cross-functional teams
- ○ Integrate data coming from multiple sources
- ○ Prepare reports and dashboards
- ○ Filter, clean, and standardize data

✔

Question 3

1/1 point (graded)

Which of the data analyst functional skills helps research and interpret data, theorize, and make forecasts?

- (●) Analytical skills
- ( ) Probing skills
- ( ) Proficiency in Statistics
- ( ) Problem-solving skills

✔

Question 4

1/1 point (graded)

In "A day in the life of a Data Analyst", what according to Sivaram Jaladi forms a large part of a Data Analyst's job?

- ( ) Interacting with stakeholders
- ( ) Creating a report
- ( ) Generating hypotheses
- (●) Cleaning and preparing data

✔

Question 5

1/1 point (graded)

In "A day in the life of a Data Analyst", what are some of the data points that were useful in analyzing the use case. (Select all that apply)

- [x] Average billing amount of complainants
- [ ] Employment history of the complainants
- [x] Serial number of the meters
- [ ] Age and education details of complainants

✔

Module 3 - The Data Ecosystem and Languages for Data Professionals

- Module Introduction and Learning Objectives
- Video: Overview of the Data Analyst Ecosystem
- Video: Types of Data
- Video: Understanding Different Types of File Formats
- Video: Sources of Data
- Video: Languages for Data Professionals
- Reading: Summary and Highlights
- Module 3: Practice Quiz
- Module 3: Graded Quiz (5 Questions)
  Graded Quiz due Jun 28, 2022, 1:42 AM GMT+8

Module Introduction

In this module, you will learn about the different components that make up a modern day data ecosystem. You will learn about the different types of data structures, file formats, and sources of data. You will also gain an understanding about the languages data professionals use in their day-to-day work.

Learning Objectives

After completing this module, you will be able to:

- Describe the different components of a modern data ecosystem.
- Explain the different types of data structures, file formats, and sources of data.
- Explain the features and use of the different languages used by data professionals.

Overview of the Data Analyst Ecosystem

A data analyst's ecosystem includes the infrastructure, software, tools, frameworks, and processes used to gather, clean, analyze, mine, and visualize data.

In this video, we will go over a quick overview of the ecosystem before going into the details of each of these topics in subsequent videos. Let's first talk about data.

Based on how well-defined the structure of the data is, data can be categorized as structured, semi-structured, or unstructured.

- Structured data follows a rigid format and can be organized neatly into rows and columns. This is the data that you see typically in databases and spreadsheets, for example.

- Semi-structured data is a mix of data that has consistent characteristics and data that doesn't conform to a rigid structure. For example, emails. An email has a mix of structured data, such as the name of the sender and recipient, but also has the contents of the email, which is unstructured data. And then

- there is unstructured data: Data that is complex, and mostly qualitative information that is impossible to reduce to rows and columns. For example, photos, videos, text files, PDFs, and social media content. The type of data drives the kind of data repositories that the data can be collected and stored in, and also the tools that can be used to query or process the data.

Data also comes in a wide-ranging variety of file formats being collected from a variety of data sources, ranging from relational and non-relational databases, to APIs, web services, data streams, social platforms, and sensor devices.

This brings us to data repositories:

A term that includes databases, data warehouses, data marts, data lakes, and big data stores. The type, format, and sources of data influence the type of data repositories that you can use to collect, store, clean, analyze, and mine the data for analysis.

If you're working with big data, for example, you will need big data warehouses, that allow you to store and process large-volume high-velocity data and also frameworks that allow you to perform complex analytics in real-time on big data.

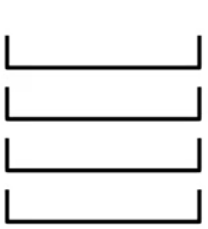
The ecosystem also includes languages that can be classified as query languages, programming languages, and shell and scripting languages. From querying and manipulating data with SQL to developing data applications with Python, and writing shell scripts for repetitive operational tasks, these are important components in a data analyst's workbench.

Automated tools, frameworks, and processes for all stages of the analytics process are part of the Data Analysts ecosystem. From tools used for gathering, extracting, transforming, and loading data into data repositories, to tools for data wrangling, data cleaning, data mining, analysis, and data visualization — it's a very diverse and rich ecosystem. Spreadsheets, Jupyter Notebooks, and IBM Cognos are just a few examples.

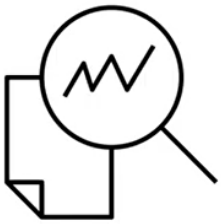
Overview of the Data Analyst Ecosystem

Overview

A Data Analyst's ecosystem includes the infrastructure, software, tools, frameworks, and processes used to

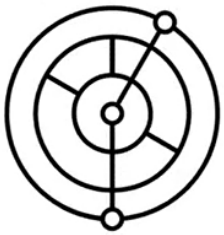


Gather Data



Clean Data

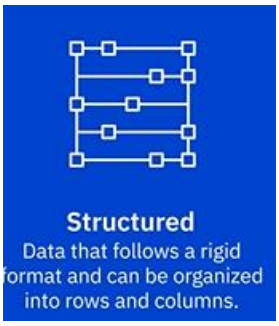
Mine Data



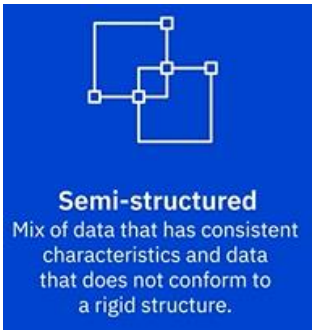
Visualize Data

Data

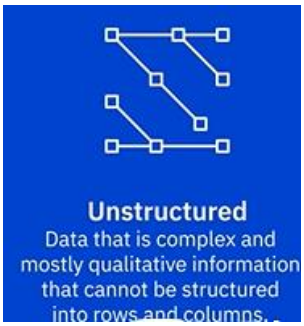
data can be categorized as structured, semi-structured, or unstructured.



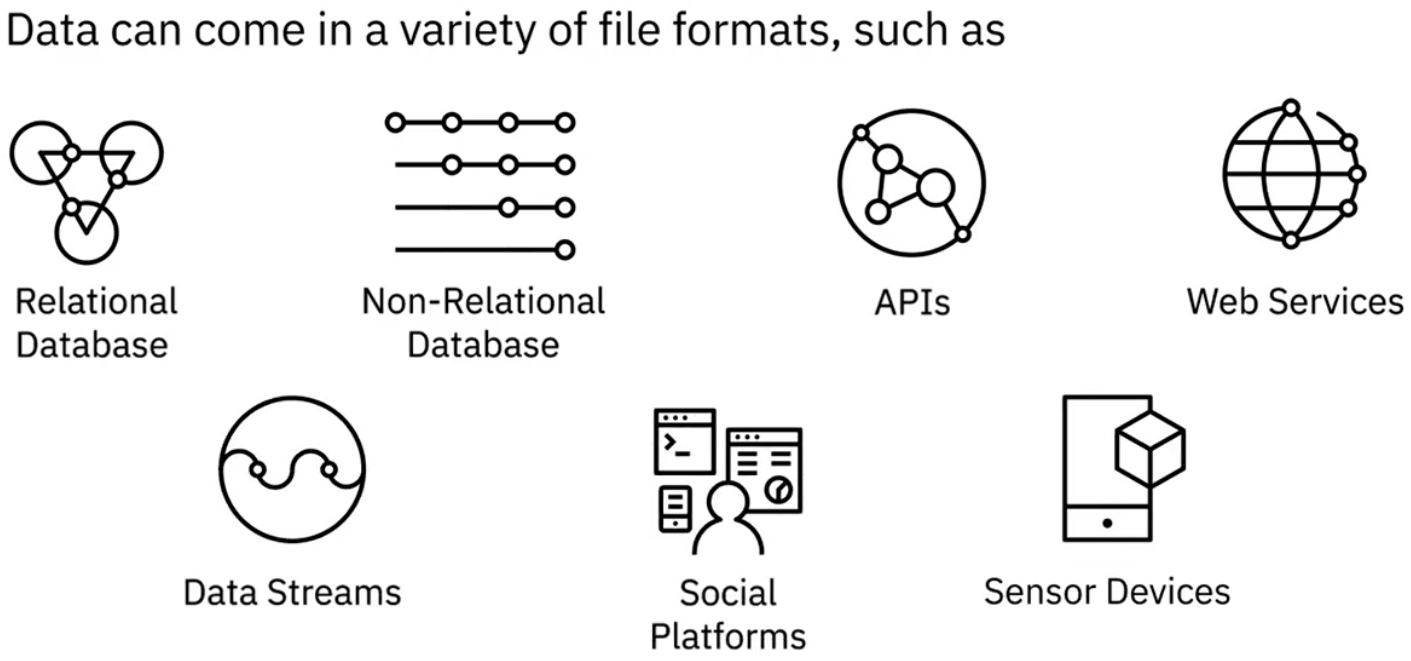
Structured data follows a rigid format and can be organized neatly into rows and columns. This is the data that you see typically in databases and spreadsheets, for example.



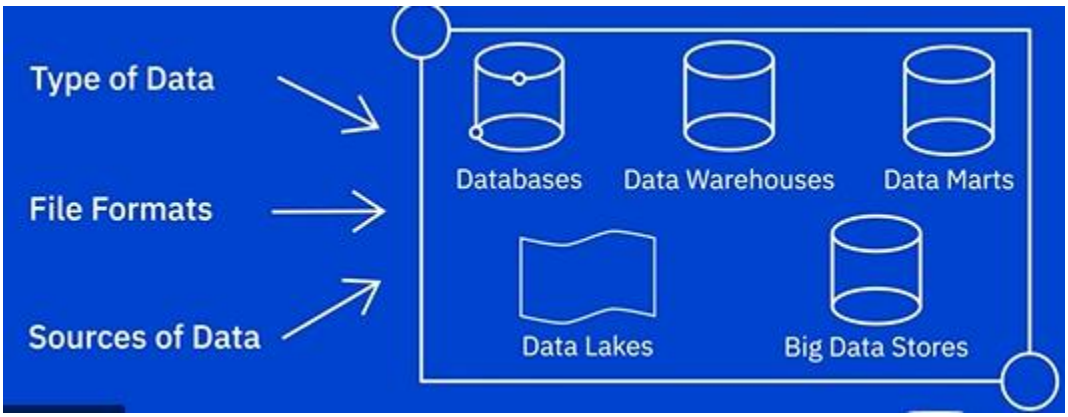
Semi-structured data is a mix of data that has consistent characteristics and data that doesn't conform to a rigid structure. For example, emails. An email has a mix of structured data, such as the name of the sender and recipient, but also has the contents of the email, which is unstructured data. And then

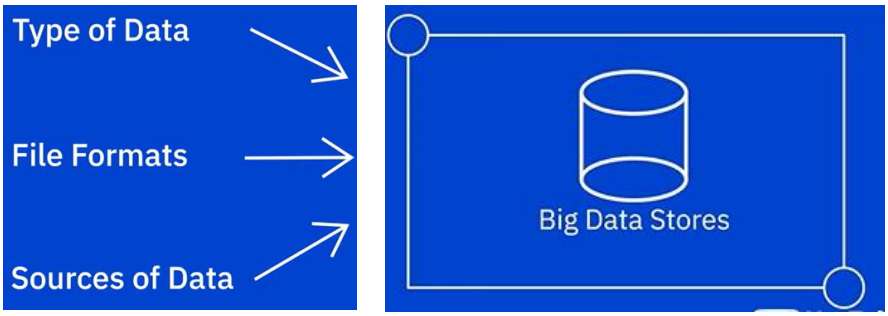


Unstructured data: Data that is complex, and mostly qualitative information that is impossible to reduce to rows and columns. For example, photos, videos, text files, PDFs, and social media content. The type of data drives the kind of data repositories that the data can be collected and stored in, and also the tools that can be used to query or process the data.



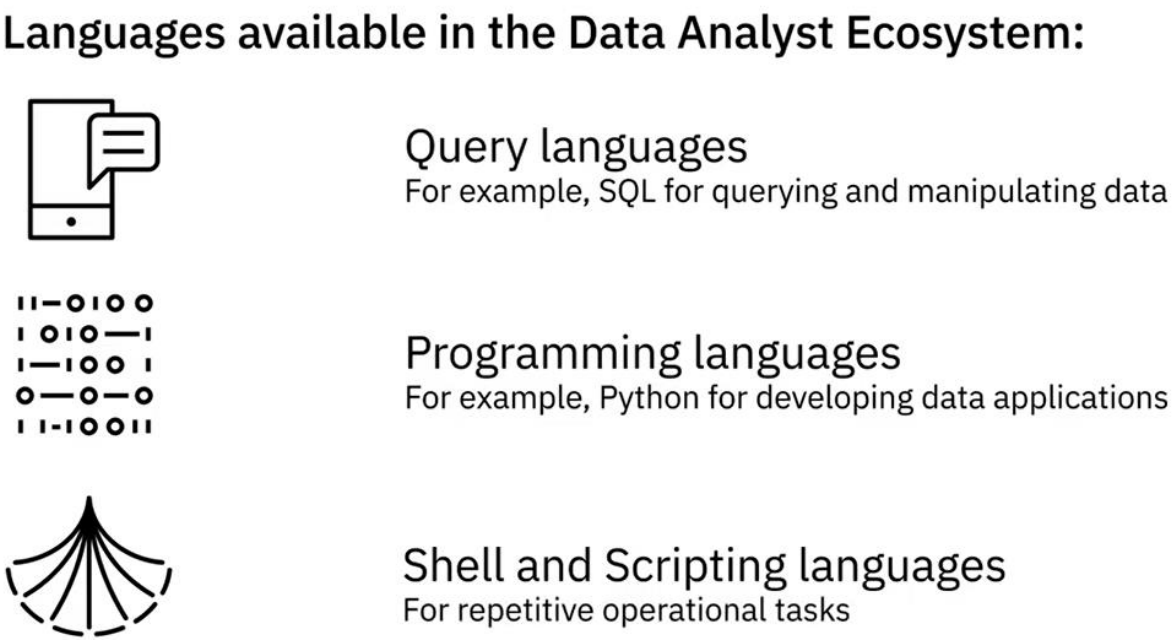
Data Repositories





If you're working with big data, for example, you will need big data warehouses, that allow you to store and process large-volume high-velocity data and also frameworks that allow you to perform complex analytics in real-time on big data.

Languages



Data Analyst Ecosystem

Automated tools, frameworks, and processes for all stages of the analytics process are part of the Data Analysts ecosystem.

Types of Data

Data is unorganized information that is processed to make it meaningful. Generally, data comprises of facts, observations, perceptions, numbers, characters, symbols, and images that can be interpreted to derive meaning.

One of the ways in which data can be categorized is by its structure-data can be: Structured; Semi-structured, or Unstructured

Structured data has a well-defined structure or adheres to a specified data model can be stored in well-defined schemas such as databases and in many cases can be represented in a tabular manner with rows and columns. Structured data is objective facts and numbers that can be collected, exported, stored, and organized in typical databases.

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.

Semi-structured data is data that has some organizational properties but lacks a fixed or rigid schema. Semi-structured data cannot be stored in the form of rows and columns as in databases. It contains tags and elements, or metadata, which is used to group data and organize it in a 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.

Unstructured data is data that does not have an easily identifiable structure and, therefore, cannot be organized in a mainstream relational database in the form of rows and columns. It does not follow any particular format, sequence, semantics, or rules.

Unstructured data can deal with the heterogeneity of sources and has a variety of business intelligence and analytics applications. 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.

To summarize:

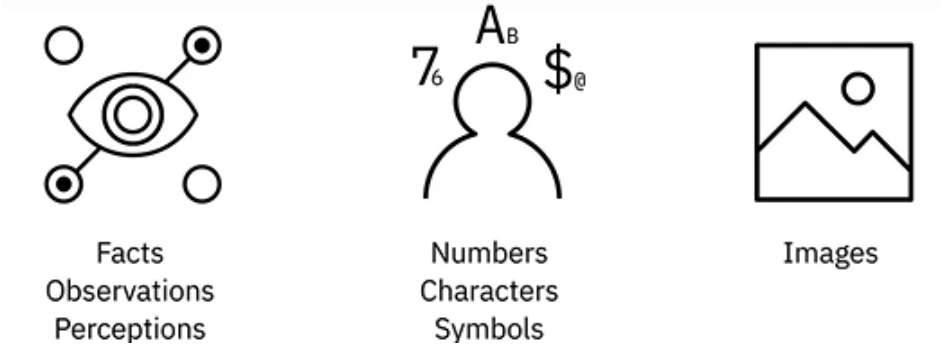
Structured data is data that is well organized in formats that can be stored in databases and lends itself to standard data analysis methods and tools;

Semi-structured data is data that is somewhat organized and relies on meta tags for grouping and hierarchy; and

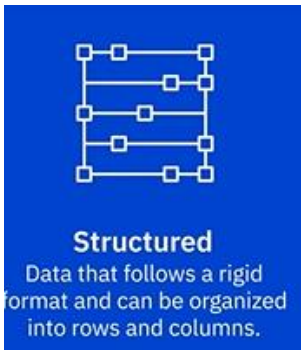
Unstructured data is data that is not conventionally organized in the form of rows and columns in a particular format.

Types of Data

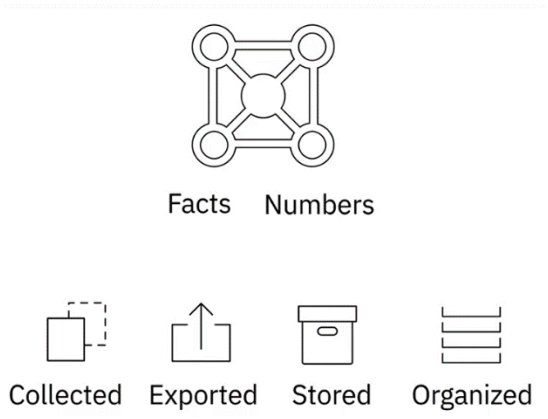
What is data?

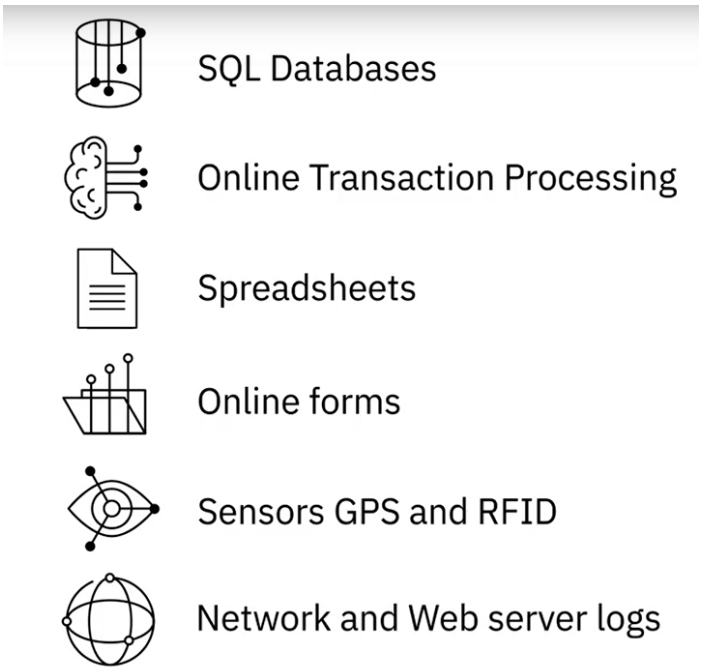


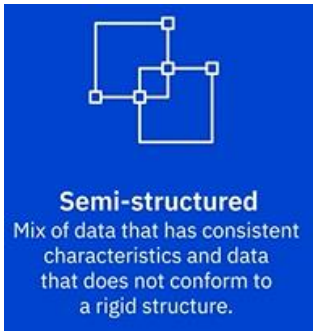
Types of Data



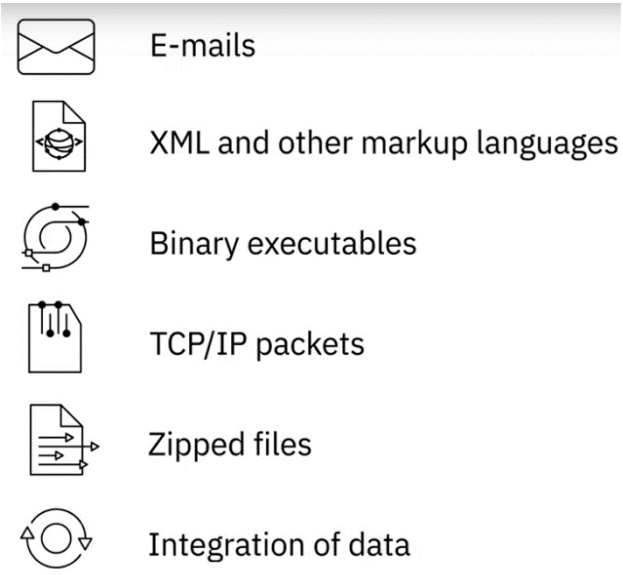
- Has a well-defined structure
- Can be stored in well-defined schemas
- Can be represented in a tabular manner with rows and columns

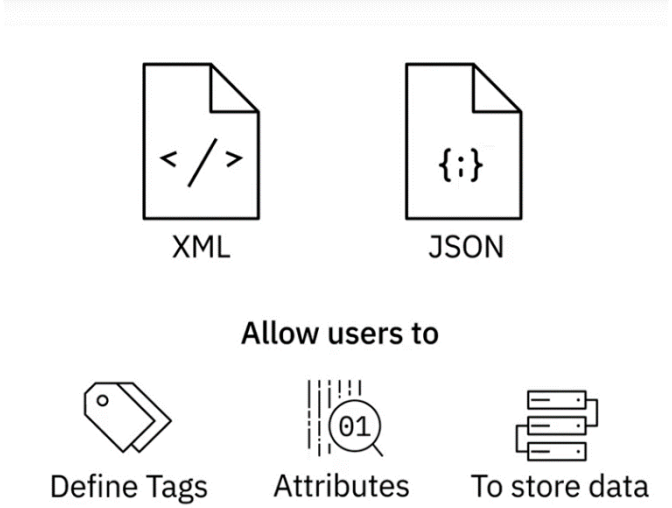




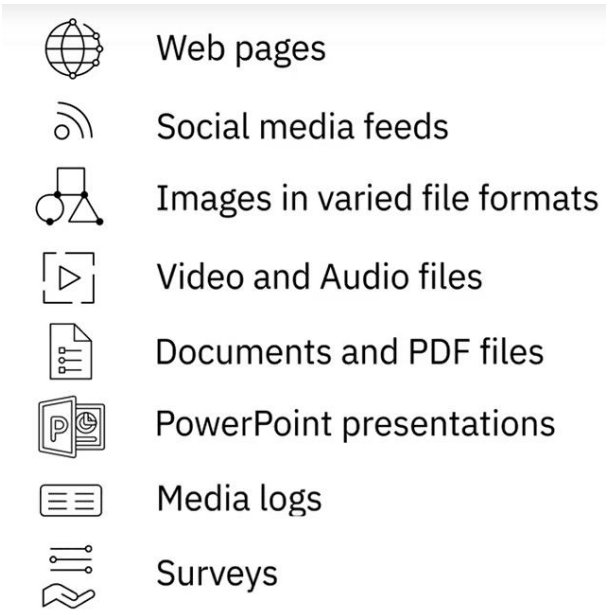


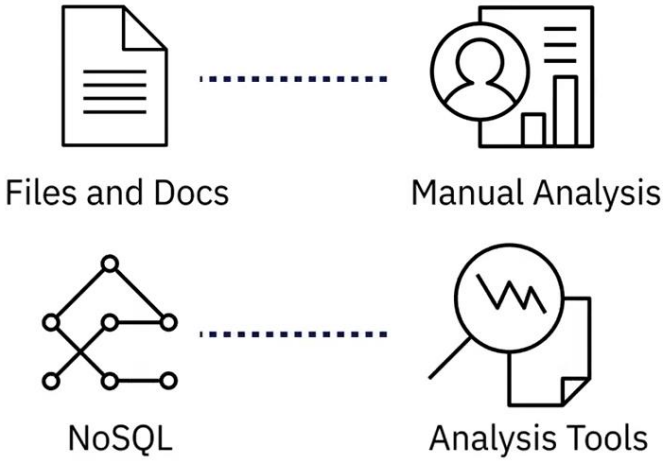
- Has some organizational properties but lacks a fixed or rigid schema
- Cannot be stored in the form of rows and columns as in databases
- Contains tags and elements, or metadata, which is used to group data and organize it in a hierarchy





- Does not have an easily identifiable structure
- Cannot be organized in a mainstream relational database in the form of rows and columns
- Does not follow any particular format, sequence, semantics, or rules





To summarize:

Structured data is data that is well organized in formats that can be stored in databases and lends itself to standard data analysis methods and tools;

Semi-structured data is data that is somewhat organized and relies on meta tags for grouping and hierarchy; and

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

Understanding Different Types of File Formats

As a data professional, you will be working with a variety of **data file types, and formats**. It is important to understand the underlying structure of file formats along with their benefits and limitations. This understanding will support you to make the right decisions on the formats best suited for your data and performance needs.

Some of the **standard file formats** that we will cover in this video include:

- 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,

Delimited text files are text files used to store data as text in which each line, or row, has values separated by a delimiter; where a delimiter is a sequence of one or more characters for specifying the boundary between independent entities or values.

Any character can be used to separate the values, but most common delimiters are the comma, tab, colon, vertical bar, and space.

Comma-separated values (or CSVs) and **tab-separated values (or TSVs) are** the most commonly used file types in this category.

In **CSVs,** the delimiter is a comma while in TSVs, the delimiter is a tab. When literal commas are present in text data and therefore cannot be used as delimiters,

TSVs serve as an alternative to CSV format.

Tab stops are infrequent in running text. Each row, or horizontal line, in the text file has a set of values separated by the delimiter, and represents a record. The first row works as a column header, where each column can have a different type of data.

For example, a column can be of date type, while another can be a string or integer type data.

Delimited files allow field values of any length and are considered a standard format for providing straightforward information schema. They can be processed by almost all existing applications.

Delimiters also represent one of various means to specify boundaries in a data stream.

Microsoft Excel Open XML Spreadsheet, or XLSX, is a Microsoft Excel Open XML file format that falls under the spreadsheet file format.

It is an XML-based file format created by Microsoft. In an .XLSX, also known as a workbook, there can be multiple worksheets. And each worksheet is organized into rows and columns, at the intersection of which is the cell. Each cell contains data.

XLSX uses the open file format, which means it is generally accessible to most other applications. It can use and save all functions available in Excel and is also known to be one of the more secure file formats as it cannot save malicious code.

Extensible Markup Language, or XML, is a markup language with set rules for encoding data. The XML file format is both readable by humans and machines. It is a self-descriptive language designed for sending information over the internet.

XML is similar to HTML in some respects, but also has differences. For example, an .XML does not use predefined tags like .HTML does. XML is platform independent and programming language independent and therefore simplifies data sharing between various systems.

Portable Document Format, or PDF, is a file format developed by Adobe to present documents independent of application software, hardware, and operating systems, which means it can be viewed the same way on any device. This format is frequently used in legal and financial documents and can also be used to fill in data such as for forms.

JavaScript Object Notation, or JSON, is a text-based open standard designed for transmitting structured data over the web. The file format is a language-independent data format that can be read in any programming language.

JSON is easy to use, is compatible with a wide range of browsers, and is considered as one of the best tools for sharing data of any size and type, even audio and video. That is one reason, many APIs and Web Services return data as JSON.

Understanding Different Types of File Formats

As a data professional, you will be working with a variety of data file types, and formats. It is important to understand the underlying structure of file formats along with their benefits and limitations. This understanding will support you to make the right decisions on the formats best suited for your data and performance needs.

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

Delimited Text files

Files used to store data as text

Each value is separated by a delimiter

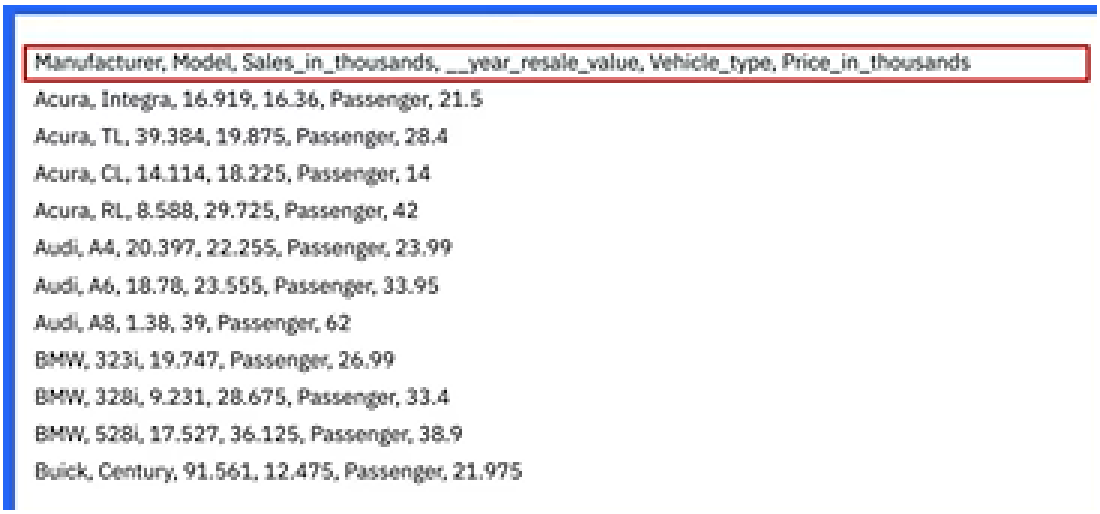
Delimiter - A sequence of one or more characters for specifying the boundary between independent entities or values.

Comma, Tab, Colon, Vertical Bar, Space

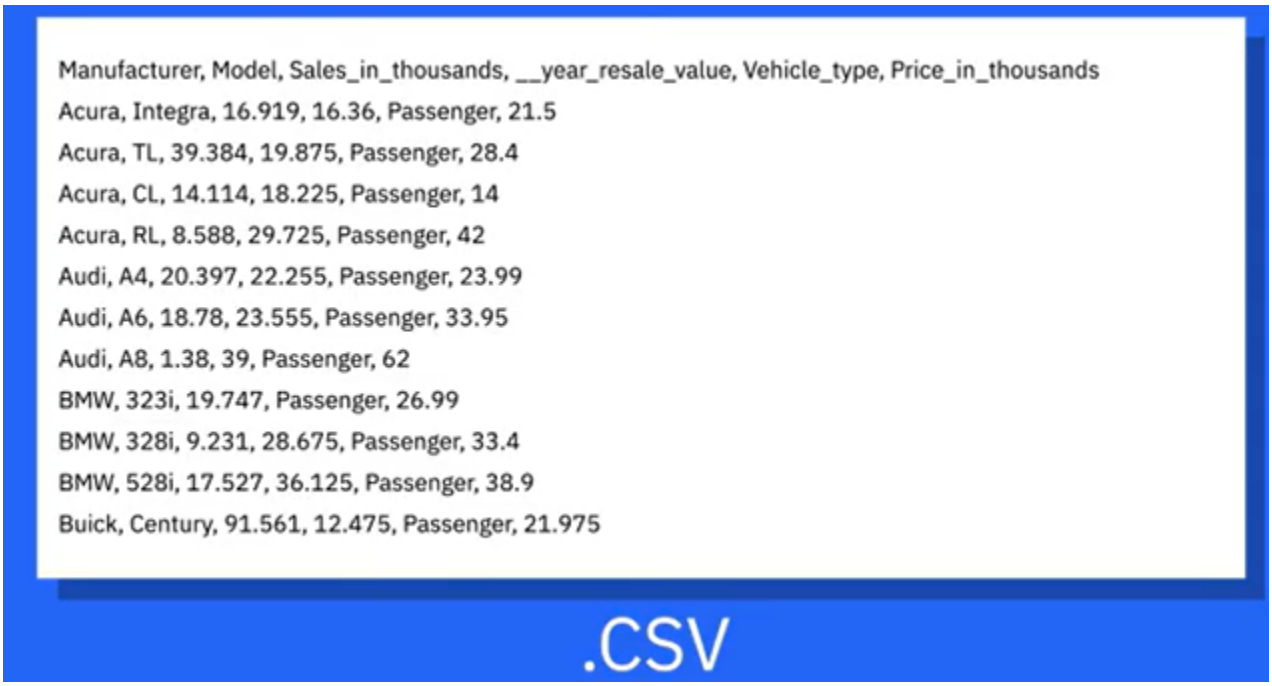
Comma-separated values

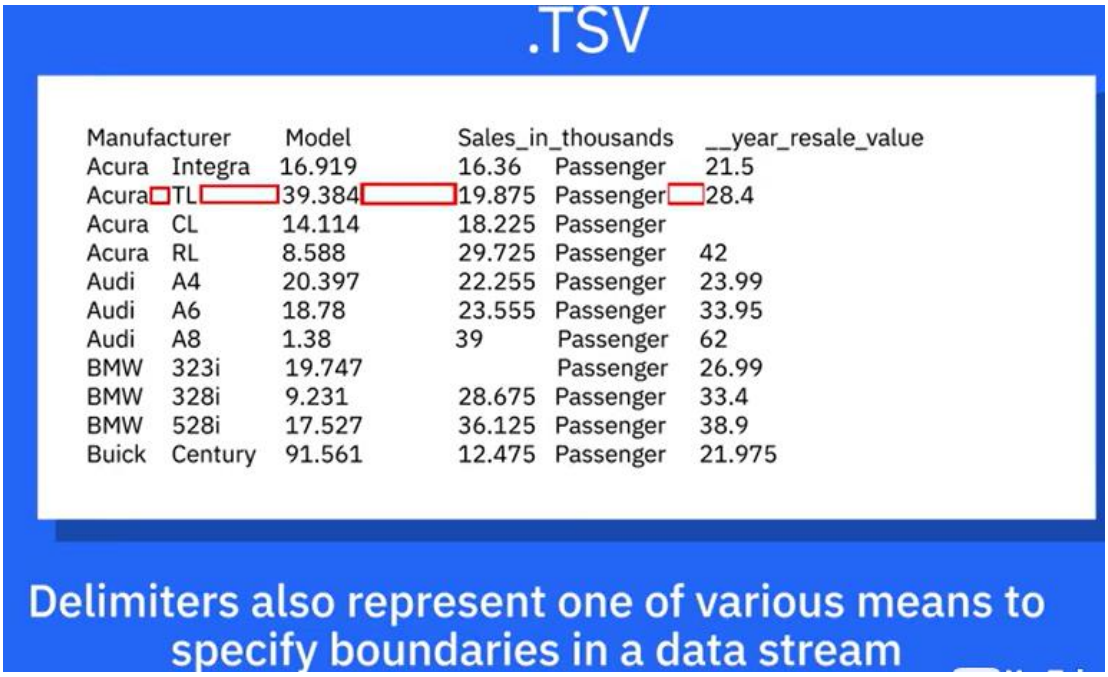
Tab-separated values

In CSVs, the delimiter is a comma while in TSVs, the delimiter is a tab. When literal commas are present in text data and therefore cannot be used as delimiters, TSVs serve as an alternative to CSV format.



Tab stops are infrequent in running text. Each row, or horizontal line, in the text file has a set of values separated by the delimiter, and represents a record. The first row works as a column header, where each column can have a different type of data. For example, a column can be of date type, while another can be a string or integer type data.

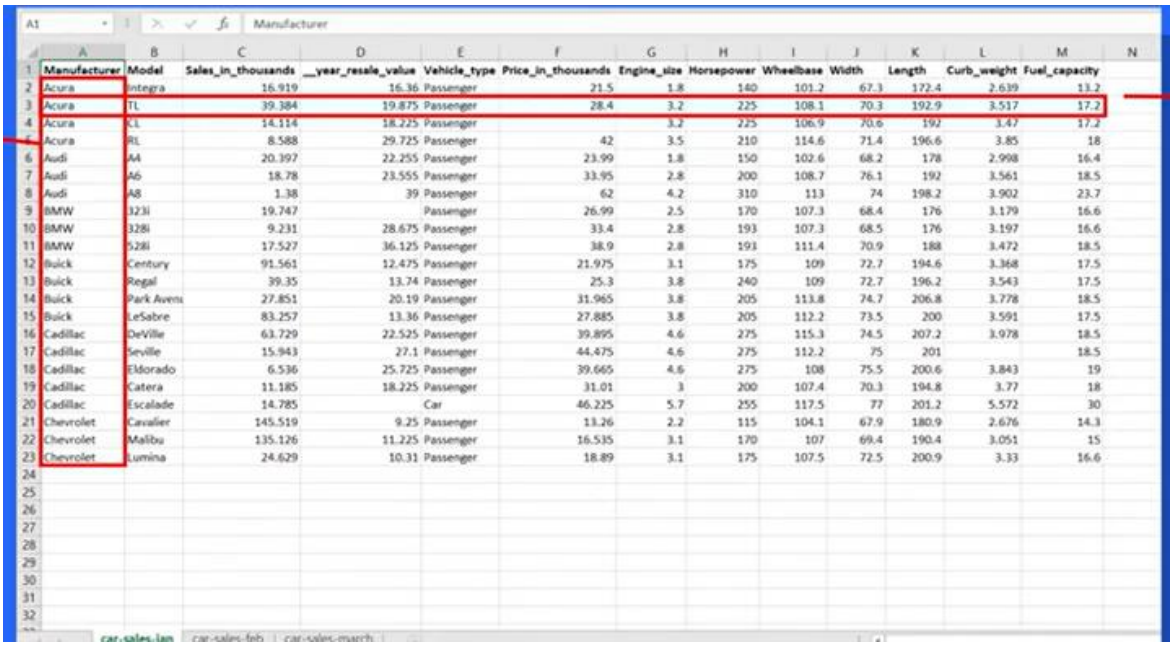




Delimited files allow field values of any length and are considered a standard format for providing straightforward information schema. They can be processed by almost all existing applications. Delimiters also represent one of various means to specify boundaries in a data stream.

Microsoft Excel Open XML Spreadsheet, or .XLSX

Microsoft Excel Open XML Spreadsheet, or XLSX, is a Microsoft Excel Open XML file format that falls under the spreadsheet file format. It is an XML-based file format created by Microsoft.



Manufacturer	Model	Sales_in_thousands	__year_resale_value	Vehicle_type	Price_in_thousands	Engine_size	Horsepower	Wheelbase	Width	Length	Curb_weight	Fuel_capacity
Acura	Integra	16.919	16.36	Passenger	21.5	1.8	140	101.2	67.3	172.4	2.639	13.2
Acura	TL	39.384	19.875	Passenger	28.4	3.2	225	108.1	70.3	192.9	3.517	17.2
Acura	CL	14.114	18.225	Passenger		3.2	225	106.9	70.6	192	3.47	17.2
Acura	RL	8.588	29.725	Passenger	42	3.5	210	114.6	71.4	196.6	3.85	18
Audi	A4	20.397	22.255	Passenger	23.99	1.8	150	102.6	68.2	178	2.998	16.4
Audi	A6	18.78	23.555	Passenger	33.95	2.8	200	108.7	76.1	192	3.561	18.5
Audi	A8	1.38	39	Passenger	62	4.2	310	113	74	198.2	3.902	23.7
BMW	323i	19.747		Passenger	26.99	2.5	170	107.3	68.4	176	3.179	16.6
BMW	328i	9.231	28.675	Passenger	33.4	2.8	193	107.3	68.5	176	3.197	16.6
BMW	528i	17.527	36.125	Passenger	38.9	2.8	193	111.4	70.9	188	3.472	18.5
Buick	Century	91.561	12.475	Passenger	21.975	3.1	175	109	72.7	194.6	3.368	17.5
Buick	Regal	39.35	13.74	Passenger	25.3	3.8	240	109	72.7	196.2	3.543	17.5
Buick	Park Avenu	27.851	20.19	Passenger	31.965	3.8	205	113.8	74.7	206.8	3.778	18.5
Buick	LeSabre	83.257	13.36	Passenger	27.885	3.8	205	112.2	73.5	200	3.591	17.5
Cadillac	DeVille	63.729	22.525	Passenger	39.895	4.6	275	115.3	74.5	207.2	3.978	18.5
Cadillac	Seville	15.943	27.1	Passenger	44.475	4.6	275	112.2	75	201		18.5
Cadillac	Eldorado	6.536	25.725	Passenger	39.665	4.6	275	108	75.5	200.6	3.843	19
Cadillac	Catera	11.185	18.225	Passenger	31.01	3	200	107.4	70.3	194.8	3.77	18
Cadillac	Escalade	14.785		Car	46.225	5.7	255	117.5	77	201.2	5.572	30
Chevrolet	Cavalier	145.519	9.25	Passenger	13.26	2.2	115	104.1	67.9	180.9	2.676	14.3
Chevrolet	Malibu	135.126	11.225	Passenger	16.535	3.1	170	107	69.4	190.4	3.051	15
Chevrolet	Lumina	24.629	10.31	Passenger	18.89	3.1	175	107.5	72.5	200.9	3.33	16.6

- XLSX uses the open file format, which means it is generally accessible to most other applications.
- It can use and save all functions available in Excel
- more secure file formats as it cannot save malicious code.

Extensible Markup Language or .XML

```
<?xml version="1.0"?>
<car-specs>

<manufacturer>Acura<manufacturer>

<model>Integra<model>

<sales_in-thousands>16.919<sales_in-thousands>

<year_resale_value>16.36<year_resale_value>

<vehicle_type>Passenger<vehicle_type>

<car-specs>
```

Extensible Markup Language, or XML, is a markup language with set rules for encoding data.

- Readable by both humans and machines
- Self-descriptive language
- Similar to .HTML in some respects
- Does not use predefined tags like .HTML does
- Platform independent
- Programming language independent
- Makes it simpler to share data between systems

Portable Document Format or PDF

USAID
FROM THE AMERICAN PEOPLE

OMB No. 0412-0004
Expiration Date: 02/29/2012

APPLICATION FOR APPROVAL OF COMMODITY ELIGIBILITY (FORM AID 11) | Transaction No. *(Assigned by USAID)*

TRANSACTION IDENTIFICATION

1. USAID Letter of Commitment No.	2. Payment Terms: U.S. Bank Letter of Credit No.	Date	Name and Address of U.S. Bank (Advising Bank)	Other Payment Terms (If any)

3. Import License No.	Date	4. Supplier's Relationship to Authorized Source Country
		☐ Corporation or Partnership Organized under Source Country Laws ☐ Individual, Citizen or Permanent Resident of Source Country ☐ Controlled Foreign Corporation ☐ Other

5. Supplier's Name and Address	6. Importer's Name and Address

7. Contract: Total Amount (Funded by USAID)	Date	8. Shipping Plans at Time of Application: a. Partial Shipment ☐ No ☐ Yes	b. Loading Port	c. Destination Port	d. Month(s) of Shipment
$					

COMMODITY IDENTIFICATION

9. Schedule B 10-Digit Code(s)	10. Commodity Description, Quantity, Size	11. Unit and Unit Price, or Total FAS/FOB Vessel Price, or FCA Price *(Named Port of Loading/Airport)*
(a)		
(b)		
(c)		
(d)		
(e)		

12. Commodity Condition:

☐ New and Unused ☐ Used - Not Rebuilt or Reconditioned ☐ Rebuilt ☐ Reconditioned ☐ Other (Specify below)

13. Source of Commodity: a. Authorized Area	b. Shipped From	c. Produced In	14. Components (Parts of the Commodity): a. From Other than 13.a Source	b. If 14.a is "Yes", Name Country	c. Cost Per Unit of 14.b Components

Portable Document Format, or PDF,Is a file format developed by Adobe to present documents independent of application software, hardware, and operating systems.

- Can be viewed the same way on any device
- Is frequently used in legal and financial documents
- Can also be used to fill in data for forms

Javascript Object Notation or .JSON

```
{
    "Employee": [
        {
          "id":"1",
          "Manufacturer": "Audi",
          "Model": "Integra",
        },
        {
          "id":"2",
          "Manufacturer": "Buick",
          "Model": "LeSabre",
        },
        {
          "id":"3",
          "Manufacturer": "Cadillac",
          "Model": "Escalade",
        }
    ]
}
```

JavaScript Object Notation, or JSON, is a text-based open standard designed for transmitting structured data over the web.

- Language-independent data format
- Can be read in any programming language
- Easy to use
- Compatible with a wide range of browsers
- Considered as one of the best tools for sharing data

Sources of Data

As we touched upon in one of our previous videos, data sources have never been as dynamic and diverse as they are today.

In this video, we will look at some **common sources** such as:

Relational Databases,

Flat files and XML Datasets,

APIs and Web Services,

Web Scraping,

Data Streams, and Feeds.

Typically, organizations 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.

External to the organization, there are other publicly and privately available datasets. For example, government organizations 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 format is CSV in which values are separated by commas.
- Spreadsheet files are a special type of flat files, that also organize 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.

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.
- 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 summarize 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 co-relating 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 organization.

Web 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.

Data streams are

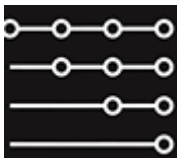
- another widely used source for aggregating constant streams of data flowing from sources such as instruments, IoT devices and applications, GPS data from cars, computer programs, websites, and social media posts.
- This data is generally timestamped and also geo-tagged for geographical identification.
- 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.

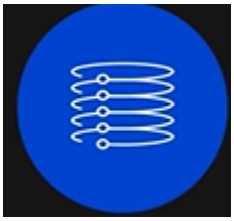
- These are typically used for capturing updated data from online forums and news sites where data is refreshed on an ongoing basis.
- Using a feed reader, which is an interface that converts RSS text files into a stream of updated data, updates are streamed to user devices.

Sources of Data

Common source of data: data sources have never been as dynamic and diverse as they are today

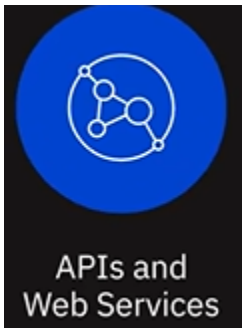




Common Sources of Data:



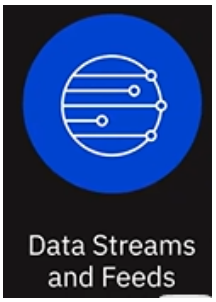
Relational Databases

Flat files and XML Databases



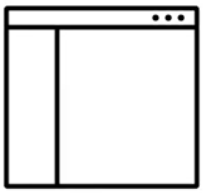
APIs and Webservices

Web scraping



Data Streams and Feeds

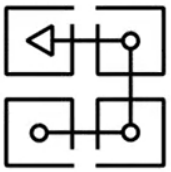
Relational Databases



Business activities

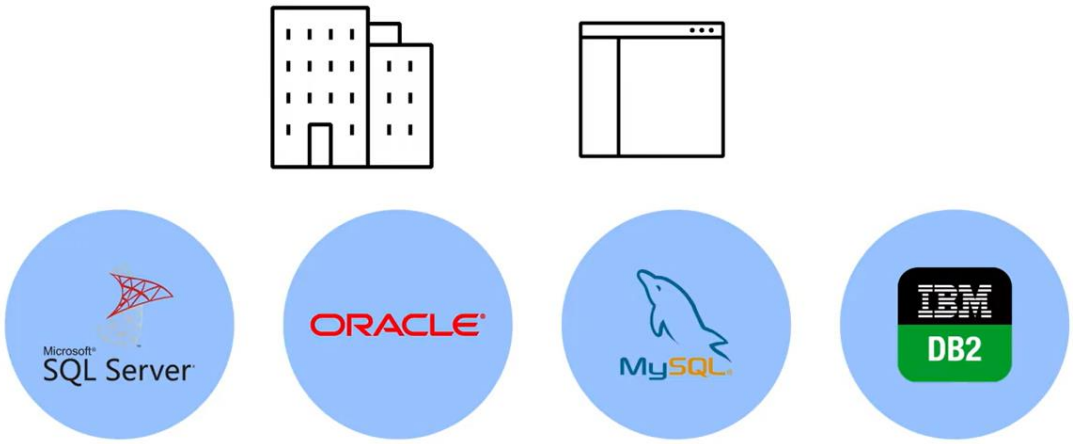
Customer transactions

Human resource activities

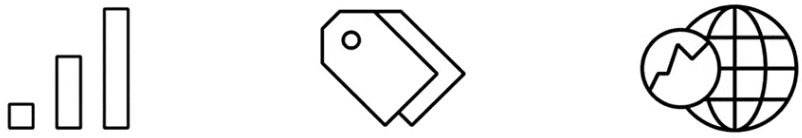


Workflows

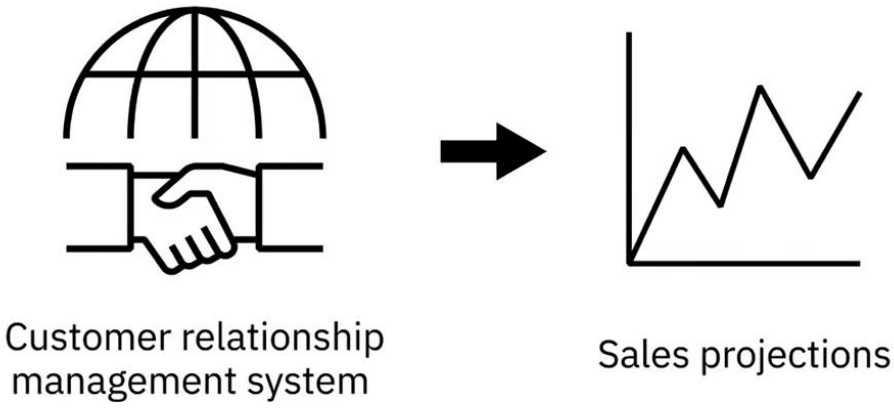
To store data in a structured way:



Store structured data that can be leveraged for analysis:

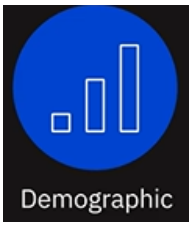


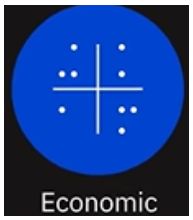
Data from a retail transactions system :



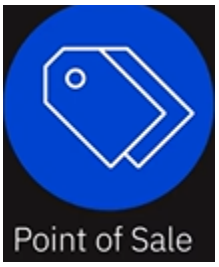
Flatfile and XML

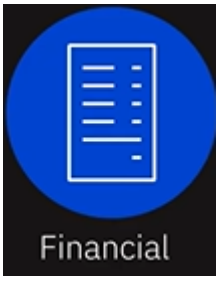
Available Public and Private Datasets

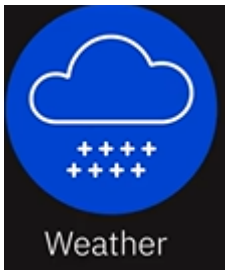




Companies that sell specific data







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
- Each line or row in one record
- Each value is separated by a delimeter

All of the data in flat file maps to a single table

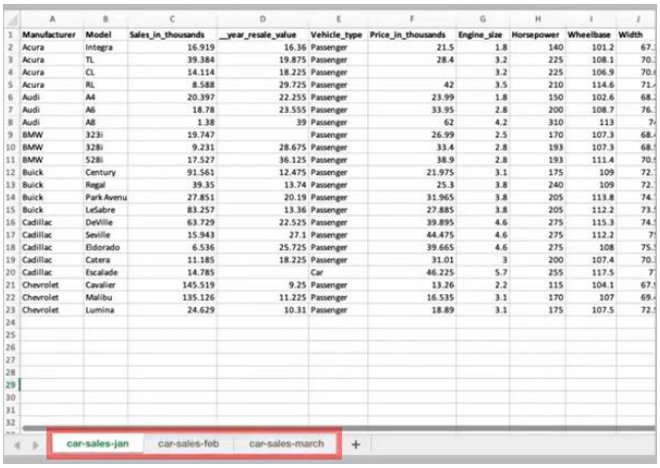
```
"Manufacturer", "Model", "Sales_in_thousands", "__year_resale_value", "Vehicle_type", "Price_in_thousands"
"Acura", "Integra", "16.919", "16.36", "Passenger", "21.5"
"Acura", "TL", "39.384", "19.875", "Passenger", "28.4"
"Acura", "CL", "14.114", "18.225", "Passenger", "14"
"Acura", "RL", "8.588", "29.725", "Passenger", "42"
"Audi", "A4", "20.397", "22.255", "Passenger", "23.99"
"Audi", "A6", "18.78", "23.555", "Passenger", "33.95"
"Audi", "A8", "1.38", "39", "Passenger", "62"
"BMW", "323i", "19.747", "Passenger", "26.99"
"BMW", "328i", "9.231", "28.675", "Passenger", "33.4"
"BMW", "528i", "17.527", "36.125", "Passenger", "38.9"
"Buick", "Century", "91.561", "12.475", "Passenger", "21.975"
```

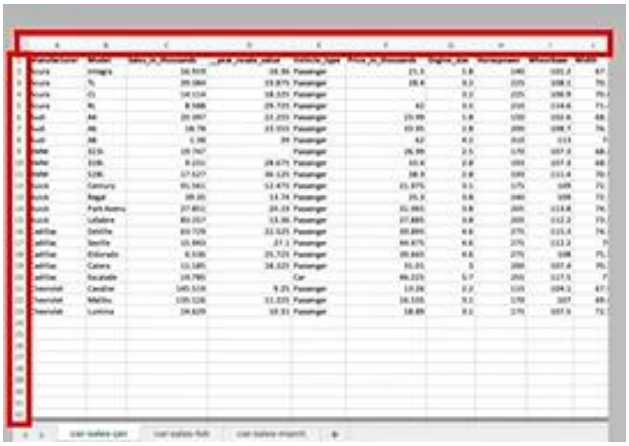
Manufacturer	Model	Sales_in_thousands	__year_resale_value	Vehicle_type	Price_in_thousands
Acura	Integra	16.919	16.36	Passenger	21.5
Acura	TL	39.384	19.875	Passenger	28.4
Acura	CL	14.114	18.225	Passenger	
Acura	RL	8.588	29.725	Passenger	42
Audi	A4	20.397	22.255	Passenger	23.99
Audi	A6	18.78	23.555	Passenger	33.95
Audi	A8	1.38	39	Passenger	62
BMW	323i	19.747		Passenger	26.99
BMW	328i	9.231	28.675	Passenger	33.4
BMW	528i	17.527	36.125	Passenger	38.9
Buick	Century	91.561	12.475	Passenger	21.975

- Most common flat file format is .CSV

```
"EMPNO","ENAME","JOB","MGR","HIREDATE","SAL","COMM","DEPTNO"
9999,"ADAMS","CLERK",7788,23-MAY-1987 12.00.00,1100,,20
7369,"SMITH","CLERK",7902,17-DEC-1980 12.00.00,800,,20
7499,"ALLEN","SALESMAN",7698,20-FEB-1981 12.00.00,1600,300,30
7521,"WARD","SALESMAN",7698,22-FEB-1981 12.00.00,1250,500,30
7566,"JONES","MANAGER",7839,02-APR-1981 12.00.00,2975,,20
7654,"MARTIN","SALESMAN",7698,28-SEP-1981 12.00.00,1250,1400,30
7698,"BLAKE","MANAGER",7839,01-MAY-1981 12.00.00,2850,,30
7782,"CLARK","MANAGER",7839,09-JUN-1981 12.00.00,2450,,10
7788,"SCOTT","ANALYST",7566,19-APR-1987 12.00.00,3000,,20
7839,"KING","PRESIDENT",,17-NOV-1981 12.00.00,5000,,10
7844,"TURNER","SALESMAN",7698,08-SEP-1981 12.00.00,1500,0,30
7876,"ADAMS","CLERK",7788,23-MAY-1987 12.00.00,1100,,20
7900,"JAMES","CLERK",7698,03-DEC-1981 12.00.00,950,,30
7902,"FORD","ANALYST",7566,03-DEC-1981 12.00.00,3000,,20
7934,"MILLER","CLERK",7782,23-JAN-1982 12.00.00,1300,,10
```

Spreadsheet files





- Spreadsheet files are a special type of flat files
- Organize data in a tabular format – rows and columns.
- Can contain multiple worksheets, and each worksheet can map to a different table.

- .XLS or .XLSX format is probably the most common spreadsheet
- Others include Google sheets, Apple Numbers, and LibreOffice.
-

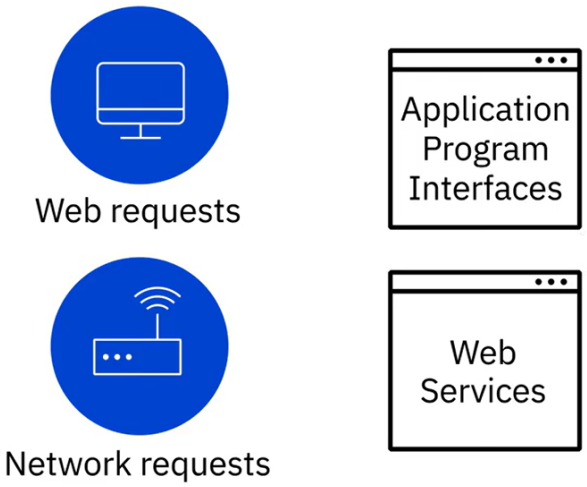
XML

- 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

```
<?xml version="1.0"?>
<car-specs>

<manufacturer>Acura<manufacturer>

<model>Integra<model>

<sales_in-thousands>16.919<sales_in-thousands>

<year_resale_value>16.36<year_resale_value>

<vehicle_type>Passenger<vehicle_type>

<car-specs>
```

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.

Popular examples of API

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

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 summarize 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 co-relating 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 organization.

Web Scraping

- Extract relevant data from unstructured sources
- Also known as Screen scraping, Web harvesting, and Web data extraction
- Downloads specific data based on defined parameters
- Can extract text, contact information, images, videos, product items, and more...

Web 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.

Popular uses:

Providing price comparisons by collecting product details from retailer, manufacturers, and eCommerce websites

Generating sales leads through public data sources

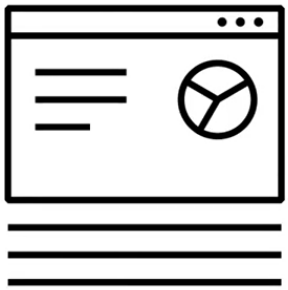
Extracting data from posts and authors on various forums and communities

Collecting training and testing datasets for machine learning models

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.

Popular web scraping tools:

- BeautifulSoup
- Scrapy
- Pandas
- Selenium



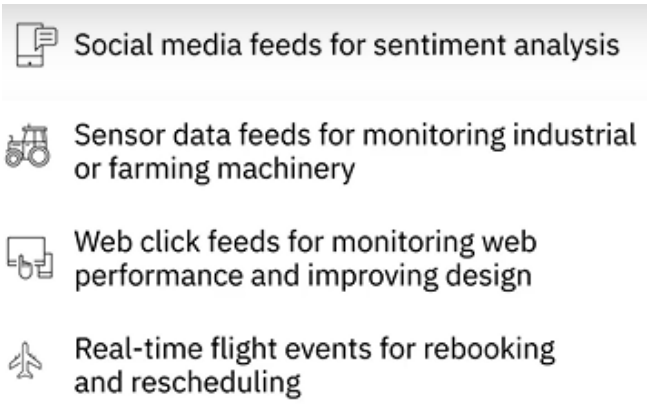
Data Streams and Feeds

Aggregating streams of data flowing from instruments, IoT devices and applications, GPS data from cars, computer programs, websites, and social media posts

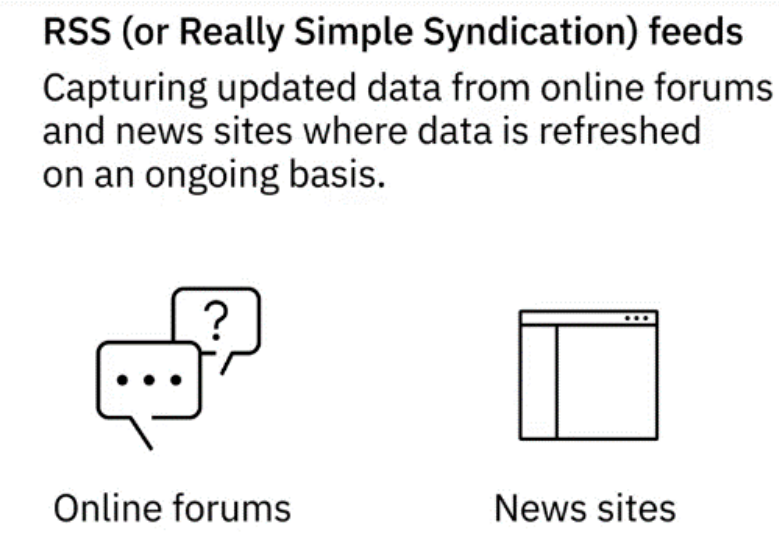
- Stock and market tickers for financial trading
- Retail transaction streams for predicting demand and supply chain management
- Surveillance and video feeds for threat detection

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.



Languages for Data Professionals

We will learn about some of the languages relevant to the work of data professionals. These can be categorized as – query languages, programming languages, and shell scripting.

Having proficiency in at least one language in each category is essential for any data professional.

Simply stated:

Query languages are designed for accessing and manipulating data in a database; for example, SQL, Programming languages are designed for developing applications and controlling application behavior; for example, Python, R, and Java; and Shell and Scripting languages, such as Unix/Linux Shell, and PowerShell, are ideal for repetitive and time-consuming operational tasks.

In the remaining video, we will examine these languages in greater depth. SQL, or Structured Query Language, is a querying language designed for accessing and manipulating information from, mostly, though not exclusively, relational databases.

Using SQL,

- we can write a set of instructions to perform operations such as Insert, update, and delete records in a database; Create new databases, tables, and views; and Write stored procedures—which means you can write a set of instructions and call them for later use

Here are some advantages of using SQL:

- SQL is portable and can be used independent of the platform, It can be used for querying data in a wide variety of databases and data repositories, although each vendor may have some variations and special extensions,
- It has a simple syntax that is similar to the English language, Its syntax allows developers to write programs with fewer lines than some of the other programming languages using basic keywords such as select, insert, into, and update,
- It can retrieve large amounts of data quickly and efficiently,
- It runs on an interpreter system, which means code can be executed as soon as it is written, making prototyping quick and easy.

SQL is one of the most popular querying language. Due to its large user community and the sheer volume of documentation accumulated over the years, it continues to provide a uniform platform, worldwide, to all its users.

Python

- is a widely-used open-source, general-purpose, high-level programming language.
- Its syntax allows programmers to express their concepts in fewer lines of code, as compared to some of the older languages.
- Python is perceived as one of the easiest languages to learn and has a large developer community.

- Because of its focus on simplicity and readability, and a low learning curve, it's an ideal tool for beginning programmers.
- It is great for performing high-computational tasks in vast amounts of data, which can otherwise be extremely time-consuming and cumbersome.
- Python provides libraries like Numpy and Pandas, which eases this task by the use of parallel processing. It has inbuilt functions for almost all of the frequently used concepts.
- Python supports multiple programming paradigms, such as object-oriented, imperative, functional, and procedural, making it suitable for a wide variety of use cases.

Now let's look at some of the reasons that make Python one of the fastest-growing programming languages in the world today.

- It is easy to learn - With Python, you have the advantage of using fewer lines of code to accomplish tasks compared to other languages.
- It is open-source — Python is free and uses a community-based model for development.
- It runs on Windows and Linux environments and can be ported to multiple platforms.
- It has widespread community support with plenty of useful analytics libraries available.
- It has several open-source libraries for data manipulation, data visualization, statistics, and mathematics, to name just a few.

Its vast array of libraries and functionalities also include:

- Pandas for data cleaning and analysis, Numpy and Scipy, for statistical analysis,
- Beautifulsoup and Scrapy for web scraping,
- Matplotlib and Seaborn to visually represent data in the form of bar graphs, histogram,

  and pie-charts,

- Opencv for image processing.

R is

- an open-source programming language and environment for data analysis, data visualization, machine learning, and statistics.
- Widely used for developing statistical software and performing data analytics, it is especially known for its ability to create compelling visualizations, giving it an edge over some of the other languages in this space.

Some of the key benefits of R include the following:

- It is an open-source platform-independent programming language,
- It can be paired with many programming languages, including Python,
- It is highly extensible, which means developers can continue to add functionalities by defining new functions,
- It facilitates the handling of structured as well as unstructured data which means it has a more comprehensive data capability,

- It has libraries such as Ggplot2 and Plotly that offer aesthetic graphical plots to its users,
- You can make reports with the data and scripts embedded in them; also, interactive web apps that allow users to play with the results and the data,
- It is dominant among other programming languages for developing statistical tools.

Java is

- an object-oriented, class-based, and platform-independent programming language originally developed by Sun Microsystems. It is among the top-ranked programming languages used today.
- Java is used in a number of processes all through data analytics, including cleaning data, importing and exporting data, statistical analysis, and data visualization.
- In fact, most of the popular frameworks and tools used for big data are typically written in Java, such as Hadoop, Hive, and Spark.
- It is perfectly suited for speed-critical projects.

A Unix/Linux Shell is a computer program written for the UNIX shell. It is a series of UNIX commands written in a plain text file to accomplish a specific task.

- Writing a shell script is fast and easy. It is most useful for repetitive tasks that may be time-consuming to execute by typing one line at a time.

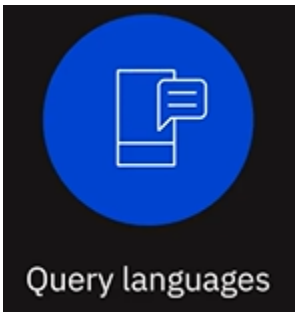
Typical operations performed by shell scripts include:

- file manipulation,
- program execution,
- system administration tasks such as disk backups and evaluating system logs,
- installation scripts for complex programs, executing routine backups, running batches,

PowerShell is

- a cross-platform automation tool and configuration framework by Microsoft that is optimized for working with structured data formats, such as JSON, CSV, XML, and REST APIs, websites, and office applications.
- It consists of a command-line shell and scripting language.
- PowerShell is object-based, which makes it possible to filter, sort, measure, group, compare, and many more actions on objects as they pass through a data pipeline.
- It is also a good tool for data mining, building GUIs, and creating charts, dashboards, and interactive reports.

Languages for Data Professionals



Query Languages

Query languages are designed for accessing and manipulating data in a database; for example, SQL,

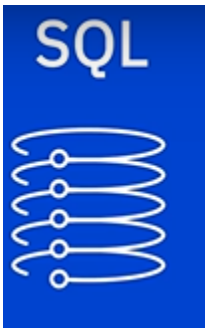
Programming Languages

Programming languages are designed for developing applications and controlling application behavior; for example, Python, R, and Java;

Shell Scripting

Shell and Scripting languages, such as Unix/Linux Shell, and PowerShell, are ideal for repetitive and time-consuming operational tasks.

1. Query Language



SQL, or Structured Query Language, is a querying language designed for accessing and manipulating information from, mostly, though not exclusively, relational databases.

Using SQL, you can:

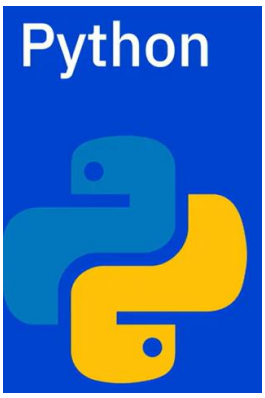
- Insert, update, and delete records in a database
- Create new databases, tables, and views
- Write stored procedures

Advantages of using SQL:

- SQL is portable and platform independent
- Can be used for querying data in a wide variety of databases and data repositories
- Has a simple syntax that is similar to the English language
- Its syntax allows developers to write programs with fewer lines of code using basic keywords
- Can retrieve large amounts of data quickly and efficiently
- Runs on an interpreter system

It runs on an interpreter system, which means code can be executed as soon as it is written, making prototyping quick and easy. SQL is one of the most popular querying language. Due to its large user community and the sheer volume of documentation accumulated over the years, it continues to provide a uniform platform, worldwide, to all its users.

2. Programming Languages



Python is a widely-used open-source, general-purpose, high-level programming language.

Its syntax allows programmers to express their concepts in fewer lines of code

An ideal tool for beginning programmers because of its focus on simplicity and readability

Great for performing high-computational tasks in large volumes of data

Python provides libraries like Numpy and Pandas, which eases this task by the use of parallel processing

Python is a widely-used open-source, general-purpose, high-level programming language.

Has in-built functions for frequently used concepts

Supports multiple programming paradigms – object-oriented, imperative, functional, and procedural

Python is one of the fastest-growing programming languages in the world.

- Easy to learn
- Open-source
- Can be ported to multiple platforms
- Has widespread community support
- Provides open-source libraries for data manipulation, data visualization, statistics, mathematics

Its vast array of libraries and functionalities also include:

- Pandas for data cleaning and analysis
- Numpy and Scipy, for statistical analysis
- Beautifulsoup and Scrapy for web scraping
- Matplotlib and Seaborn to visually represent data in the form of bar graphs, histogram, and pie-charts
- Opency for image processing

R is an open-source programming language and environment for data analysis, data visualization, machine learning, and statistics.

Widely used for:

- Developing statistical software
- Performing data analytics
- Creating compelling visualizations

Key benefits:

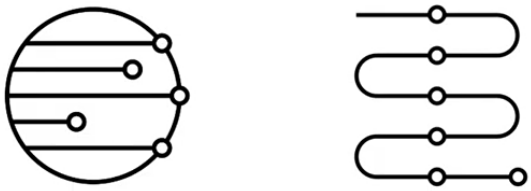
- Open-source
- Platform-independent
- Can be paired with many programming languages
- Highly extensible
- Facilitates the handling of structured and unstructured data
- Includes libraries such as Ggplot2 and Plotly that offer aesthetic graphical plots to its users
- Allows data and scripts to be embedded in reports
- Allows creation of interactive web apps
- Can be used for developing statistical tools

Java is an object-oriented, class-based, and platform-independent programming language originally developed by Sun Microsystems.

- One of the top-ranked programming languages used today
- Used in a number of data analytics processes – cleaning data, importing and exporting data, statistical analysis, data visualization
- Used in the development of big data frameworks and tools – Hadoop, Hive, Spark
- Well-suited for speed-critical projects

3.Shell Scripting

A Unix/Linux Shell is a computer program written for the UNIX shell. It is a series of UNIX commands written in a plain text file to accomplish a specific task.



A Unix/Linux Shell is a computer program written for the UNIX shell. It is a series of UNIX commands written in a plain text file to accomplish a specific task. Writing a shell script is fast and easy. It is most useful for repetitive tasks that may be time-consuming to execute by typing one line at a time.

Typical operations performed by shell scripts include:

- File manipulation
- Program execution
- System administration tasks such as disk backups and evaluating system logs
- Installation scripts for complex programs
- Executing routine backups
- Running batches

PowerShell

PowerShell is a cross-platform automation tool and configuration framework by Microsoft that is optimized for working with structured data formats, such as JSON, CSV, XML, and REST APIs, websites, and office applications.

- Consists of command-line shell and scripting language
- Is object-based and can be used to filter, sort, measure, group, and compare objects as they pass through a data pipeline
- Used for data mining, building GUIs, creating charts, dashboards, and interactive reports

Reading: Summary and Highlights

In this lesson, you have learned the following information:

A data analyst ecosystem includes the infrastructure, software, tools, frameworks, and processes used to gather, clean, analyze, mine, and visualize data.

Based on how well-defined the structure of the data is, data can be categorized as:

- Structured Data, that is data which is well organized in formats that can be stored in databases.
- Semi-Structured Data, that is data which is partially organized and partially free form.
- Unstructured Data, that is data which can not be organized conventionally into rows and columns.

Data comes in a wide-ranging variety of file formats, such as delimited text files, spreadsheets, XML, PDF, and JSON, each with its own list of benefits and limitations of use.

Data is extracted from multiple data sources, ranging from relational and non-relational databases to APIs, web services, data streams, social platforms, and sensor devices.

Once the data is identified and gathered from different sources, it needs to be staged in a data repository so that it can be prepared for analysis. The type, format, and sources of data influence the type of data repository that can be used.

Data professionals need a host of languages that can help them extract, prepare, and analyze data. These can be classified as:

- Querying languages, such as SQL, used for accessing and manipulating data from databases.
- Programming languages such as Python, R, and Java, for developing applications and controlling application behavior.
- Shell and Scripting languages, such as Unix/Linux Shell, and PowerShell, for automating repetitive operational tasks.

Quiz: Practice Quiz

Bookmarked

Question 1

1/1 point (ungraded)

What data type is typically found in databases and spreadsheets?

- Unstructured data
- Structured data
- Social media content
- Semi-structured data

✔

Question 2

1/1 point (ungraded)

Which of these data sources is an example of semi-structured data?

- Network and web logs
- Social media feeds
- Emails
- Documents

✔

Question 3

1/1 point (ungraded)

Which one of the provided file formats is commonly used by APIs and Web Services to return data?

- (●) JSON
- ( ) XML
- ( ) Delimited file
- ( ) XLS

✔

Question 4

1/1 point (ungraded)

What is one example of the relational databases discussed in the video?

- ( ) XML
- ( ) Spreadsheet
- ( ) Flat files
- (●) SQL Server

✔

Question 5

1/1 point (ungraded)

Which of the following languages is one of the most popular querying languages in use today?

- ○ R
- ◉ SQL
- ○ Python
- ○ Java

✔

Quiz: Graded Quiz

Bookmarked

Graded Quiz due Jun 28, 2022 01:42 +08

Question 1

1/1 point (graded)

In the data analyst's ecosystem, languages are classified by type. What are shell and scripting languages most commonly used for?

- ○ Querying data
- ○ Building apps
- ◉ Automating repetitive operational tasks
- ○ Manipulating data

✔

Question 2

1/1 point (graded)

Which of the following is an example of unstructured data?

- ◉ Video and audio files
- ○ XML
- ○ Zipped files
- ○ Spreadsheets

✔

Submit | You have used 1 of 2 attempts | Reset | Show answer

✔ Correct (1/1 point)

Question 3

1/1 point (graded)

Which one of these file formats is independent of software, hardware, and operating systems, and can be viewed the same way on any device?

- ○ XML
- ◉ PDF
- ○ Delimited text file
- ○ XLSX

✔

Question 4

1/1 point (graded)

Which data source can return data in plain text, XML, HTML, or JSON among others?

- ○ PDF
- ○ Delimited text file
- ○ XML
- ◉ API

✔

Question 5

1/1 point (graded)

According to the video "Languages for Data Professionals," which of the programming languages supports multiple programming paradigms, such as object-oriented, imperative, functional, and procedural, making it suitable for a wide variety of use cases?

- ○ PowerShell
- ◉ Python
- ○ Java
- ○ Unix/Linux Shell

✔

Module Introduction

In this module, you will learn about different types of data repositories such as Databases, Data Warehouses, Data Marts, and Data Lakes. You will learn about data pipelines, and the ETL (Extract, Transform, and Load) process, using which data is extracted, transformed, and loaded into data repositories. You will also gain an understanding of Big Data and Big Data processing tools such as Hadoop, Hadoop Distributed File System (HDFS), Hive, and Spark.

Learning Objectives

After completing this module, you will be able to:

- Describe and differentiate between relational and non-relational database management systems.
- 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 and summarize the features and use of some of the Big Data processing tools.

Overview of Data Repositories

A **data repository** is a general term used to refer to data that has been collected, organized, and isolated so that it can be used for business operations or mined for reporting and data analysis. It can be a small or large database infrastructure with one or more databases that collect, manage, and store data sets.

In this video, we will provide an overview of the different types of repositories your data might reside in, such as databases, data warehouses, and big data stores, and examine them in greater detail in further videos.

Let's begin with databases.

A **database i**s a collection of data, or information, designed for the input, storage, search and retrieval, and modification of data.

And a **Database Management System, or DBMS,** is a

- set of programs that creates and maintains the database. It allows you to store, modify, and extract information from the database using a function called querying.
- For example, if you want to find customers who have been inactive for six months or more, using the query function, the database management system will retrieve data of all customers from the database that have been inactive for six months and more.
- Even though a database and DBMS mean different things the terms are often used interchangeably.

There are **different types of databases**. Several **factors influence the choice of database**, such as the data type and structure, querying mechanisms, latency requirements, transaction speeds, and intended use of the data. It's important to mention two main types of databases here—relational and non-relational databases.

Relational databases, also referred to as RDBMSes,

- build on the organizational principles of flat files, with data organized into a tabular format with rows and columns following a well-defined structure and schema.
- However, unlike flat files, RDBMSes are optimized for data operations and querying involving many tables and much larger data volumes.

Structured Query Language, or SQL, is the standard querying language for relational databases. Then we have non-relational databases, also known as NoSQL, or "Not Only SQL".

Non-relational databases emerged in response to the volume, diversity, and speed at which data is being generated today, mainly influenced by advances in cloud computing, the Internet of Things, and social media proliferation.

- Built for speed, flexibility, and scale, non-relational databases made it possible to store data in a schema-less or free-form fashion.

NoSQL is widely used for processing big data. A data warehouse works as a central repository that merges information coming from disparate sources and consolidates it through the extract, transform, and load

process, also known as the ETL process, into one comprehensive database for analytics and business intelligence.

At a very high-level, the ETL process helps you to extract data from different data sources, transform the data into a clean and usable state, and load the data into the enterprise's data repository.

Related to Data Warehouses are the **concepts of Data Marts** and **Data Lakes**, which we will cover later. Data Marts and Data Warehouses have historically been relational, since much of the traditional enterprise data has resided in RDBMSes. However, with the emergence of NoSQL technologies and new sources of data, non-relational data repositories are also now being used for Data Warehousing.

Another category of data repositories are **Big Data Stores**, that include distributed computational and storage infrastructure to store, scale, and process very large data sets. Overall, data repositories help to isolate data and make reporting and analytics more efficient and credible while also serving as a data archive.

Overview of Data Repositories

Introduction

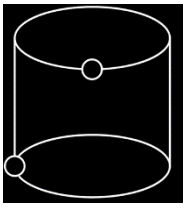
A data repository is a general term used to refer to data that has been collected, organized, and isolated so that it can be used for business operations or mined for reporting and data analysis.

It can be a small or large database infrastructure with one or more databases that collect, manage, and store data sets.

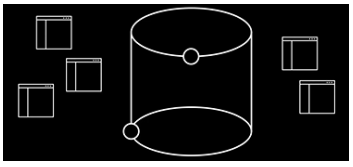
Types of data repositories Include:

- Databases
- Data Warehouses
- Big Data Stores

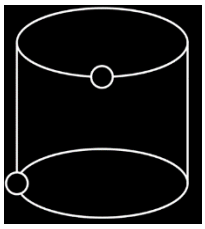
Databases



A database is a collection of data, or information, designed for the input, storage, search and retrieval, and modification of data.



Database Management System, or DBMS, is a set of programs that creates and maintains the database. It allows you to store, modify, and extract information from the database using a function called querying.

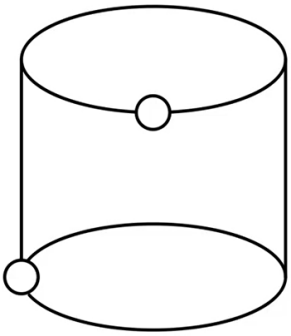


Even though a database and DBMS mean different things the terms are often used interchangeably.

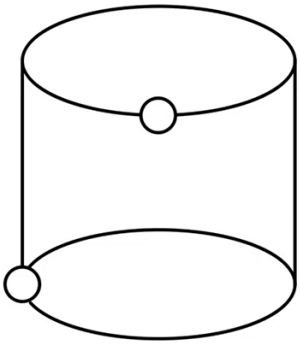
Factors governing choice of database include:

- Data type
- Data structure
- Querying mechanisms
- Latency requirements
- Transaction speed
- Intended use of the data

Two Types of Databases:

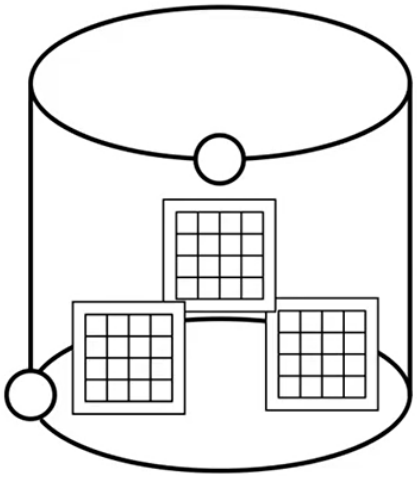


Relational



Non-relational

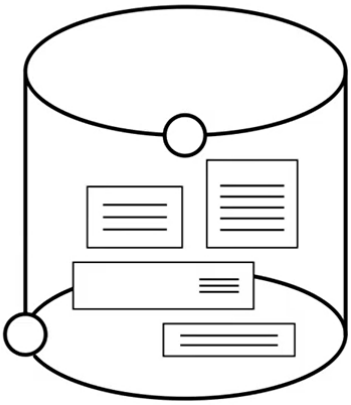
Relational Databases:



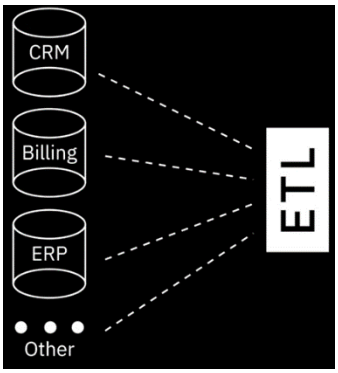
- Data is organized into a tabular format with rows and columns
- Well-defined structure and schema
- Optimized for data operations and querying
- Use SQL as the standard querying language

Non- Relational Databases:

- Emerged in response to the volume, diversity, and speed at which data is being generated today
- Built for speed, flexibility, and scale
- Data can be stored in a schema-less form
- Widely used for processing big data



Data Warehouse:

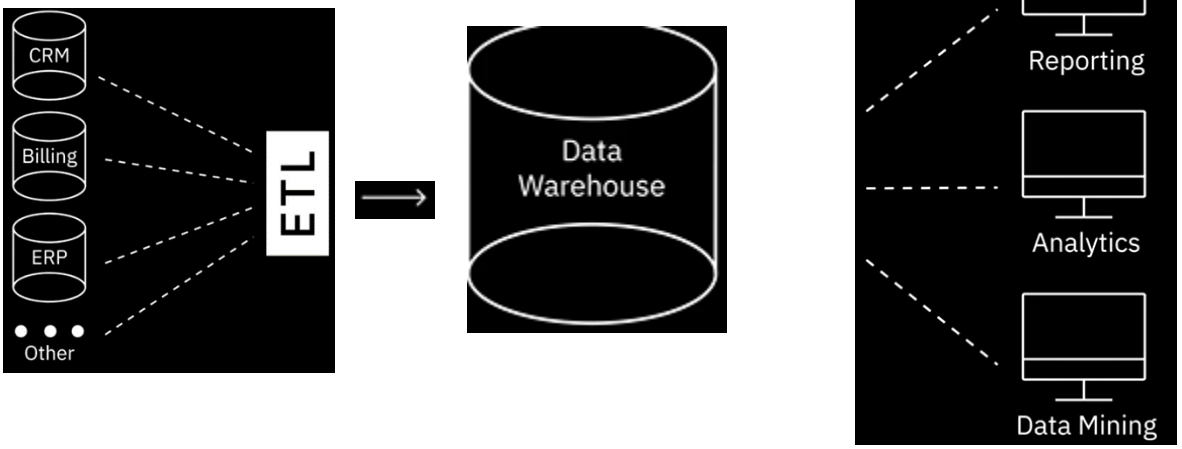


Consolidates the data through the extract, transform, and load process, also known as the ETL process, into one comprehensive database for analytics and business intelligence.

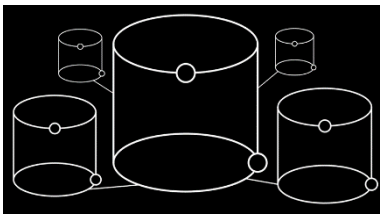
the ETL process helps you to extract data from different data sources;

transform the data into a clean and

usable state, and load the data into the enterprise's data repository.



Big Data Stores



distributed computational and storage infrastructure to store, scale, and process very large datasets

Summary:

Data repositories help to isolate data and make reporting and analytics more efficient and credible while also serving as a data archive.

RDBMS

A **relational database** is a

- collection of data organized into a table structure, where the tables can be linked, or related, based on data common to each. Tables are made of rows and columns, where rows are the “records”, and the columns the “attributes”.

Let’s take the example of a customer table that maintains data about each customer in a company.

The columns, or attributes, in the customer table are the Company ID, Company Name, Company Address, and Company Primary Phone; and Each row is a customer record.

Now let’s understand what we mean by tables being linked, or related, based on data common to each. Along with the customer table, the company also maintains transaction tables that contain data describing multiple individual transactions pertaining to each customer.

The **columns** for the transaction table might include the Transaction Date, Customer ID, Transaction Amount, and Payment Method.

The customer table and the transaction tables can be related based on the common Customer ID field. You can query the customer table to produce reports such as a customer statement that consolidates all transactions in a given period.

This capability of relating tables based on common data enables you to retrieve an entirely new table from data in one or more tables with a single query. It also allows you to understand the relationships among all available data and gain new insights for making better decisions.

- Relational databases use structured query language, or SQL, for querying data. We’ll learn more about SQL later in this course.

- Relational databases build on the organizational principles of flat files such as spreadsheets, with data organized into rows and columns following a well-defined structure and schema.

But this is where the similarity ends.

- Relational databases, by design, are ideal for the optimized storage, retrieval, and processing of data for large volumes of data, unlike spreadsheets that have a limited number of rows and columns. Each table in a relational database has a unique set of rows and columns and relationships can be defined between tables, which minimizes data redundancy. Moreover, you can restrict database fields to specific data types and values, which minimizes irregularities and leads to greater consistency and data integrity.
- Relational databases use SQL for querying data, which gives you the advantage of processing millions of records and retrieving large amounts of data in a matter of seconds. Moreover, the security architecture of relational databases provides controlled access to data and also ensures that the standards and policies for governing data can be enforced. Relational databases range from small desktop systems to massive cloud-based systems.

They can be either:

open-source and internally supported,

open-source with commercial support, or

commercial closed-source systems.

IBM DB2, Microsoft SQL Server, MySQL, Oracle Database, and PostgreSQL are some of the popular

relational databases.

Cloud-based relational databases, also referred to as Database-as-a-Service, are gaining wide

use as they have access to the limitless compute and storage capabilities offered by the cloud.

Some of the popular cloud relational databases include Amazon Relational Database Service

(RDS), Google Cloud SQL, IBM DB2 on Cloud, Oracle Cloud, and SQL Azure.

- RDBMS is a mature and well-documented technology, making it easy to learn and find qualified talent.

One of the most significant advantages of the relational database approach is

- its ability to create meaningful information by joining tables. Some of its other advantages include:
- Flexibility: Using SQL, you can add new columns, add new tables, rename relations, and make other changes while the database is running and queries are happening.

- Reduced redundancy: Relational databases minimize data redundancy. For example, the information of a customer appears in a single entry in the customer table, and the transaction table pertaining to the customer stores a link to the customer table.
- Ease of backup and disaster recovery: Relational databases offer easy export and import options, making backup and restore easy.
- Exports can happen while the database is running, making restore on failure easy.
- Cloud-based relational databases do continuous mirroring, which means the loss of data on restore can be measured in seconds or less.
- ACID-compliance: ACID stands for Atomicity, Consistency, Isolation, and Durability. And ACID compliance implies that the data in the database remains accurate and consistent despite failures, and database transactions are processed reliably.

Now we'll look at some use cases for relational databases:

- Online Transaction Processing: OLTP applications are focused on transaction-oriented tasks that run at high rates. Relational databases are well suited for OLTP applications because they can accommodate a large number of users;
- they support the ability to insert, update, or delete small amounts of data; and
- they also support frequent queries and updates as well as fast response times.

Data warehouses:

In a data warehousing environment, relational databases can be optimized for online analytical processing (or OLAP), where historical data is analyzed for business intelligence.

IoT solutions: Internet of Things (IoT) solutions require speed as well as the ability to collect and process data from edge devices, which need a lightweight database solution.

This brings us to the limitations of RDBMS:

- RDBMS does not work well with semi-structured and unstructured data and is, therefore, not suitable for extensive analytics on such data.
- For migration between two RDBMSs, schemas and type of data need to be identical between the source and destination tables.
- Relational databases have a limit on the length of data fields, which means if you try to enter more information into a field than it can accommodate, the information will not be stored.

Despite the limitations and the evolution of data in these times of big data, cloud computing, IoT devices, and social media, RDBMS continues to be the predominant technology for working with structured data.

RDBMS

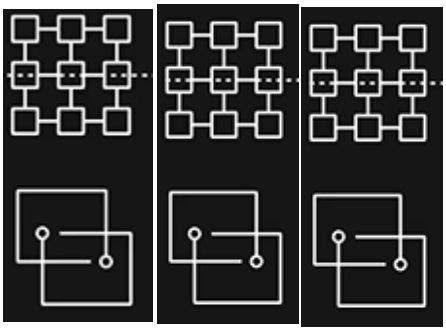
What is a relational database?

A relational database is a collection of data organized into a table structure, where the tables can be linked, or related, based on data common to each.

Customer ID	Customer Name	Customer Address	Customer Phone
01234	Jim H.	------	----------
02345	Pam B.	------	----------

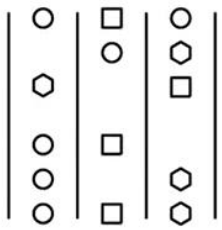
Transaction Date	Customer ID	Transaction Amount	Payment Method
------	01234	----------	------
----------	02345	------	----------

Along with the customer table, the company also maintains transaction tables that contain data describing multiple individual transactions pertaining to each customer. The columns for the transaction table might include the Transaction Date, Customer ID, Transaction Amount, and Payment Method. The customer table and the transaction tables can be related based on the common Customer ID field. You can query the customer table to produce reports such as a customer statement that consolidates all transactions in a given period.



This capability of relating tables based on common data enables you to retrieve an entirely new table from data in one or more tables with a single query.

Relational databases use structured query language, or SQL, for querying data.

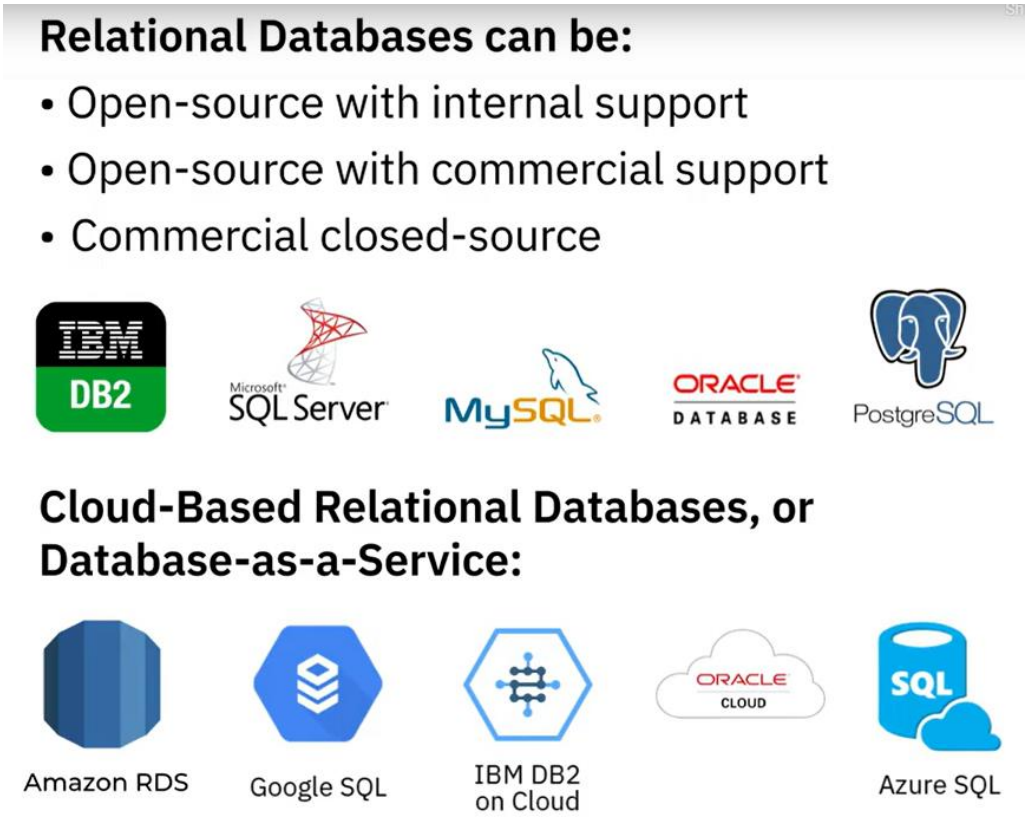


Similarities between relational databases and spreadsheets:
Relational databases build on the organizational principles of flat files such as spreadsheets, with data organized into rows and columns following a well-defined structure and schema.

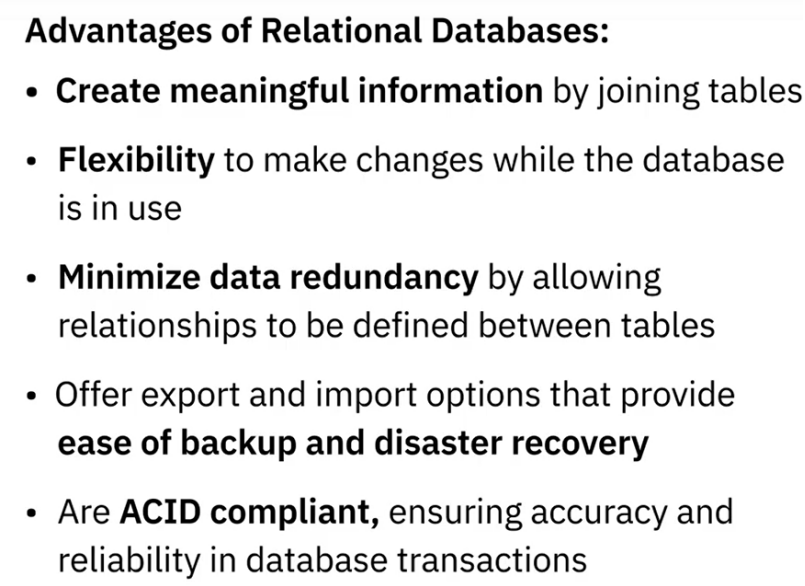
- Ideal for the optimized storage, retrieval, and processing of data for large volumes of data
- Each table has a unique set of rows and columns
- Relationships can be defined between tables
- Fields can be restricted to specific data types and values
- Can retrieve millions of records in seconds using SQL for querying data
- Security architecture of relational databases provides greater access control and governance

Examples of Relational Databases

Relational databases range from small desktop systems to massive cloud-based systems.



Advantages of the Relational Database Approach



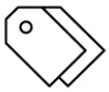
ACID stands for **Atomicity, Consistency, Isolation, and Durability.**

And ACID compliance implies that the data in the database remains accurate and consistent despite failures, and database transactions are processed reliably.

Use Cases for RDBMS

1.

Relational Databases are well suited for:



Online Transaction Processing (OLTP) application
Can support transaction-oriented tasks that run at high rates and
- Accommodate large number of users
- Manage small amounts of data
- Support frequent queries and fast

2.

Relational Databases are well suited for:

Data Warehouses
Can be optimized for online analytical processing (OLAP)

IoT Solutions
Provide the speed and ability to collect and process data from edge devices

Limitations of RDBMS

Limitations of RDBMS:
- Does not work well with semi-structured and unstructured data
- Migration between two RDBMS's is possible only when the source and destination tables have identical schemas and data types
- Entering a value greater than the defined length of a data field results in loss of information

No SQL

NoSQL, which stands for “not only SQL,” or sometimes “non SQL” is a non-relational database design that provides flexible schemas for the storage and retrieval of data.

NoSQL databases have existed for many years but have only recently become more popular in the era of cloud, big data, and high-volume web and mobile applications. They are chosen today for their attributes around scale, performance, and ease of use. It's important to emphasize that the "No" in "NoSQL" is an abbreviation for "not only" and not the actual word "No."

NoSQL databases are built for specific data models and have flexible schemas that allow programmers to create and manage modern applications. They do not use a traditional row/column/table database design with fixed schemas, and typically not use the structured query language (or SQL) to query data, although some may support SQL or SQL-like interfaces.

NoSQL allows data to be stored in a schema-less or free-form fashion. Any data, be It structured, semi-structured, or unstructured, can be stored in any record. Based on the model being used for storing data, there are four common types of NoSQL databases.

Key-value store, document-based, column-based, and graph-based.

Key-value store.

- Data in a key-value database is stored as a collection of key-value pairs. The key represents an attribute of the data and is a unique identifier. Both keys and values can be anything from simple integers or strings to complex JSON documents.
- Key-value stores are great for storing user session data and user preferences, making real-time recommendations and targeted advertising, and in-memory data caching. However, if you want to be able to query the data on specific data value, need relationships between data values, or need to have multiple unique keys, a key-value store may not be the best fit. Redis, Memcached, and DynamoDB are some well-known examples in this category.

Document-based:

- Document databases store each record and its associated data within a single document. They enable flexible indexing, powerful ad hoc queries, and analytics over collections of documents.
- Document databases are preferable for eCommerce platforms, medical records storage, CRM platforms, and analytics platforms. However, if you’re looking to run complex search queries and multi-operation transactions, a document-based database may not be the best option for you. MongoDB, DocumentDB, CouchDB, and Cloudant are some of the popular document-based databases.

Column-based:

Column-based models store data in cells grouped as columns of data instead of rows. A logical grouping of columns, that is, columns that are usually accessed together, is called a column family.

For example, a customer's name and profile information will most likely be accessed together but not their purchase history. So, customer name and profile information data can be grouped into a column family. Since column databases store all cells corresponding to a column as a continuous disk entry, accessing and searching the data becomes very fast.

Column databases can be great for systems that require heavy write requests, storing time-series data, weather data, and IoT data. But if you need to use complex queries or change your querying patterns frequently, this may not be the best option for you. The most popular column databases are Cassandra and HBase.

Graph-based:

Graph-based databases use a graphical model to represent and store data. They are particularly useful for visualizing, analyzing, and finding connections between different pieces of data.

The circles are nodes, and they contain the data. The arrows represent relationships. Graph databases are an excellent choice for working with connected data, which is data that contains lots of interconnected relationships.

Graph databases are great for social networks, real-time product recommendations, network diagrams, fraud detection, and access management.

But if you want to process high volumes of transactions, it may not be the best choice for you, because graph databases are not optimized for large-volume analytics queries. Neo4J and CosmosDB are some of the more popular graph databases.

NoSQL was created in response to the limitations of traditional relational database technology. The primary advantage of NoSQL is its ability to handle large volumes of structured, semi-structured, and unstructured data.

Some of its other advantages include:

The ability to run as distributed systems scaled across multiple data centers, which enables them to take advantage of cloud computing infrastructure;

An efficient and cost-effective scale-out architecture that provides additional capacity and performance with the addition of new nodes; and

Simpler design, better control over availability, and improved scalability that enables you to be more agile, more flexible, and to iterate more quickly.

To summarize the key differences between relational and non-relational databases:

RDBMS schemas rigidly define how all data inserted into the database must be typed and composed, whereas NoSQL databases can be schema-agnostic, allowing unstructured and semi-structured data to be stored and manipulated.

Maintaining high-end, commercial relational database management systems is expensive whereas NoSQL databases are specifically designed for low-cost commodity hardware.

Relational databases, unlike most NoSQL, support ACID-compliance, which ensures reliability of transactions and crash recovery.

RDBMS is a mature and well-documented technology, which means the risks are more or less perceivable as compared to NoSQL, which is a relatively newer technology.

Nonetheless, NoSQL databases are here to stay, and are increasingly being used for mission critical applications.

No SQL (Not only SQL)

What is a NoSQL database?

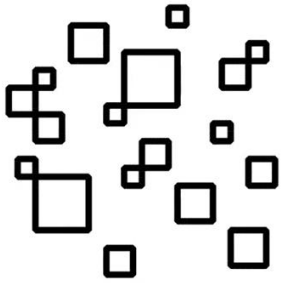
NoSQL, which stands for “not only SQL,” or sometimes “non SQL” is a non-relational database design that provides flexible schemas for the storage and retrieval of data.

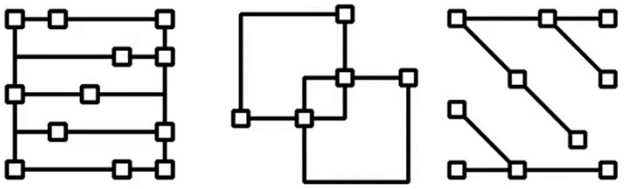
It's important to emphasize that the "No" in "NoSQL" is an abbreviation for "not only" and not the actual word "No."

- NoSQL databases are built for specific data models
- Have flexible schemas that allow programmers to create and manage modern applications.
- They do not use a traditional row/column/table database design with fixed schemas, and
- Typically not use the structured query language (or SQL) to query data, although some may support SQL or SQL-like interfaces.

NoSQL allows data to be stored in
a schema-less or free-form fashion.

NoSQL allows data to be stored in
a schema-less or free-form fashion.





NoSQL allows data to be stored in a schema-less or free-form fashion. Any data, be It structured, semi-structured, or unstructured, can be stored in any record.

Four common types of NoSQL databases:

Based on the model being used for storing data, there are four common types of NoSQL databases:

- Key-value store
- Document Based
- Column Based
- Graph Based

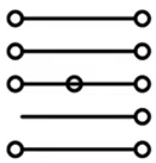
1. Key-value Store

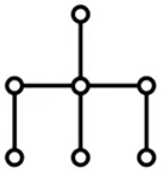
Key-value store:

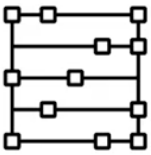
- Data in a key-value database is stored as a collection of key-value pairs.
- A key represents an attribute of the data and is a unique identifier.
- Both keys and values can be anything from simple integers or strings to complex JSON documents.
- Great for storing user session data, user preferences, real-time recommendations, targeted advertising, in-memory data caching.

Not a great fit if you want to:

- Query data on specific data value
- Need relationships between data values
- Need multiple unique keys







Examples:

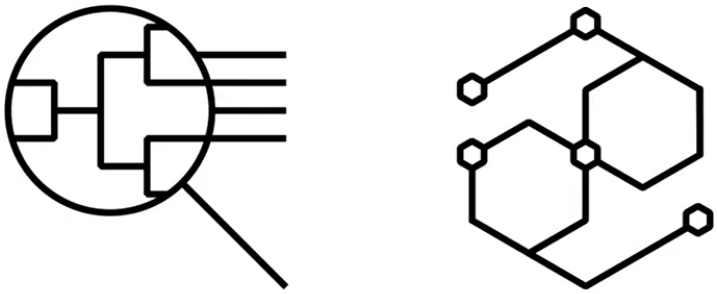
Redis Memcached DynamoDB

2. Document Based

- Document databases store each record and its associated data within a single document.
- They enable flexible indexing, powerful ad hoc queries, and analytics over collections of documents.
- Preferred for eCommerce platforms, medical records storage, CRM platforms, and analytics platforms.

Not a great fit if you want to:

- Run complex search queries
- Perform multi-operation transactions



Popular document-based databases:

MongoDB DocumentDB CouchDB Cloudant

3. Column-based

- Data is stored in cells grouped as columns of data instead of rows.
- A logical grouping of columns is referred to as a column family.

- All cells corresponding to a column are saved as a continuous disk entry, making access and search easier and faster.
- Great for systems that require heavy write requests, storing time-series data, weather data, and IoT data.

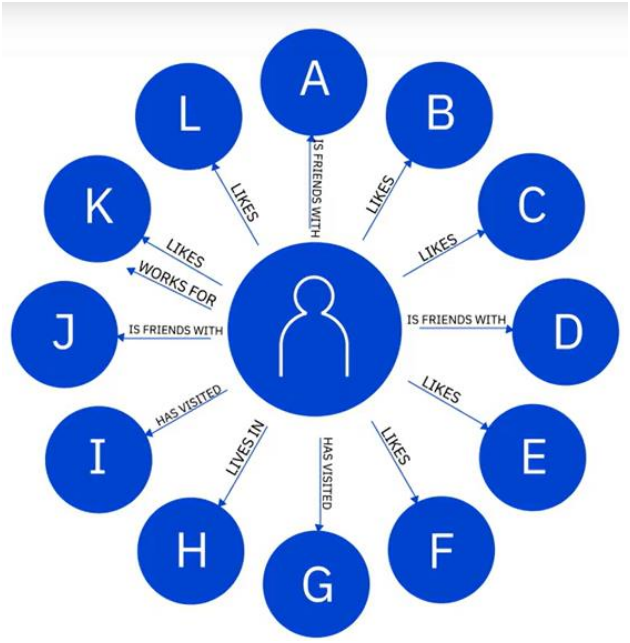
Not a great fit if you want to:

- Run complex queries
- Change querying patterns frequently

Popular Column databases:

3. Graph-based

- Graph-based databases use a graphical model to represent and store data.
- Useful for visualizing, analyzing, and finding connections between different pieces of data.



An excellent choice for working with connected data

The circles are nodes, and they contain the data. The arrows represent relationships.

Great for:

Social networks

Product recommendations

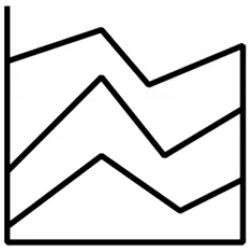
Network diagrams

Fraud detection

Access management

Not a great fit if you want to:

- Process high volumes of transactions



But if you want to process high volumes of transactions, it may not be the best choice for you, because graph databases are not optimized for large-volume analytics queries.

Popular Graph based databases:

Neo4J

CosmosDB

Advantages of NoSQL:

- Its ability to handle large volumes of structured, semi-structured, and unstructured data
- Its ability to run as a distributed system scaled across multiple data centers
- An efficient and cost-effective scale-out architecture that provides additional capacity and performance with the addition of new nodes
- Simpler design, better control over availability, and improved scalability that makes it agile, flexible, and support quick iterations

Key differences

Relational databases

- RDBMS schemas rigidly define how all data inserted into the database must be typed and composed
- Maintaining high-end, commercial relational database management systems can be expensive
- Support ACID-compliance, which ensures reliability of transactions and crash recovery
- A mature and well-documented technology, which means the risks are more or less perceivable

Non-Relational databases

- NoSQL databases can be schema-agnostic, allowing unstructured and semi-structured data to be stored and manipulated
- Specifically designed for low-cost commodity hardware
- Most NoSQL databases are not ACID compliant
- A relatively newer technology

Data Marts, Data Lakes, ETL, and Data Pipelines

Earlier in the course, we examined databases, data warehouses, and big data stores. Now we'll go a little deeper in our exploration of data warehouses, data marts, and data lakes; and also learn about the ETL process and data pipelines.

A data warehouse works like a multi-purpose storage for different use cases. By the time the data comes into the warehouse, it has already been modeled and structured for a specific purpose, meaning it is analysis ready.

As an organization, you would opt for a data warehouse when you have massive amounts of data from your operational systems that needs to be readily available for reporting and analysis.

Data warehouses serve as the single source of truth—storing current and historical data that has been cleansed, conformed, and categorized. A data warehouse is a multi-purpose enabler of operational and performance analytics.

A data mart is a sub-section of the data warehouse, built specifically for a particular business function, purpose, or community of users. The idea is to provide stakeholders data that is most relevant to them, when they need it.

For example, the sales or finance teams accessing data for their quarterly reporting and projections. Since a data mart offers analytical capabilities for a restricted area of the data warehouse, it offers isolated security and isolated performance. The most important role of a data mart is business-specific reporting and analytics.

A Data Lake is a storage repository that can store large amounts of structured, semi-structured, and unstructured data in their native format, classified and tagged with metadata. So, while a data warehouse stores data processed for a specific need, a data lake is a pool of raw data where each data element is given a unique identifier and is tagged with metatags for further use.

You would opt for a data lake if you generate, or have access to, large volumes of data on an ongoing basis, but don't want to be restricted to specific or pre-defined use cases. Unlike data warehouses, a data lake would retain all source data, without any exclusions. And the data could include all types of data sources and types.

Data lakes are sometimes also used as a staging area of a data warehouse. The most important role of a data lake is in predictive and advanced analytics. Now we come to the process that is at the heart of gaining value from data—the Extract, Transform, and Load process, or ETL. ETL is how raw data is converted into analysis-ready data.

It is an automated process in which you gather raw data from identified sources, extract the information that aligns with your reporting and analysis needs, clean, standardize, and transform that data into a format that is usable in the context of your organization; and load it into a data repository.

While ETL is a generic process, the actual job can be very different in usage, utility, and complexity. Extract is the step where data from source locations is collected for transformation.

Data extraction could be through: Batch processing, meaning source data, is moved in large chunks from the source to the target system at scheduled intervals.

Tools for batch processing include Stitch and Blendo. Stream processing, which means source data is pulled in real-time from the source and transformed while it is in transit and before it is loaded into the data repository. Tools for stream processing include Apache Samza, Apache Storm, and Apache Kafka.

Transform involves the execution of rules and functions that converts raw data into data that can be used for analysis. For example, making date formats and units of measurement consistent across all sourced data, removing duplicate data, filtering out data that you do not need, enriching data, for example, splitting full name to first, middle, and last names, establishing key relationships across tables, applying business rules and data validations.

Load is the step where processed data is transported to a destination system or data repository. It could be: Initial loading, that is, populating all the data in the repository, Incremental loading, that is, applying ongoing updates and modifications as needed periodically; or Full refresh, that is, erasing contents of one or more tables and reloading with fresh data.

Load verification, which includes data checks for missing or null values, server performance, and monitoring load failures, are important parts of this process step. It is vital to keep an eye on load failures and ensure the right recovery mechanisms are in place.

ETL has historically been used for batch workloads on a large scale. However, with the emergence of streaming ETL tools, they are increasingly being used for real-time streaming event data as well. It’s common to see the terms ETL and data pipelines used interchangeably.

And although both move data from source to destination, data pipeline is a broader term that encompasses the entire journey of moving data from one system to another, of which ETL is a subset.

Data pipelines can be architected for batch processing, for streaming data, and a combination of batch and streaming data. In the case of streaming data, data processing or transformation, happens in a continuous flow.

This is particularly useful for data that needs constant updating, such as data from a sensor monitoring traffic. A data pipeline is a high performing system that supports both long-running batch queries and smaller interactive queries.

The destination for a data pipeline is typically a data lake, although the data may also be loaded to different target destinations, such as another application or a visualization tool.

There are a number of data pipeline solutions available, most popular among them being Apache

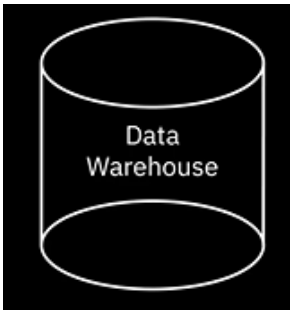
Data Marts, Data Lakes, ETL, and Data Pipelines

Introduction

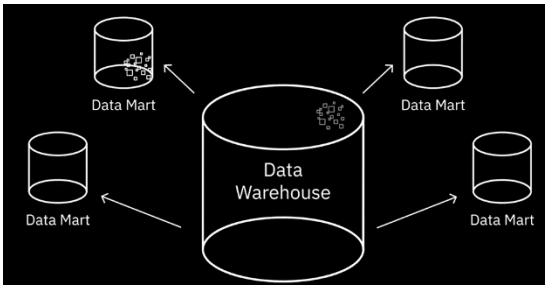
Databases, Data Warehoused and Big Data Stores

Data Warehouses, Data Marts, Data Lakes, ETL Process and Data Pipelines

A data warehouse works like a multi-purpose storage for different use cases. By the time the data comes into the warehouse, it has already been modeled and structured for a specific purpose, meaning it is analysis ready.

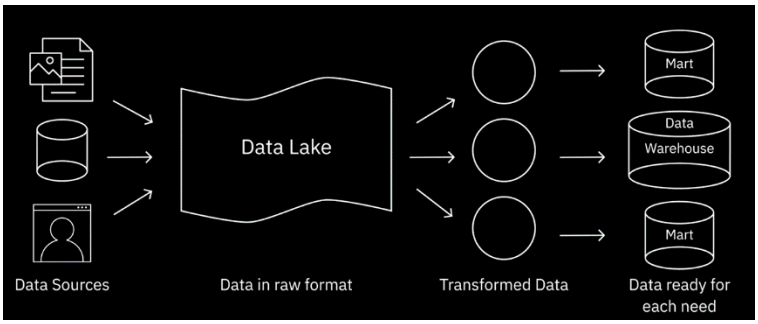


A data warehouse is a multi-purpose enabler of operational and performance analytics.



A data mart is a sub-section of the data warehouse, built specifically for a particular business function, purpose, or community of users. The idea is to provide stakeholders data that is most relevant to them, when they need it. For example, the sales or finance teams accessing data for their quarterly reporting and projections. Since a data mart offers analytical capabilities for a

restricted area of the data warehouse, it offers isolated security and isolated performance. The most important role of a data mart is business-specific reporting and analytics.



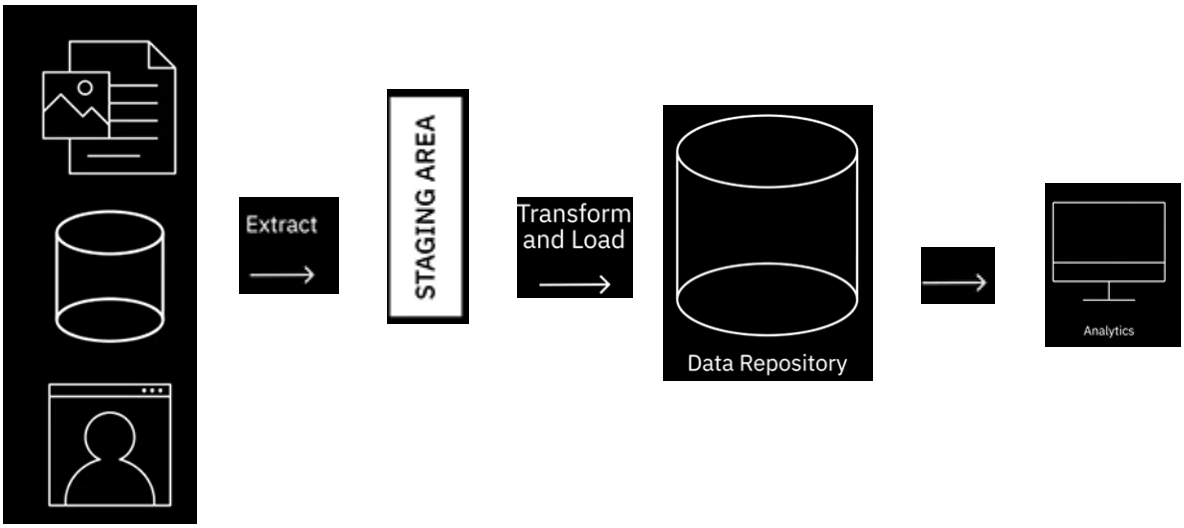
A Data Lake is a storage repository that can store large amounts of structured, semi-structured, and unstructured data in their native format, classified and tagged with metadata. So, while a data warehouse stores data processed for a specific need, a data lake is a pool of raw data where each data element is given a unique identifier and is tagged with metatags for further use. You would opt for a data lake if you generate, or have access to, large volumes of data on an ongoing basis, but don't want to be restricted to specific or pre-defined use cases. Unlike data warehouses, a data lake would retain all source data, without any exclusions. And the data could include all types of data sources and types. Data lakes are sometimes also used as a staging area of a data warehouse. The most important role of a data lake is in predictive and advanced analytics.

Extract, Transform and Load (ETL) Process

ETL is how raw data is converted into analysis-ready data.

- It is an automated process in which you gather raw data from identified sources,
- extract the information that aligns with your reporting and analysis needs,
- clean, standardize, and transform that data into a format that is usable in the context of your organization; and
- load it into a data repository

While ETL is a generic process, the actual job can be very different in usage, utility, and complexity.



Data extraction could be through:

- Batch processing, meaning source data, is moved in large chunks from the source to the target system at scheduled intervals. Tools for batch processing include Stitch and Blendo.

- Stream processing, which means source data is pulled in real-time from the source and transformed while it is in transit and before it is loaded into the data repository. Tools for stream processing include Apache Samza, Apache Storm, and Apache Kafka.

Transform involves the execution of rules and functions that converts raw data into data that can be used for analysis.

For example, making date formats and units of measurement consistent across all sourced data, removing duplicate data, filtering out data that you do not need, enriching data, for example, splitting full name to first, middle, and last names, establishing key relationships across tables, applying business rules and data validations.

Load is the step where processed data is transported to a destination system or data repository.

It could be:

- Initial loading, that is, populating all the data in the repository,
- Incremental loading, that is, applying ongoing updates and modifications as needed periodically;
- Full refresh, that is, erasing contents of one or more tables and reloading with fresh data.

Load verification, which includes data checks for missing or null values, server performance, and monitoring load failures, are important parts of this process step. It is vital to keep an eye on load failures and ensure the right recovery mechanisms are in place.

ETL has historically been used for batch workloads on a large scale. However, with the emergence of streaming ETL tools, they are increasingly being used for real-time streaming event data as well.

It’s common to see the terms ETL and data pipelines used interchangeably. And although both move data from source to destination, data pipeline is a broader term that encompasses the entire journey of moving data from one system to another, of which ETL is a subset.

Data Pipeline

Data pipelines can be architected for batch processing, for streaming data, and a combination of batch and streaming data.

In the case of streaming data, data processing or transformation, happens in a continuous flow. This is particularly useful for data that needs constant updating, such as data from a sensor monitoring traffic.

A data pipeline is a high performing system that supports both long-running batch queries and smaller interactive queries. The destination for a data pipeline is typically a data lake, although the data may also be loaded to different target destinations, such as another application or a visualization tool. There are a number of data pipeline solutions available, most popular among them being Apache Beam and DataFlow.

- that encompasses the entire journey of moving data from one system to another, of which ETL is a subset.
- Data pipelines can be architected for batch processing, for streaming data, and a combination of batch and streaming data. In the case of streaming data, data processing or transformation, happens in a continuous flow. This is particularly useful for data that needs constant updating, such as data from a sensor monitoring traffic.
- A data pipeline is a high performing system that supports both long-running batch queries and smaller interactive queries.
- The destination for a data pipeline is typically a data lake, although the data may also be loaded to different target destinations, such as another application or a visualization tool.
- There are a number of data pipeline solutions available, most popular among them being Apache Beam and DataFlow.

Viewpoints: Considerations for choice of Data Repository

- A number of different factors influence the selection of the right data repository
- Type of data – structured, semi structured or unstructured
- Schema of the data
- Performance requirements
- Whether you're working with data at rest or streaming data in motion)
- Data need to be encrypted
- Volume of data and whether you need a Big Data System
- Storage requirement
    - Frequency of data access
    - Keep in vault for a long time
- Standards set by your organization on the databases and data repositories that can be used

Considerations for choice of Data Repository

- Capacity the data repository is required to handle
- Type of access:
  - At short intervals
  - Run long-running queries
- Purpose of data repository:
  - Transactional
  - Analytical
  - Archival
  - Data Warehousing
- Compatibility of the data repository with the existing ecosystem of programming language tools and processes
- Security features of the data repository
- Scalability from a long term perspective

Considerations for choice of Data Repository

- Very few organizations use one data repository
- We have preferred enterprise relational database, an open source relational database, and unstructured data source
- It's important to think about the skills you have or want to foster
- Cost of various solutions
- Hosting platform is an important consideration – AWS RDS, Amazon's Aurora, Google's relational offerings
- How data needs to be stored
- How data needs to be retrieved
- Where the data should be stored

Considerations for choice of Data Repository

- Structure of data
- Nature of the application volume of data being ingested
- Volume of data being ingested
- Depending on the use case a relational database may not be a good fit

- For ingesting large volumes:
- AWS Document stores such as Mongo DB
    - Wide column- stores such as Cassandra
    - For product recommendation engine or network of people on social media, graph data structures such as Neo4J and apache TinkerPop
- For mining data for analytics Hadoop engine with MapReduce may be a good fit

Foundations of Big Data

In this digital world, everyone leaves a trace. From our travel habits to our workouts and entertainment, the increasing number of internet connected devices that we interact with on a daily basis record vast amounts of data about us there's even a name for it Big Data.

Ernst and Young offers the following definition:

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 drive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value. There is no one definition of big data but there are certain elements that are common across the different definitions, such as **velocity, volume, variety, veracity, and value. These are the V's of big data**

Velocity is the speed at which data accumulates. Data is being generated extremely fast in a process that never stops. Near or real-time streaming, local, and cloud-based technologies can process information very quickly.

Volume is the scale of the data or the increase in the amount of data stored. Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure.

Variety

is the diversity of the data.

- Structured data fits neatly into rows and columns in relational databases, while unstructured data is not organized in a predefined way like tweets, blog posts, pictures, numbers, and video.
- Variety also reflects that data comes from different sources; machines, people, and processes, both internal and external to organizations. Drivers are mobile technologies social media, wearable technologies, geo technologies video, and many, many more.

Veracity

- is the quality and origin of data and its conformity to facts and accuracy.
- Attributes include consistency, completeness, integrity, and ambiguity.
- Drivers include cost and the need for traceability. With the large amount of data available, the debate rages on about the accuracy of data in the digital age. Is the information real or is it false?

Value

- is our ability and need to turn data into value.
- Value isn't just profit. It may have medical or social benefits, as well as customer, employee or personal satisfaction.

- The main reason that people invest time to understand big data is to derive value from it. Let's look at some examples of the V's in action.

1. Velocity. Every 60 seconds, hours of footage are uploaded to YouTube, which is generating data. Think about how quickly data accumulates over hours, days, and years.

2.Volume.

- The world population is approximately 7 billion people and the vast majority are now using digital devices. Mobile phones, desktop and laptop computers, wearable devices, and so on.
- These devices all generate, capture, and store data approximately 2.5 quintillion bytes every day. That's the equivalent of 10 million blu-ray DVDs.

3.Variety.

Let's think about the different types of data. Text, pictures, film, sound, health data from wearable devices, and many different types of data from devices connected to the internet of things.

4. Veracity

- Eighty percent of data is considered to be unstructured and we must devise ways to produce reliable and accurate insights. The data must be categorized, analyzed, and visualized.

Data scientists, today, derive insights from big data and cope with the challenges that these massive data sets present.

The scale of the data being collected means that it's not feasible to use conventional data analysis tools, however, alternative tools that leverage distributed computing power can overcome this problem.

Tools such as Apache Spark, Hadoop, and its ecosystem provides ways to extract, load, analyze, and process the data across distributed compute resources, providing new insights and knowledge. This gives organizations more ways to connect with their customers and enrich the services they offer.

So next time you strap on your smartwatch, unlock your smartphone, or track your workout, remember your data is starting a journey that might take it all the way around the world, through big data analysis and back to you.

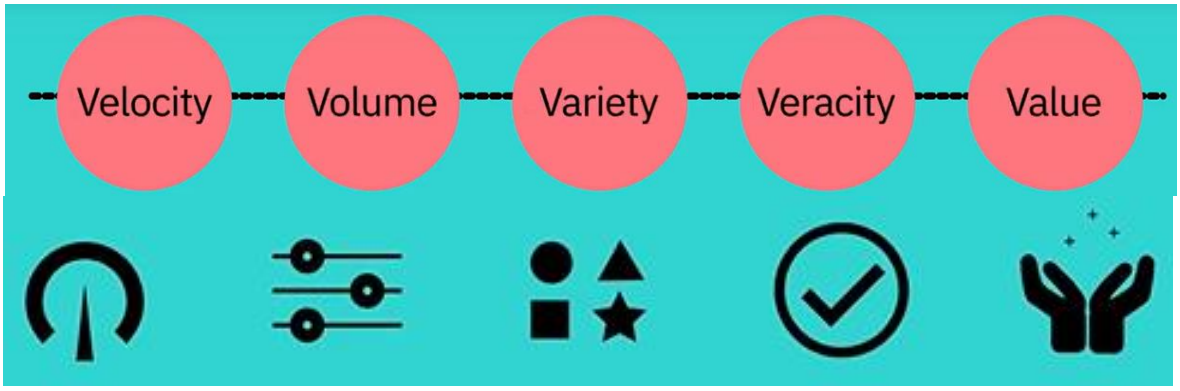
Foundations of Big Data

Big Data

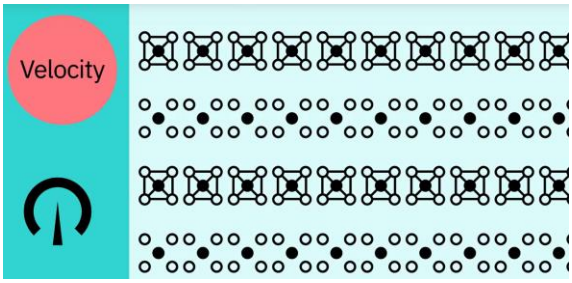
"Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate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

Ernst and Young

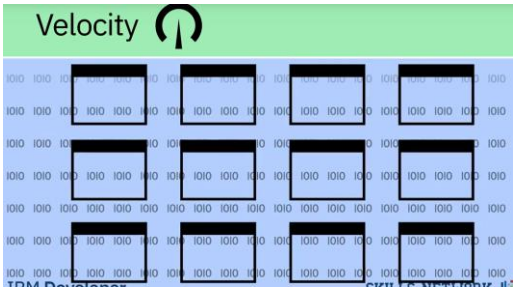
The V's of Big Data



1. Velocity

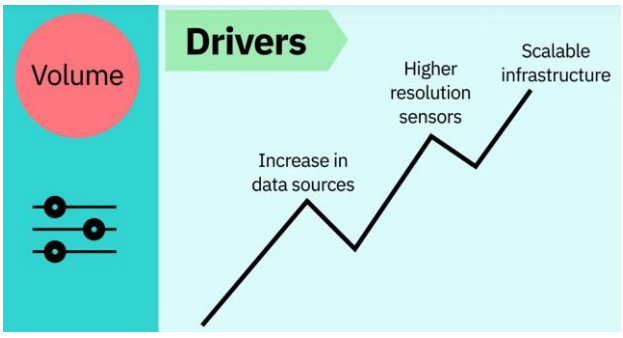


Velocity is the speed at which data accumulates. Data is being generated extremely fast in a process that never stops. Near or real-time streaming, local, and cloud-based technologies can process information very quickly.

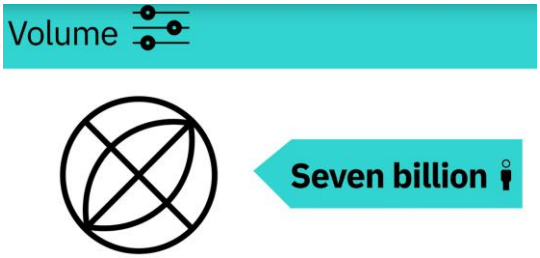


Every 60 seconds, hours of footage are uploaded to YouTube, which is generating data. Think about how quickly data accumulates over hours, days, and years.

2.Volume

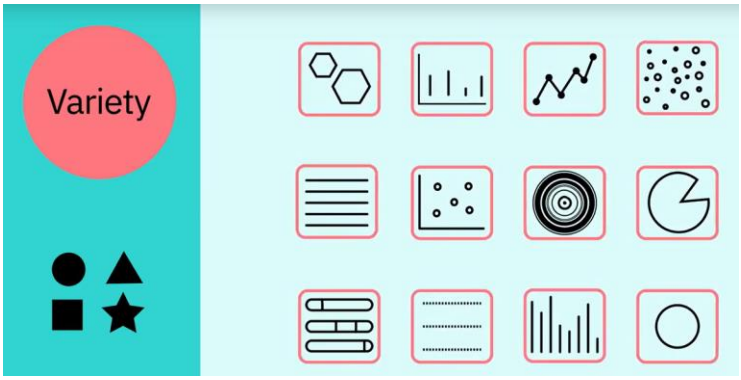


Volume is the scale of the data or the increase in the amount of data stored. Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure.



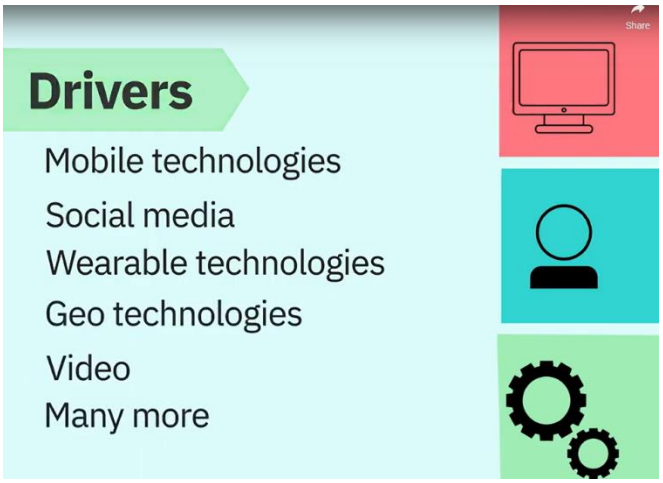
- The world population is approximately 7 billion people and the vast majority are now using digital devices. Mobile phones, desktop and laptop computers, wearable devices, and so on.
- These devices all generate, capture, and store data approximately 2.5 quintillion bytes every day. That's the equivalent of 10 million blu-ray DVDs.

3.Variety

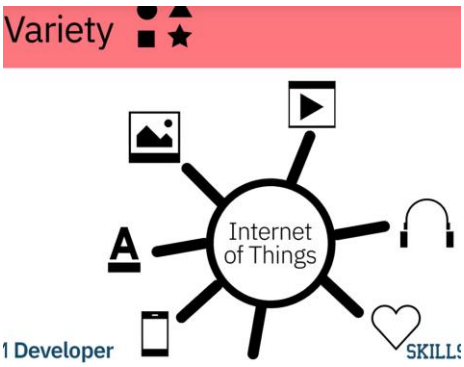


is the diversity of the data.

- Structured data fits neatly into rows and columns in relational databases, while unstructured data is not organized in a predefined way like tweets, blog posts, pictures, numbers, and video.
- Variety also reflects that data comes from different sources; machines, people, and processes, both internal and external to organizations.

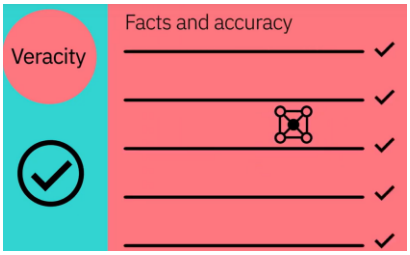


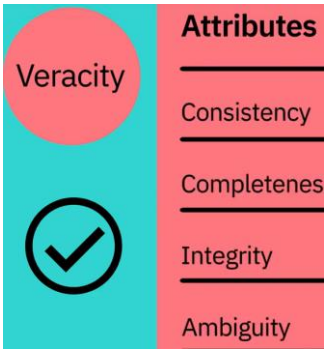
Drivers are mobile technologies social media, wearable technologies, geo technologies video, and many, many more.

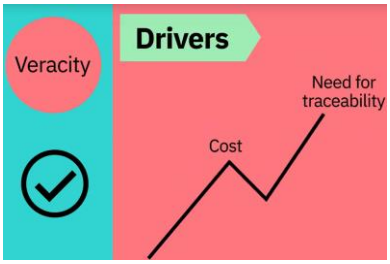


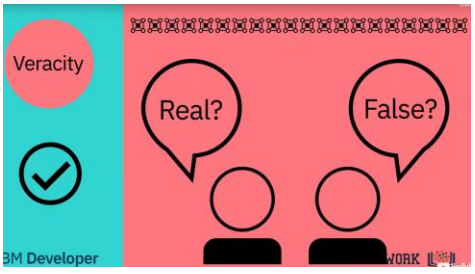
Text, pictures, film, sound, health data from wearable devices, and many different types of data from devices connected to the internet of things.

4.Veracity

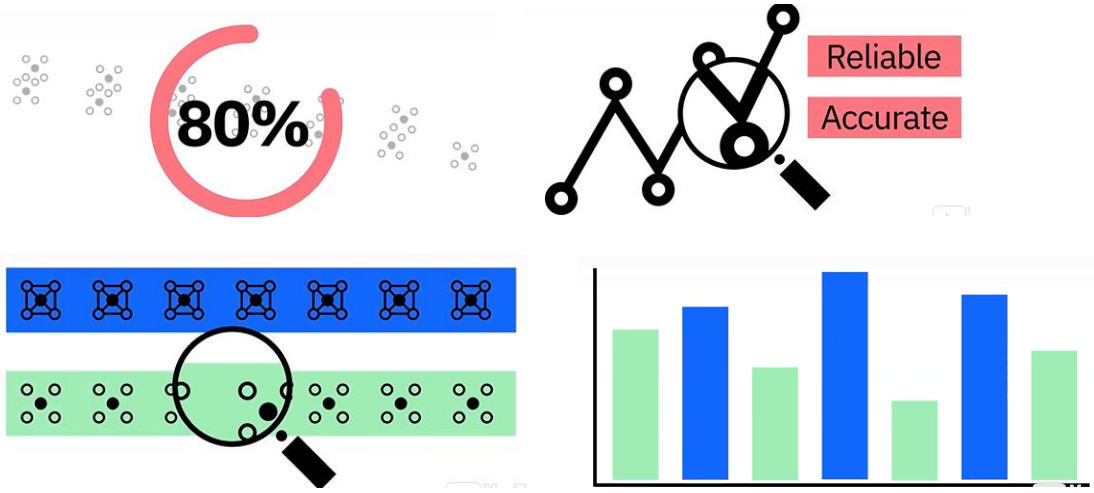








- is the quality and origin of data and its conformity to facts and accuracy.
- Attributes include consistency, completeness, integrity, and ambiguity.
- Drivers include cost and the need for traceability. With the large amount of data available, the debate rages on about the accuracy of data in the digital age. Is the information real or is it false?

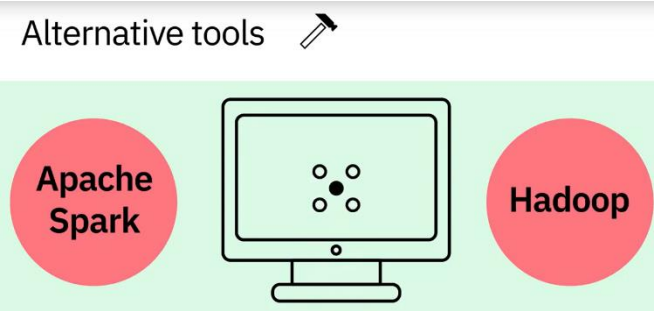


- Eighty percent of data is considered to be unstructured and we must devise ways to produce reliable and accurate insights. The data must be categorized, analyzed, and visualized.

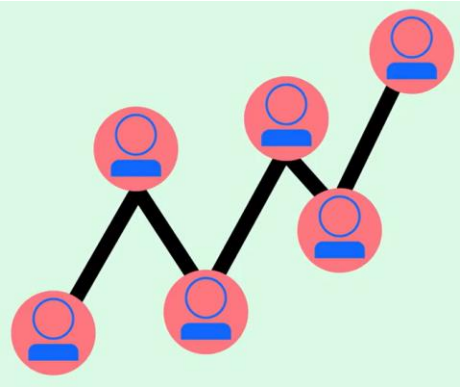
5.Value

- is our ability and need to turn data into value.
- Value isn't just profit. It may have medical or social benefits, as well as customer, employee or personal satisfaction.

Data scientists, today, derive insights from big data and cope with the challenges that these massive data sets present. The scale of the data being collected means that it's not feasible to use conventional data analysis tools, however, alternative tools that leverage distributed computing power can overcome this problem.



Tools such as Apache Spark, Hadoop, and its ecosystem provides ways to extract, load, analyze, and process the data across distributed compute resources, providing new insights and knowledge.



This gives organizations more ways to connect with their customers and enrich the services they offer. So next time you strap on your smartwatch, unlock your smartphone, or track your workout, remember your data is starting a journey that might take it all the way around the world, through big data analysis and back to you.

Big Data Processing Tools

The 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.

In some of the other videos, we discussed Big Data technologies such as NoSQL databases and Data Lakes. In this video, we are going to talk about three open source technologies and the role they play in big data analytics—Apache Hadoop, Apache Hive, and Apache Spark.

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

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

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

1. Hadoop,

- a java-based open-source framework, allows distributed storage and processing of large datasets across clusters of computers.
- In Hadoop distributed system, a node is a single computer, and a collection of nodes forms a cluster.
- Hadoop can scale up from a single node to any number of nodes, each offering local storage and computation.
- Hadoop provides a reliable, scalable, and cost-effective solution for storing data with no format requirements.

Using Hadoop, you can:

- Incorporate emerging data formats, such as streaming audio, video, social media sentiment, and clickstream data, along with structured, semi-structured, and unstructured data not traditionally used in a data warehouse.
- Provide real-time, self-service access for all stakeholders.
- Optimize and streamline costs in your enterprise data warehouse by consolidating data across the organization and moving “cold” data, that is, data that is not in frequent use, to a Hadoop-based system.

One of the four main components of Hadoop is **Hadoop Distributed File System, or HDFS,**

- which is a storage system for big data that runs on multiple commodity hardware connected through a network.
- HDFS provides scalable and reliable big data storage by partitioning files over multiple nodes.
- It splits large files across multiple computers, allowing parallel access to them. Computations can, therefore, run in parallel on each node where data is stored.
- It also replicates file blocks on different nodes to prevent data loss, making it fault-tolerant.

Let's understand this through an **example.**

Consider a file that includes phone numbers for everyone in the United States; the numbers for people with last name starting with A might be stored on server 1, B on server 2, and so on. With Hadoop, pieces of this phonebook would be stored across the cluster. To reconstruct the entire phonebook, your program would need the blocks from every server in the cluster.

- HDFS also replicates these smaller pieces onto two additional servers by default, ensuring availability when a server fails, In addition to higher availability, this offers multiple benefits.
- It allows the Hadoop cluster to break up work into smaller chunks and run those jobs on all servers in the cluster for better scalability.
- Finally, you gain the benefit of data locality, which is the process of moving the computation closer to the node on which the data resides. This is critical when working with large data sets because it minimizes network congestion and increases throughput.

Some of the **other benefits** that come from **using HDFS** include:

- Fast recovery from hardware failures, because HDFS is built to detect faults and automatically recover.
- Access to streaming data, because HDFS supports high data throughput rates.
- Accommodation of large data sets, because HDFS can scale to hundreds of nodes, or computers, in a single cluster.
- Portability, because HDFS is portable across multiple hardware platforms and compatible with a variety of underlying operating systems.

Hive is

- an open-source data warehouse software for reading, writing, and managing large data set files that are stored directly in either HDFS or other data storage systems such as Apache HBase.
- Hadoop is intended for long sequential scans and, because Hive is based on Hadoop, queries have very high latency—which means Hive is less appropriate for applications that need very fast response times.
- Hive is read-based, and therefore not suitable for transaction processing that typically involves a high percentage of write operations.
- Hive is better suited for data warehousing tasks such as ETL, reporting, and data analysis and includes tools that enable easy access to data via SQL.

Spark,

- a general-purpose data processing engine designed to extract and process large volumes of data for a wide range of applications, including Interactive Analytics, Streams Processing, Machine Learning, Data Integration, and ETL.
- It takes advantage of in-memory processing to significantly increase the speed of computations and spilling to disk only when memory is constrained.
- Spark has interfaces for major programming languages, including Java, Scala, Python, R, and SQL.

- It can run using its standalone clustering technology as well as on top of other infrastructures such as Hadoop. And it can access data in a large variety of data sources, including HDFS and Hive, making it highly versatile.
- The ability to process streaming data fast and perform complex analytics in real-time is the key use case for Apache Spark.

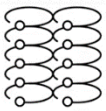
Big Data Processing Tools

The 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.

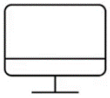
A collection of tools that provides distributed storage and processing of big data.

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

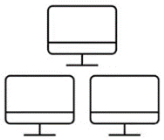
is a distributed data analytics framework designed to perform complex data analytics in real-time.



Distributed storage and processing of large datasets across clusters of computers.



Node



Cluster

Hadoop provides a **reliable, scalable,** and **cost-effective** solution for storing data with no format requirements.

Benefits include:

Better real-time data-driven decisions:
Incorporates emerging data formats not traditionally used in data warehouses

Improved data access and analysis:
Provides real-time, self-service access to stakeholders

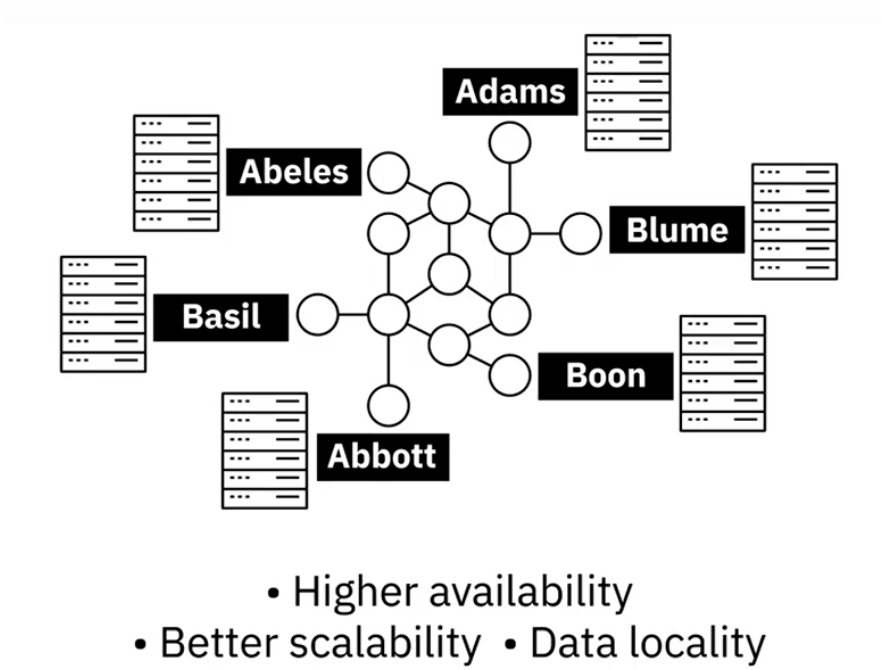
Data offload and consolidation:
Optimizes and streamlines costs by consolidating data, including cold data, across the organization

Hadoop Distributed File System, or HDFS, is a storage system for big data that runs on multiple commodity hardware connected through a network.

Provides scalable and reliable big data storage by partitioning files over multiple nodes

Splits large files across multiple computers, allowing parallel access to them

Replicates file blocks on different nodes to prevent data loss



• Higher availability
• Better scalability • Data locality

Consider a file that includes phone numbers for everyone in the United States; the numbers for people with last name starting with A might be stored on server 1, B on server 2, and so on. With Hadoop, pieces of this phonebook would be stored across the cluster. To reconstruct the entire phonebook, your program would need the blocks from every server in the cluster.

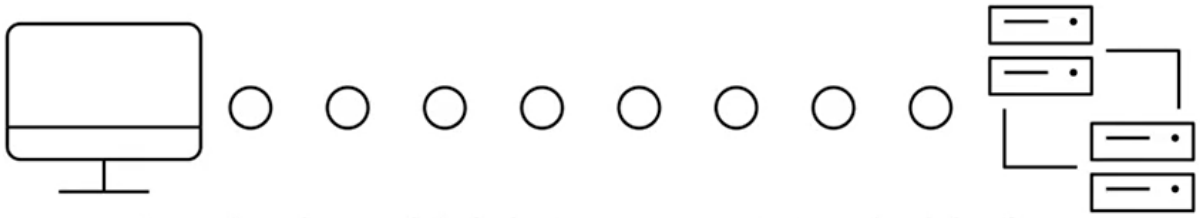
HDFS also replicates these smaller pieces onto two additional servers by default, ensuring availability when a server fails, In addition to higher availability, this offers multiple benefits. It allows the Hadoop cluster to break up work into smaller chunks and run those jobs on all servers in the cluster for better scalability.

Finally, you gain the benefit of data locality, which is the process of moving the computation closer to the node on which the data resides. This is critical when working with large data sets because it minimizes network congestion and increases throughput.

Benefits that come from using HDFS include:

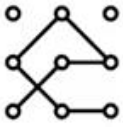
- Fast recovery from hardware failures, because HDFS is built to detect faults and automatically recover.
- Access to streaming data, because HDFS supports high data throughput rates.
- Accommodation of large data sets, because HDFS can scale to hundreds of nodes, or computers, in a single cluster.
- Portability, because HDFS is portable across multiple hardware platforms and compatible with a variety of underlying operating systems.

Hive is an open-source data warehouse software for reading, writing, and managing large data set files that are stored directly in either HDFS or other data storage systems such as Apache HBase.



Queries have high latency → Not suitable for applications that need fast response times

Read-based → Not suitable for transaction processing that involves a high percentage of write operations.



Hive is better suited for →

- Data warehousing tasks such as ETL, reporting, and data analysis

Spark is a general-purpose data processing engine designed to extract and process large volumes of data for a wide range of applications.

- Interactive Analytics
- Streams Processing
- Machine Learning
- Data Integration
- ETL

Key attributes:

- Has in-memory processing which significantly increases speed of computations
- Provides interfaces for major programming languages such as Java, Scala, Python, R, and SQL
- Can run using its standalone clustering technology
- Can also run on top of other infrastructures, such as Hadoop
- Can access data in a large variety of data sources, including HDFS and Hive
- Processes streaming data fast
- Performs complex analytics in real-time

Reading: Summary and Highlights

In this lesson, you have learned the following information:

A Data Repository is a general term that refers to data that has been collected, organized, and isolated so that it can be used for reporting, analytics, and also for archival purposes.

The different types of Data Repositories include:

- Databases, which can be relational or non-relational, each following a set of organizational principles, the types of data they can store, and the tools that can be used to query, organize, and retrieve data.
- Data Warehouses, that consolidate incoming data into one comprehensive storehouse.
- Data Marts, that are essentially sub-sections of a data warehouse, built to isolate data for a particular business function or use case.
- Data Lakes, that serve as storage repositories for large amounts of structured, semi-structured, and unstructured data in their native format.
- Big Data Stores, that provide distributed computational and storage infrastructure to store, scale, and process very large data sets.

ETL, or Extract Transform and Load, Process is an automated process that converts raw data into analysis-ready data by:

- Extracting data from source locations.
- Transforming raw data by cleaning, enriching, standardizing, and validating it.
- Loading the processed data into a destination system or data repository.

Data Pipeline, sometimes used interchangeably with ETL, encompasses the entire journey of moving data from the source to a destination data lake or application, using the ETL process.

Big Data refers to the vast amounts of data that is being produced each moment of every day, by people, tools, and machines. The sheer velocity, volume, and variety of data challenge the tools and systems used for conventional data. These challenges led to the emergence of processing tools and platforms designed specifically for Big Data, such as Apache Hadoop, Apache Hive, and Apache Spark.

Quiz: Practice Quiz

Bookmarked

Question 1

1/1 point (ungraded)

Structured Query Language, or SQL, is the standard querying language for what type of data repository?

- ○ Flat Files
- ◉ RDBMS
- ○ Data lake
- ○ NoSQL

✔

Question 2

1/1 point (ungraded)

In use cases for RDBMS, what is one of the reasons that relational databases are so well suited for OLTP applications?

- ○ Allow you to make changes in the database even while a query is being executed
- ○ Minimize data redundancy
- ◉ Support the ability to insert, update, or delete small amounts of data
- ○ Offer easy backup and restore options

✔

Question 3

1/1 point (ungraded)

Which NoSQL database type stores each record and its associated data within a single document and also works well with Analytics platforms?

- ○ Key-value store
- ◉ Document-based
- ○ Graph-based
- ○ Column-based

Question 4

1/1 point (ungraded)

What type of data repository is used to isolate a subset of data for a particular business function, purpose, or community of users?

- ○ Data Lake
- ○ Data Pipeline
- ◉ Data Mart
- ○ Data Warehouse

✔

Question 5

1/1 point (ungraded)

What does the attribute “Velocity” imply in the context of Big Data?

- ○ Scale of data
- ○ Diversity of data
- ◉ The speed at which data accumulates
- ○ Quality and origin of data

✔

Question 6

1/1 point (ungraded)

Which of the Big Data processing tools provides distributed storage and processing of Big Data?

- ○ Hive
- ◉ Hadoop
- ○ Spark
- ○ ETL

✔

Quiz: Graded Quiz

Bookmarked

Graded Quiz due Jul 12, 2022 03:44 +08

Question 1

1/1 point (graded)

Data Marts and Data Warehouses have typically been relational, but the emergence of what technology has helped to let these be used for non-relational data?

- ○ SQL
- ○ Data Lake
- ◉ NoSQL
- ○ ETL

✔

Question 2

1/1 point (graded)

What is one of the most significant advantages of an RDBMS?

- ○ Enforces a limit on the length of data fields
- ◉ Is ACID-Compliant
- ○ Requires source and destination tables to be identical for migrating data
- ○ Can store only structured data

✔

Question 3

1/1 point (graded)

Which one of the NoSQL database types uses a graphical model to represent and store data, and is particularly useful for visualizing, analyzing, and finding connections between different pieces of data?

- ○ Document-based
- ◉ Graph-based
- ○ Key value store
- ○ Column-based

✔

Question 4

1/1 point (graded)

Which of the data repositories serves as a pool of raw data and stores large amounts of structured, semi-structured, and unstructured data in their native formats?

- ○ Data Warehouses
- ◉ Data Lakes
- ○ Relational Databases
- ○ Data Marts

✔

Question 5

1/1 point (graded)

What does the attribute "Veracity" imply in the context of Big Data?

- ◉ Accuracy and conformity of data to facts
- ○ Diversity of the type and sources of data
- ○ Scale of data
- ○ The speed at which data accumulates

✔

Question 6

1/1 point (graded)

Apache Spark is a general-purpose data processing engine designed to extract and process Big Data for a wide range of applications. What is one of its key use cases?

- ◉ Perform complex analytics in real-time
- ○ Fast recovery from hardware failures
- ○ Consolidate data across the organization
- ○ Scalable and reliable Big Data storage

✔

Module 5

Introduction

In this module, you will learn about the process of identifying and gathering data from disparate sources. Data has never been as diverse as it is now, and it is continually evolving. This module will introduce you to the different methods and tools available for gathering data from different data sources, such as, databases, the web, sensor data, and data exchanges. You will also learn about the features and characteristics of some of the popular tools used for gathering and importing data.

Learning Objectives

After completing this module, you will be able to:

- Describe the process and steps you need to take to identify, gather, and import data from disparate sources .
- List different types of sources of data.
- Describe the different methods and tools available for importing data from disparate data sources into destination data repositories.

Identifying Data for Analysis

At this stage, you have an understanding of the problem and the desired outcome—you know “Where you are” and “Where you want to be.“

You also have a well-defined metric—you know “What will be measured,” and “How it will be measured.”

The next step is for you is to identify the data you need for your use case. The process of identifying data begins by determining the information you want to collect.

In this step, you make decisions regarding (a) the specific information you need; and (b) the possible sources for this data. Your goals determine the answers to these questions.

Let’s take the example of a product company that wants to create targeted marketing campaigns based on the age group that buys their products the most. Their goal is to design reach-outs that appeal most to this segment and encourages them to further influence their friends and peers into buying these products.

Based on this use case, some of the obvious information that you will identify includes the customer profile, purchase history, location, age, education, profession, income, and marital status, for example. To ensure you gain even greater insights into this segment, you may also decide to collect the customer complaint data for this segment to understand the kind of issues they face because this could discourage them from recommending your products.

To know how satisfied they were with the resolution of their issues, you could collect the ratings from the customer service surveys. Taking this a step forward, you may want to understand how these customers talk about your products on social media and how many of their connections engage with them in these discussions, for example, the likes, shares, and comments their posts receive. The next step in the process is to define a plan for collecting data.

You need to establish a timeframe for collecting the data you have identified. Some of the data you need may be required on an ongoing basis and some over a defined period of time. For collecting website visitor data, for example, you may need to have the numbers refreshed in real-time. But if you’re tracking data for a specific event, you have a definite beginning and end date for collecting the data.

In this step, you can also define how much data would be sufficient for you to reach a credible analysis. Is the volume defined by the segment, for example, all customers within the age range of 21 to 30 years; or a dataset of a hundred thousand customers within the age range of 21 to 30.

You can also use this step to define the dependencies, risks, mitigation plan, and several other such factors that are relevant to your initiative. The purpose of the plan should be to establish the clarity you need for execution.

The third step in the process is for you to determine your data collection methods. In this step, you will identify the methods for collecting the data you need. You will define how you will collect the data from the data sources you have identified, such as internal systems, social media sites, or third-party data providers.

Your methods will depend on the type of data, the timeframe over which you need the data, and the volume of data. Once your plan and data collection methods are finalized, you can implement your data collection strategy and start collecting data.

You will be making updates to your plan as you go along because conditions evolve as you implement the plan on the ground. 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.

None of these are one-time considerations but are relevant through the life cycle of the data analysis process. Working with data from disparate sources without considering how it measures against the quality metric can lead to failure. In order to be reliable, data needs to be free of errors, accurate, complete, relevant, and accessible.

You need to define the quality traits, the metric, and the checkpoints in order to ensure that your analysis is going to be based on quality data. You also need to watch out for issues pertaining to data governance, such as, security, regulation, and compliances.

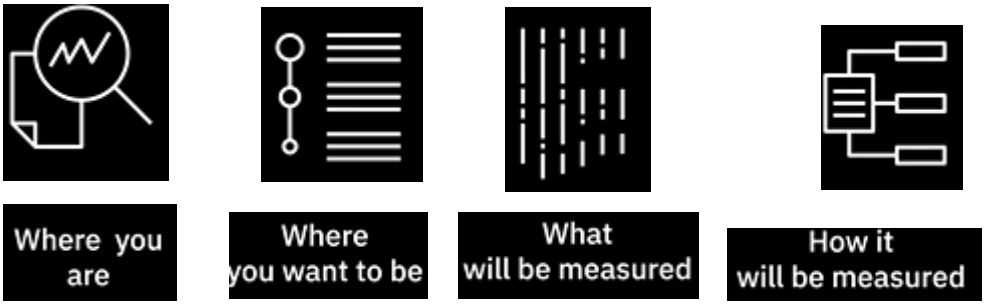
Data Governance policies and procedures relate to the usability, integrity, and availability of data. Penalties for non-compliance can run into millions of dollars and can hurt the credibility of not just your findings, but also your organization.

Another important consideration is data privacy. Data you collect needs to check the boxes for confidentiality, license for use, and compliance to mandated regulations. Checks, validations, and an auditable trail needs to be planned. Loss of trust in the data used for analysis can compromise the process, result in suspect findings, and invite penalties.

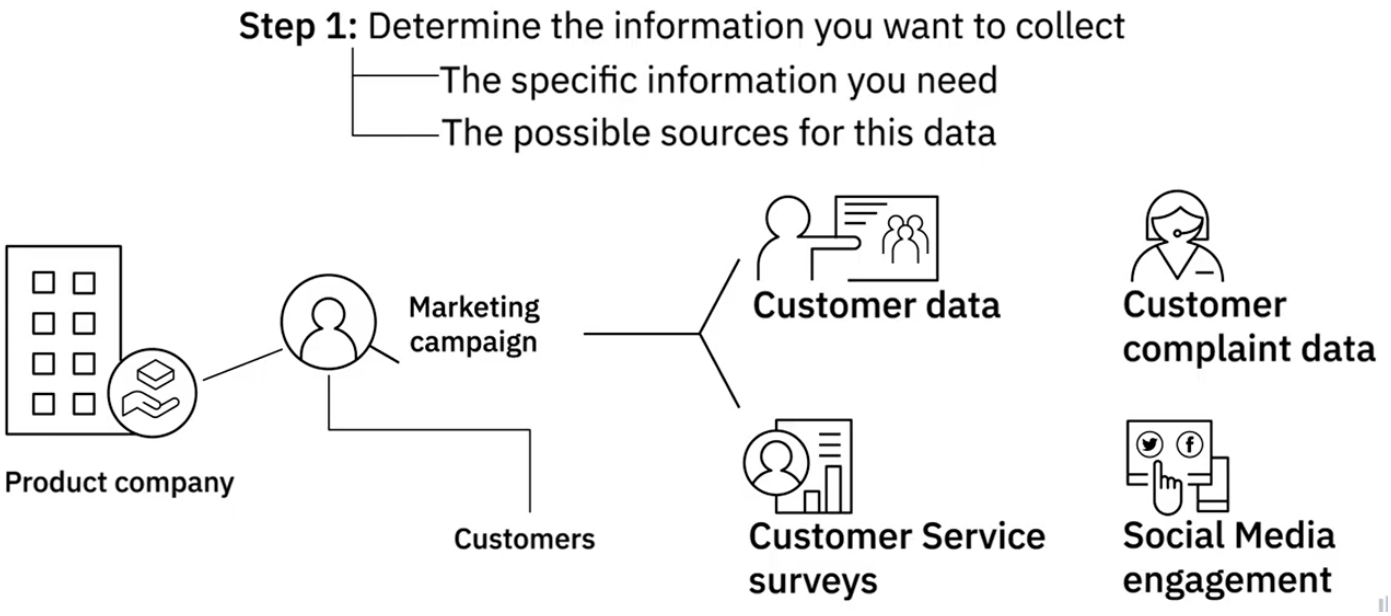
Identifying the right data is a very important step of the data analysis process. Done right, it will ensure that you are able to look at a problem from multiple perspectives and your findings are credible and reliable.

Identifying Data for Analysis

Overview



Process for Identifying Data



The next step is for you is to identify the data you need for your use case. The process of identifying data begins by determining the information you want to collect. In this step, you make decisions regarding (a) the specific information you need; and (b) the possible sources for this data. Your goals determine the answers to these questions.

Let's take the example of a product company that wants to create targeted marketing campaigns based on the age group that buys their products the most. Their goal is to design reach-outs that appeal most to this segment and encourages them to further influence their friends and peers into buying these products.

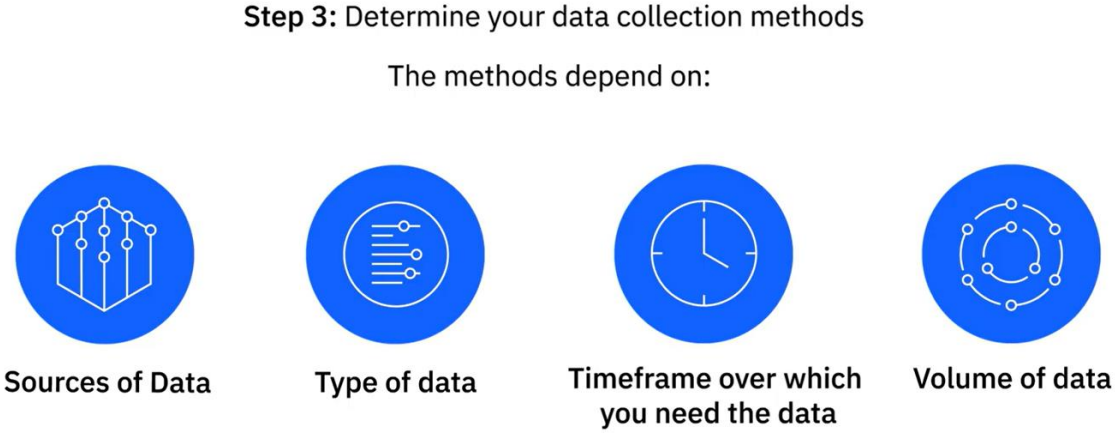
Based on this use case, some of the obvious information that you will identify includes the customer profile, purchase history, location, age, education, profession, income, and marital status, for example. To ensure you gain even greater insights into this segment, you may also decide to collect the customer complaint data for this segment to understand the kind of issues they face because this could discourage them from recommending your products.

To know how satisfied they were with the resolution of their issues, you could collect the ratings from the customer service surveys. Taking this a step forward, you may want to understand how these customers talk about your products on social media and how many of their connections engage with them in these discussions, for example, the likes, shares, and comments their posts receive.

The next step in the process is to define a plan for collecting data. You need to establish a timeframe for collecting the data you have identified. Some of the data you need may be required on an ongoing basis and some over a defined period of time. For collecting website visitor data, for example, you may need to have the numbers refreshed in real-time. But if you're tracking data for a specific event, you have a definite beginning and end date for collecting the data.

In this step, you can also define how much data would be sufficient for you to reach a credible analysis. Is the volume defined by the segment, for example, all customers within the age range of 21 to 30 years; or a dataset of a hundred thousand customers within the age range of 21 to 30.

You can also use this step to define the dependencies, risks, mitigation plan, and several other such factors that are relevant to your initiative. The purpose of the plan should be to establish the clarity you need for execution.



The third step in the process is for you to determine your data collection methods. In this step, you will identify the methods for collecting the data you need. You will define how you will collect the data from the data sources you have identified, such as internal systems, social media sites, or third-party data providers. Your methods will depend on the type of data, the timeframe over which you need the data, and the volume of data.

Once your plan and data collection methods are finalized, you can implement your data collection strategy and start collecting data. You will be making updates to your plan as you go along because conditions evolve as you implement the plan on the ground.

Key Considerations

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. None of these are one-time considerations but are relevant through the life cycle of the data analysis process.

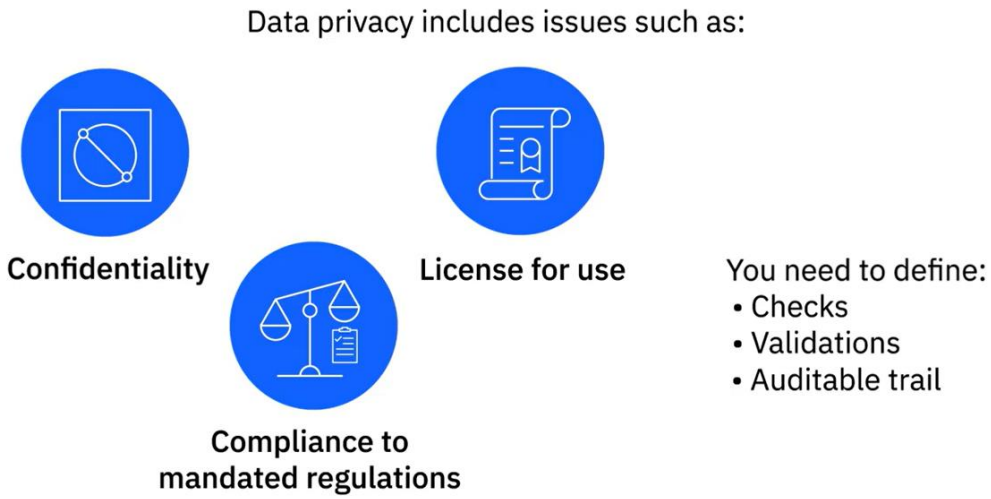
Data Quality

Working with data from disparate sources without considering how it measures against the quality metric can lead to failure. In order to be reliable, data needs to be free of errors, accurate, complete, relevant, and accessible. You need to define the quality traits, the metric, and the checkpoints in order to ensure that your analysis is going to be based on quality data.

Data Governance

You also need to watch out for issues pertaining to data governance, such as, security, regulation, and compliances. Data Governance policies and procedures relate to the usability, integrity, and availability of data. Penalties for non-compliance can run into millions of dollars and can hurt the credibility of not just your findings, but also your organization.

Data Privacy



Data you collect needs to check the boxes for confidentiality, license for use, and compliance to mandated regulations. Checks, validations, and an auditable trail needs to be planned. Loss of trust in the data used for analysis can compromise the process, result in suspect findings, and invite penalties.

Conclusion

Identifying the right data is a very important step of the data analysis process. Done right, it will ensure that you are able to look at a problem from multiple perspectives and your findings are credible and reliable.

Data Sources

Data sources can be internal or external to the organization, and they can be primary, secondary or third party sources of data. Let's look at a couple of examples to understand what we mean by primary, secondary and 3rd party sources of data.

The term primary data refers to information obtained directly by you from the source. This could be from internal sources such as data from the organization, CRM, HR or workflow applications. It could also include data you gather directly through surveys, interviews, discussions, observations and focus groups.

Secondary data refers to information retrieved from existing sources, such as external databases, research articles, publications, training material and Internet searches, or financial records available as public data. This could also include data collected through externally conducted surveys, interviews, discussions, observations and focus groups.

Third party data is data you purchased from aggregators who collect data from various sources and combine it into comprehensive datasets purely for the purpose of selling the data.

Now will look at some of the different sources from which you could be gathering data. Databases can be a source of primary, secondary and 3rd party data. Most organizations have internal applications for managing their processes, workflows and customers.

External databases are available on a subscription basis or for purchase. A significant number of businesses have or are currently moving to the cloud, which is increasingly becoming a source for accessing real time information and on demand insights.

The Web is a source of publicly available data that is available to companies. And individuals for free or commercial use. The Web is a rich source of data available in the public domain. These could include textbooks, government records, papers, and articles that are for public consumption, social media sites, and interactive platforms such as Facebook, Twitter, Google, YouTube. An Instagram are increasingly being used to source user data and opinions.

Businesses are using these data sources for quantitative and qualitative insights. An existing and potential customers. Sensor data produced by wearable devices, smart buildings, smart cities, smart phones, medical devices, even household appliances is a widely used source of data. Data exchange is a source of 3rd party data that involves the voluntary sharing of data between data providers and data consumers, individuals, organizations and governments could be both data providers and data consumers.

The data that is exchanged could include data coming from business applications, sensor devices, social media activity, location data, or consumer behavior data. Surveys gather information through questionnaires distributed to a select group of people. For example, gauging the interest of existing customers in spending on an updated version of a product.

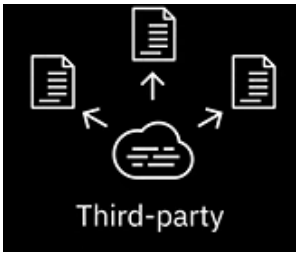
Surveys can be web or paper based. Census data is also a commonly used source for gathering household data, such as wealth and income or population data, for example. Interviews are source for gathering qualitative data, such as the participants opinions and experiences. For example, an interview conducted to understand the day-to-day

challenges faced by a customer service executive. Interviews could be telephonic over the Web or face to face observation. Studies include monitoring participants in a specific environment or while performing a particular task. For example, observing users navigate an E Commerce site to assess the. Ease with which they are able to find products and make a purchase data from surveys, interviews, an observation.

Studies could be available as primary, secondary and 3rd party data. Data sources have never been as dynamic and diverse as they are today. They are also evolving continuously. Supplementing your primary data with secondary and 3rd party data sources can help you explore problems and solutions in new and meaningful ways.

Data Sources

Data sources can be internal or external to the organization



Primary Data

The term primary data refers to information obtained directly by you from the source.

- Data from the organization, CRM, HR or workflow applications
- Data you gather directly through surveys, interviews, discussions, observations and focus groups.

Secondary Data

- External databases
- Research articles, publications, training material and Internet searches, or financial records available as public data
- Data collected through externally conducted surveys, interviews, discussions, observations and focus groups.

Third Party Data

Third party data is data you purchased from aggregators who collect data from various sources and combine it into comprehensive datasets purely for the purpose of selling the data.

Sources for Gathering Data

1. Databases

Databases can be a source of primary, secondary, and third-party data.

- Internal applications for managing processes, workflows, and customers.
- External databases available on a subscription basis or for purchase.

2. Web

Web is a source of publicly available data that is available to companies and individuals for free or commercial use.

- Textbooks
- Government records
- Papers and articles for public consumption

3. Social media sites and interactive platforms

Social media sites and Interactive platforms such as Facebook, Twitter, Google, YouTube, and Instagram are increasingly being used to source user data and opinions.

4. Source of data

Sensor data produced by wearable devices, smart buildings, smart cities, smartphones, medical devices, even household appliances, is a widely used source of data.

5. Data Exchange

Data Exchange is a source of third-party data that involves the voluntary sharing of data between data providers and data consumers. Individuals, organizations, and governments could be both data providers and data consumers.

- Data from business applications
- Sensor devices
- Social media activity
- Location data
- Consumer behavior data

6. Surveys

Surveys gather information through questionnaires distributed to a select group of people.

7. Census

Census data is popularly used for gathering household data such as wealth and income or population data.

8.Interviews

Interviews are a source for gathering qualitative data such as the participant’s opinions and experiences. Interviews can be telephonic, over the web, or face-to-face.

9. Observation Studies

Observation studies include monitoring participants in a specific environment or while performing a particular task.

10. Sources of Gathering Data

- Dynamic
- Diverse
- Continuously evolving

How to Gather and Import Data

In this video, we will learn about the different methods and tools available for gathering data from the data sources discussed earlier in the course—such as databases, the web, sensor data, data exchanges, and several other sources leveraged for specific data needs.

We will also learn about importing data into different types of data repositories. SQL, or Structured Query Language, is a querying language used for extracting information from relational databases.

SQL offers simple commands to specify what is to be retrieved from the database, the table from which it needs to be extracted, grouping records with matching values, dictating the sequence in which the query results are displayed, and limiting the number of results that can be returned by the query, amongst a host of other features and functionalities.

Non-relational databases can be queried using SQL or SQL-like query tools. Some non-relational databases come with their own querying tools such as CQL for Cassandra and GraphQL for Neo4J.

Application Programming Interfaces (or APIs) are also popularly used for extracting data from a variety of data sources. APIs are invoked from applications that require the data and access an end-point containing the data. End-points can include databases, web services, and data marketplaces. APIs are also used for data validation. For example, a data analyst may utilize an API to validate postal addresses and zip codes.

Web scraping, also known as screen scraping or web harvesting, is used for downloading specific data from web pages based on defined parameters. Among other things, web scraping is used to extract data such as text, contact information, images, videos, podcasts, and product items from a web property.

RSS feeds are another source typically used for capturing updated data from online forums and news sites where data is refreshed on an ongoing basis.

Data streams are a popular source for aggregating constant streams of data flowing from sources such as instruments, IoT devices and applications, and GPS data from cars. Data streams and feeds are also used for extracting data from social media sites and interactive platforms.

Data Exchange platforms allow the exchange of data between data providers and data consumers. Data Exchanges have a set of well-defined exchange standards, protocols, and formats relevant for exchanging data. These platforms not only facilitate the exchange of data, they also ensure that security and governance are maintained. They provide data licensing workflows, de-identification and protection of personal information, legal frameworks, and a quarantined analytics environment.

Examples of popular data exchange platforms include AWS Data Exchange, Crunchbase, Lotame, and Snowflake.

Numerous other data sources can be tapped into for specific data needs. For marketing trends and ad spending, for example, research firms like Forrester and

Business Insider are known to provide reliable data. Research and advisory firms such as Gartner and Forrester are widely trusted sources for strategic and operational guidance.

Similarly, there are many trusted names in the areas of user behavior data, mobile and web usage, market surveys, and demographic studies. Data that has been identified and gathered from the various data sources now needs to be loaded or imported into a data repository before it can be wrangled, mined, and analyzed.

The importing process involves combining data from different sources to provide a combined view and a single interface using which you can query and manipulate the data. Depending on the data type, the volume of data, and the type of destination repository, you may need varying tools and methods. Specific data repositories are optimized for certain types of data.

Relational databases store structured data with a well-defined schema. If you're using a relational database as the destination system, you will only be able to store structured data, such as data from OLTP systems, spreadsheets, online forms, sensors, network and web logs.

Structured data can also be stored in NoSQL. Semi-structured data is data that has some organizational properties but not a rigid schema, such as, data from emails, XML, zipped files, binary executables, and TCP/IP protocols.

Semi-structured can be stored in NoSQL clusters. XML and JSON are commonly used for storing and exchanging semi-structured data. JSON is also the preferred data type for web services.

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

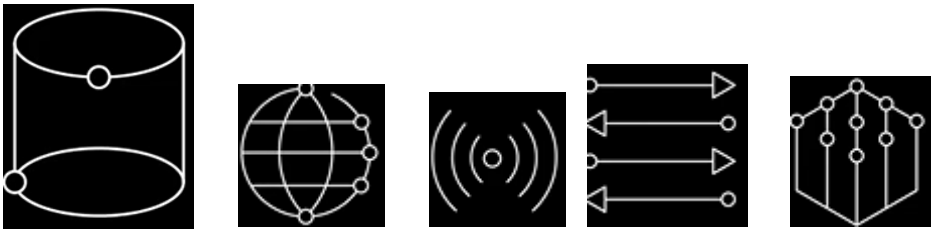
data.

Tools such as Talend and Informatica, and programming languages such as Python and R, and their libraries, are widely used for importing data.

How to Gather and Import Data

Overview

Gathering data from the data sources discussed earlier in the course—such as databases, the web, sensor data, data exchanges, and several other sources leveraged for specific data needs.



- importing data into different types of data repositories

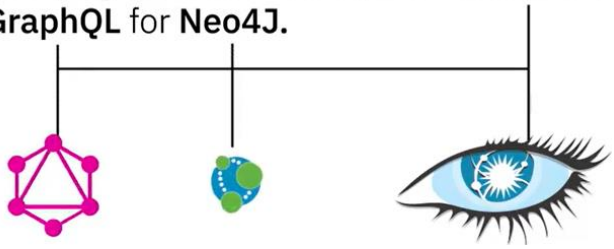
SQL, or **Structured Query Language,** is a querying language used for extracting information from relational databases.

Offers simple commands to specify

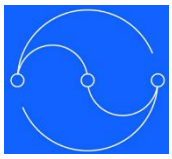
- What is to be retrieved from the database
- Table from which it needs to be extracted
- Grouping records with matching values
- Dictating the sequence in which the query results are displayed
- Limiting the number of results that can be returned by the query

Non-relational databases can be queried using **SQL** or **SQL**-like query tools.

Some non-relational databases come with their own querying tools such as **CQL** for **Cassandra** and **GraphQL** for **Neo4J.**



APIs



Application Programming Interfaces (or APIs)

Popularly used for extracting data from a variety of data sources.

Are invoked from applications that require the data and access an endpoint containing the data. Endpoints can include databases, web services, and data marketplaces.

Also used for data validation.

Extracting data from the web

Web Scraping (Screen Scraping, Web Harvesting)

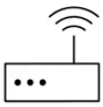
For downloading specific data from web pages based on defined parameters.

For extracting data such as text, contact information, images, videos, podcasts, and product items from a web property.

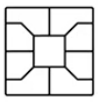
RSS feeds are used for capturing updated data from online forums and news sites where data is refreshed on an ongoing basis.

Sensor data

Data streams are a popular source for aggregating constant streams of data flowing from sources such as



Instruments



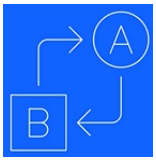
IoT devices

Applications

GPS data from cars

Data streams and feeds are also used for extracting data from social media sites and interactive platforms.

Data Exchanges



Data Exchange platforms

Allow the exchange of data between data providers and data consumers

Have a set of well-defined exchange standards, protocols, and formats relevant for exchanging data

Facilitate the exchange of data

Ensure that security and governance are maintained

Data Exchange platforms

Provide data licensing workflows, de-identification and protection of personal information, legal frameworks, and a quarantined analytics environment

Examples of popular data exchange platforms include

AWS DataExchange

Crunchbase

Lotame

Snowflake

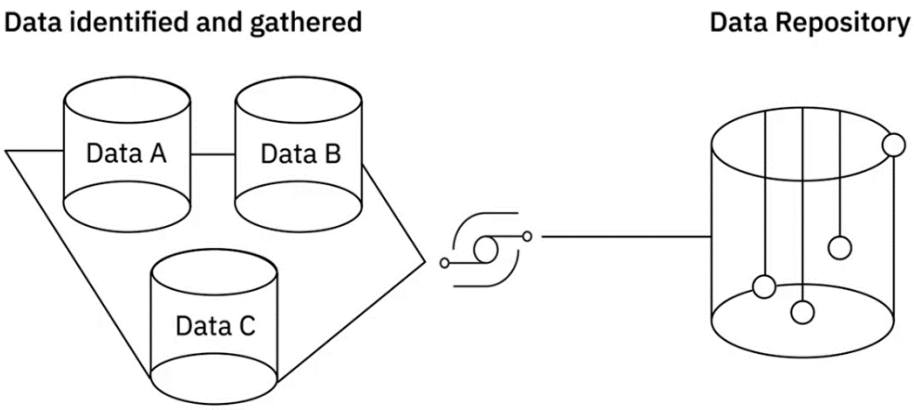
Other sources

Numerous other data sources can be tapped into for specific data needs.

For example

- Forrester and Business Insider for marketing trends and ad spending
- Research and advisory firms such as Gartner and Forrester for strategic and operational guidance
- Agencies for user behavior data, mobile and web usage, market surveys, and demographic studies.

Importing Data



Data that has been identified and gathered from the various data sources now needs to be loaded or imported into a data repository before it can be wrangled, mined, and analyzed. The importing process involves combining data from different sources to provide a combined view and a single interface using which you can query and manipulate the data. Depending on the data type, the volume of data, and the type of destination repository, you may need varying tools and methods.

Data Types and Destination Repositories

Specific data repositories are optimized for certain types of data.

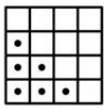
Structured data

Relational databases store structured data with a well-defined schema

Sources include data from OLTP systems, spreadsheets, online forms, sensors, network and web logs

Can also be stored in NoSQL databases

Specific data repositories are optimized for certain types of data.



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

Specific data repositories are optimized for certain types of data.

Unstructured data

Sources include web pages, social media feeds, images, videos, documents, media logs, and surveys

Can be stored in NoSQL databases and data lakes

ETL tools and data pipelines provide automated functions that facilitate the process of importing data.

Tools for importing data: talend Informatica python R

Reading: Summary and Highlights

In this lesson, you have learned:

- The process of identifying data begins by determining the information that needs to be collected, which in turn is determined by the goal you seek to achieve.
- 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.

Quiz: Practice Quiz

Question 1

1/1 point (ungraded)

What are the requirements in order for data to be reliable? (Select all that apply)

- [x] Data should be free of all errors
- [ ] Data should be structured
- [ ] Data should be easy to collect
- [x] Data should be relevant

✔

Question 2

1/1 point (ungraded)

What type of data is produced by wearable devices, smart buildings, and medical devices?

- ○ Observation study data
- ◉ Sensor data
- ○ Census data
- ○ Survey data

✔

Question 3

1/1 point (ungraded)

What type of data is semi-structured and has some organizational properties but not a rigid schema?

- ○ Web logs
- ○ Data from OLTP systems
- ◉ Emails
- ○ Online forms

✔

Quiz: Graded Quiz

Bookmark this page

Graded Quiz due Jul 14, 2022 11:44 +08

Question 1

1/1 point (graded)

What are some of the steps in the process of "Identifying Data"? (Select all that apply) .

- [x] Determine the information you want to collect
- [x] Define a plan for collecting data
- [ ] Determine the visualization tools that you will use
- [ ] Define the checkpoints

✔

Question 2

1/1 point (graded)

What type of data refers to information obtained directly from the source?

- (•) Primary data
- ( ) Sensor data
- ( ) Third-party data
- ( ) Secondary data

✔

Submit You have used 1 of 2 attempts

✔ Correct (1/1 point)

Question 3

1/1 point (graded)

Web scraping is used to extract what type of data?

- ○ Images, videos, and data from NoSQL databases
- ○ Text, videos, and data from relational databases
- ○ Data from news sites and NoSQL databases
- ◉ Text, videos, and images

✔

Question 4

1/1 point (graded)

Data obtained from an organization's internal CRM, HR, and workflow applications is classified as:

- ○ Third-party data
- ◉ Primary data
- ○ Copyright-free data
- ○ Secondary data

✔

Submit You have used 1 of 2 attempts Reset

✔ Correct (1/1 point)

Question 5

1/1 point (graded)

Which of the provided options offers simple commands to specify what is to be retrieved from a relational database?

- (●) SQL
- ( ) Web Scraping
- ( ) API
- ( ) RSS Feed

✔

Module Introduction

In this module, you will learn about the process and key tasks involved in data wrangling and data cleaning. You will also learn about the features and use cases for some of the popularly used software and tools for data wrangling.

Learning Objectives

After completing this module, you will be able to:

Describe the different phases of the data wrangling process and key tasks in each of these phases.

Describe the different tools and techniques required for wrangling and cleaning data so as to make it analysis-ready.

Describe the data cleaning workflow and activities performed at each stage of the workflow.

What is Data Wrangling?

Data wrangling,

- also known as data munging, is an iterative process that involves data exploration, transformation, validation, and making it available for a credible and meaningful analysis.
- It includes a range of tasks involved in preparing raw data for a clearly defined purpose, where raw data at this stage is data that has been collated through various data sources in a data repository.
- Data wrangling captures a range of tasks involved in preparing data for analysis.

Typically, it is a **4-step process that involves—Discovery**, **Transformation, Validation,** and **Publishing.**

The **1. Discovery phase**,

- also known as the Exploration phase, is about understanding your data better with respect to your use case.
- The objective is to figure out specifically how best you can clean, structure, organize, and map the data you have for your use case.

The next phase, which is the **2. Transformation phase**,

- forms the bulk of the data wrangling process. It involves the tasks you undertake to transform the data, such as structuring, normalizing, denormalizing, cleaning, and enriching the data.

Let's begin with the first transformation task – **2.1 Structuring**.

- This task includes actions that change the form and schema of your data.
- The incoming data can be in varied formats. You might, for example, have some data coming from a relational database and some data from Web APIs.
- In order to merge them, you will need to change the form or schema of your data. This change may be as simple as changing the order of fields within a record or dataset or as complex as combining fields into complex structures.
- Joins and Unions are the most common structural transformations used to combine data from one or more tables. How they combine the data is different. Joins combine columns. When two tables are joined together, columns from the first source table are combined with columns from the second source table—in the same row. So, each row in the resultant table contains columns from both tables.
- Unions combine rows. Rows of data from the first source table are combined with rows of data from the second source table into a single table. Each row in the resultant table is from one source table or another.

Transformation can also include **2.2 normalization** and **denormalization of data.**

- Normalization focuses on cleaning the database of unused data and reducing redundancy and inconsistency. Data coming from transactional systems, for example, where a number of insert, update, and delete operations are performed on an ongoing basis, are highly normalized.
- Denormalization is used to combine data from multiple tables into a single table so that it can be queried faster. For example, normalized data coming from transactional systems is typically denormalized before running queries for reporting and analysis.

Another transformation type is **2.3 Cleaning**.

- Cleaning tasks are actions that fix irregularities in data in order to produce a credible and accurate analysis.
- Data that is inaccurate, missing, or incomplete can skew the results of your analysis and need to be considered. It could also be that the data is biased, or has null values in relevant fields, or have outliers.
- For example, you may want to find out the demographic information on the sale of a certain product, but the data you have received does not capture the gender. You either need to source this data point and merge it with your existing dataset, or you may need to remove, and not consider the records with this field missing. We will explore many more examples of data cleaning further on in the course.

2.4 Enriching the data—is the fourth type of transformation.

- When you consider the data you have, to look at additional data points that could make your analysis more meaningful, you are looking at enriching your data.
- For example, in a large organization with information fragmented across systems, you may need to enrich the dataset provided by one system with information available in other systems, or even public datasets.
- Consider a scenario where you sell IT peripherals to businesses and want to analyze the buying patterns of your customers over the last five years. You have the customer master and transaction tables from where you've captured the customer information and purchase history.
- Supplementing your dataset with the performance data of these businesses, possibly available as a public dataset, could be valuable for you to understand factors influencing their purchase decisions.
- Inserting metadata also enriches data. For example, computing a sentiment score from a customer feedback log, collecting geo-based weather data from a resorts location to analyze occupancy trends, or capturing published time and tags for a blog post.

After transformation, the next phase in Data Wrangling is **3. Validation**.

- This is where you check the quality of the data post structuring, normalizing, cleaning, and enriching.
- Validation rules refer to repetitive programming steps used to verify the consistency, quality, and security of the data you have.

This brings us to **4. Publishing**—the fourth phase of the data wrangling process.

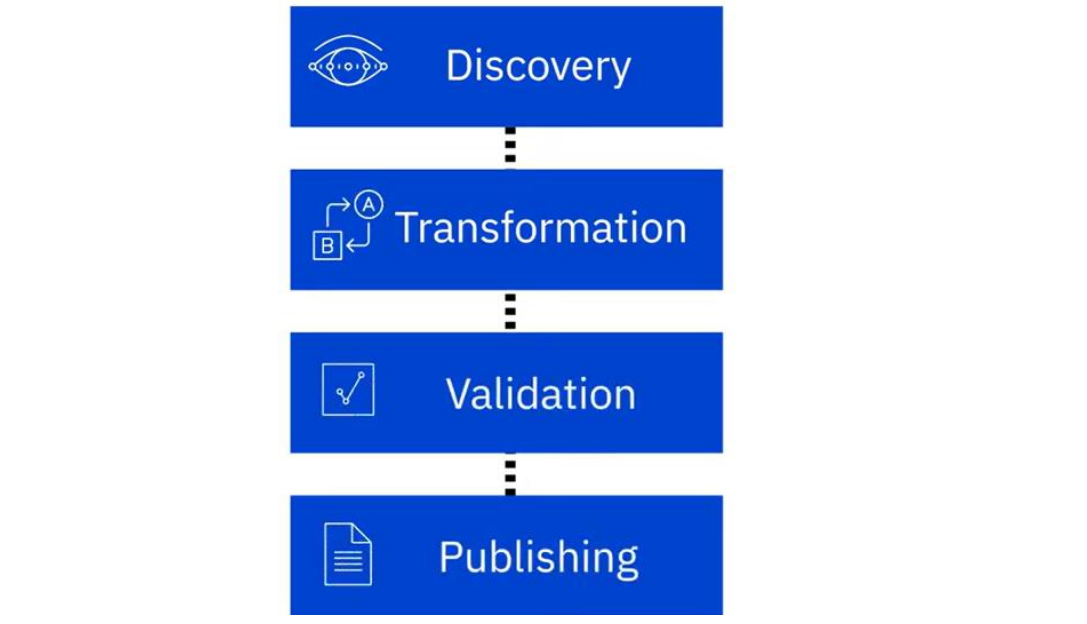
- Publishing involves delivering the output of the wrangled data for downstream project needs.
- What is published is the transformed and validated version of the input dataset along with the metadata about the dataset.

- Lastly, it is important to note the criticality of documenting the steps and considerations you have taken to convert the raw data to analysis-ready data.
- All phases of data wrangling are iterative in nature. In order to replicate the steps and to revisit your considerations for performing these steps, it is vital that you document all considerations and actions.

What is Data Wrangling?

Overview

also known as data munging, is an iterative process that involves data exploration, transformation, validation, and making it available for a credible and meaningful analysis.

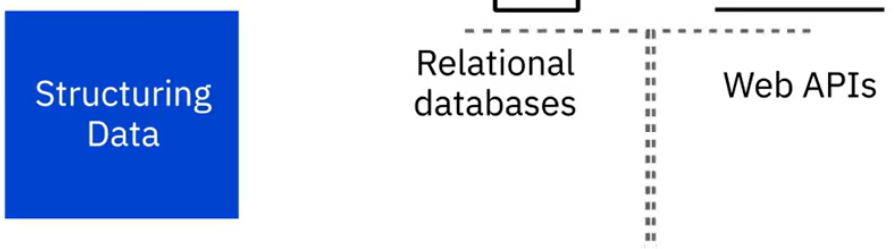


1. Discovery or Exploration Phase

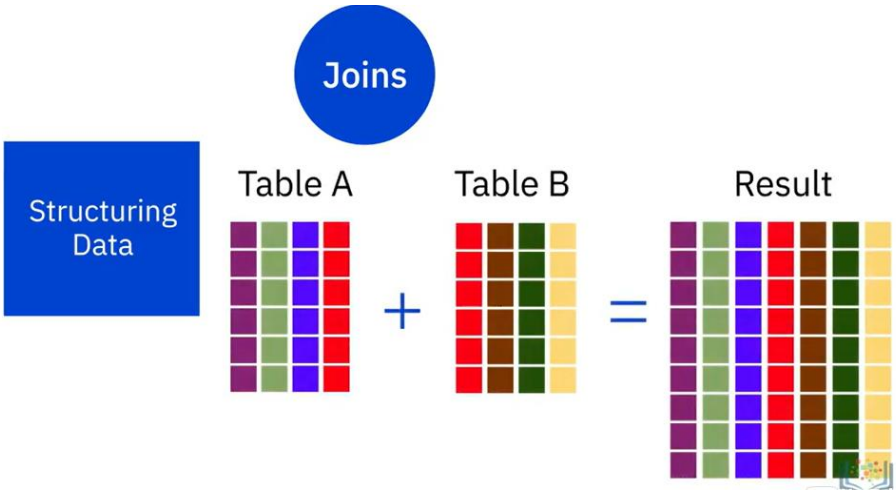
- Examining and understanding your data with respect to your use case
- Creating a plan for cleaning, structuring, organizing, and mapping your data

2. Transformation Phase

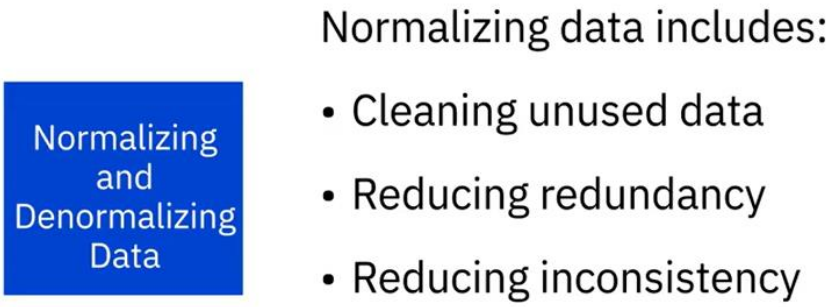
2.1 Structuring Data

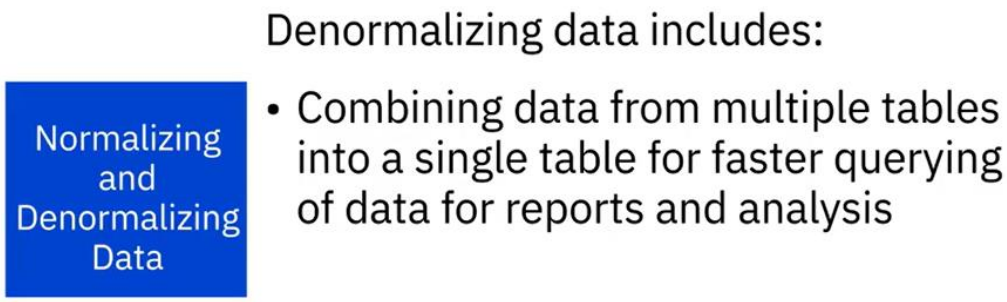


This task includes actions that change the form and schema of your data. The incoming data can be in varied formats. You might, for example, have some data coming from a relational database and some data from Web APIs. In order to merge them, you will need to change the form or schema of your data. This change may be as simple as changing the order of fields within a record or dataset or as complex as combining fields into complex structures.



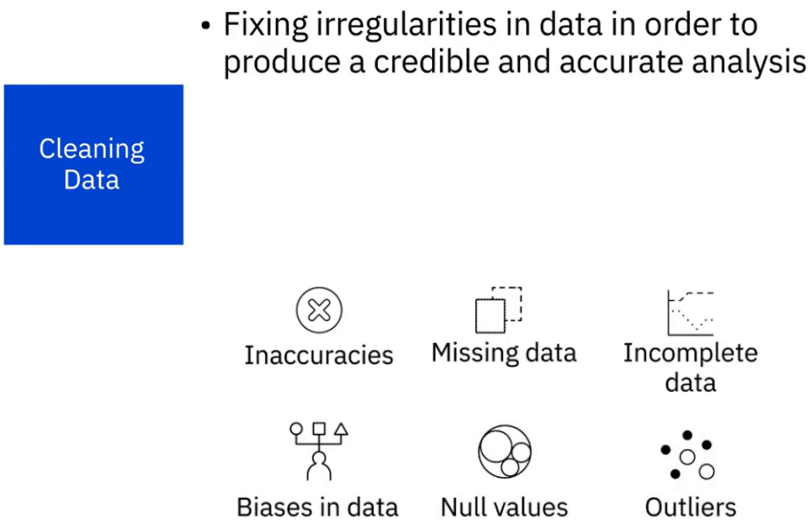
Joins and Unions are the most common structural transformations used to combine data from one or more tables. How they combine the data is different. Joins combine columns. When two tables are joined together, columns from the first source table are combined with columns from the second source table—in the same row. So, each row in the resultant table contains columns from both tables. Unions combine rows. Rows of data from the first source table are combined with rows of data from the second source table into a single table. Each row in the resultant table is from one source table or another.





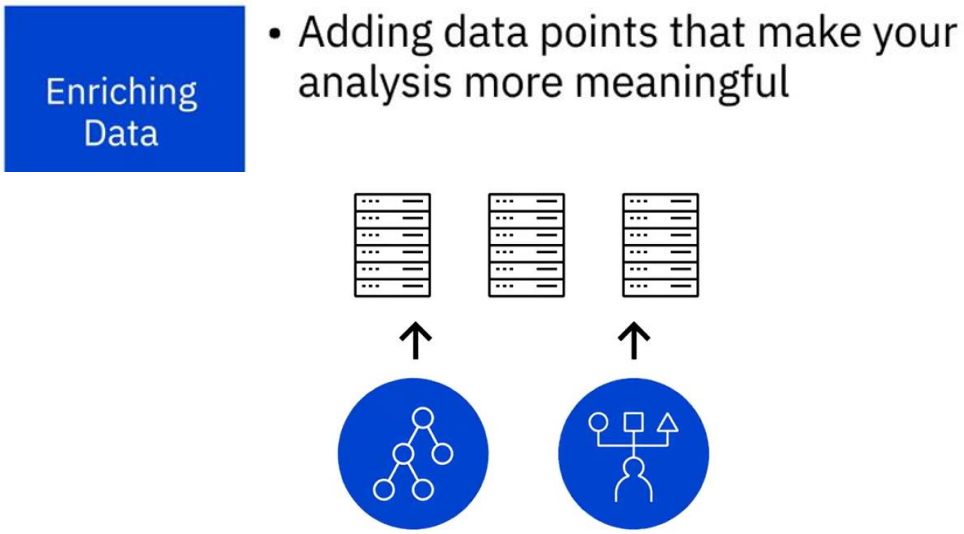
Normalization focuses on cleaning the database of unused data and reducing redundancy and inconsistency. Data coming from transactional systems, for example, where a number of insert, update, and delete operations are performed on an ongoing basis, are highly normalized. Denormalization is used to combine data from multiple tables into a single table so that it can be queried faster.

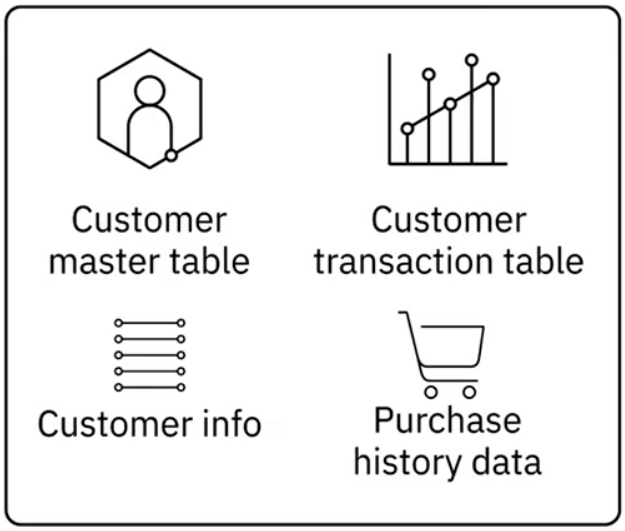
For example, normalized data coming from transactional systems is typically denormalized before running queries for reporting and analysis.



Cleaning tasks are actions that fix irregularities in data in order to produce a credible and accurate analysis. Data that is inaccurate, missing, or incomplete can skew the results of your analysis and need to be considered. It could also be that the data is biased, or has null values in relevant fields, or have outliers.

For example, you may want to find out the demographic information on the sale of a certain product, but the data you have received does not capture the gender. You either need to source this data point and merge it with your existing dataset, or you may need to remove, and not consider the records with this field missing.





Enriching the data—is the fourth type of transformation. When you consider the data you have, to look at additional data points that could make your analysis more meaningful, you are looking at enriching your data. For example, in a large organization with information fragmented across systems, you may need to enrich the dataset provided by one system with information available in other systems, or even public datasets. Consider a scenario where you sell IT peripherals to businesses and want to analyze the buying patterns of your customers over the last five years. You have the customer master and transaction tables from where you've captured the customer information and purchase history. Supplementing your dataset with the performance data of these businesses, possibly available as a public dataset, could be valuable for you to understand factors influencing their purchase decisions. Inserting metadata also enriches data. For example, computing a sentiment score from a customer feedback log, collecting geo-based weather data from a resorts location to analyze occupancy trends, or capturing published time and tags for a blog post.

3. Validation Phase

- Checking the quality of data after structuring, normalizing, denormalizing, cleaning, and enriching of data
- Verifying consistency, quality, and security of data

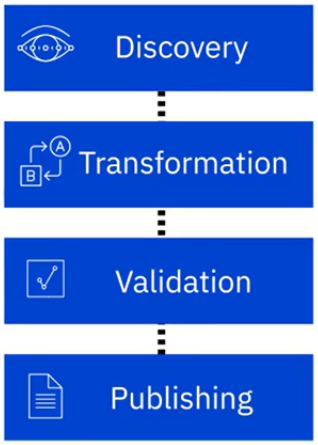
4. Publishing Phase

- Delivering the output of the wrangled data for downstream project needs

What is published is the transformed and validated version of the input dataset along with the metadata about the dataset.

Documentation Phase

- It is important to document the steps you took, and your considerations for taking those steps, to convert raw data into analysis-ready data.



Tools for Data Wrangling

Some of the popularly used **data wrangling software and tools**, such as: **Excel Power Query / Spreadsheets, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python and R.**

Let's begin with the most basic software used for manual wrangling—Spreadsheets.

Spreadsheets

- such as Microsoft Excel and Google Sheets have a host of features and in-built formulae that can help you identify issues, clean, and transform data.
- Add-ins are available that allow you to import data from several different types of sources and clean and transform data as needed—such as Microsoft Power Query for Excel and Google Sheets Query function for Google Sheets.

OpenRefine

- is an open-source tool that allows you to import and export data in a wide variety of formats, such as TSV, CSV, XLS, XML, and JSON.
- Using OpenRefine, you can clean data, transform it from one format to another, and extend data with web services and external data.
- OpenRefine is easy to learn and easy to use. It offers menu-based operations, which means you don't need to memorize commands or syntax.

Google DataPrep

- is an intelligent cloud data service that allows you to visually explore, clean, and prepare both structured and unstructured data for analysis.
- It is a fully managed service, which means you don't need to install or manage the software or the infrastructure.

DataPrep

- is extremely easy to use. With every action that you take, you get suggestions on what your ideal next step should be.
- DataPrep can automatically detect schemas, data types, and anomalies.

Watson Studio Refinery,

- available via IBM Watson Studio, allows you to discover, cleanse, and transform data with built-in operations.
- It transforms large amounts of raw data into consumable, quality information that's ready for analytics.
- Data Refinery offers the flexibility of exploring data residing in a spectrum of data sources.
- It detects data types and classifications automatically and also enforces applicable data governance policies automatically.
-

Trifacta Wrangler

- is an interactive cloud-based service for cleaning and transforming data.
- It takes messy, real-world data and cleans and rearranges it into data tables, which can then be exported to Excel, Tableau, and R.
- It is known for its collaboration features, allowing multiple team members to work simultaneously.

Python has a huge library and set of packages that offer powerful data manipulation capabilities.

- **Jupyter Notebook** is an open-source web application widely used for data cleaning and transformation, statistical modeling, also data visualization.
- **Numpy, or Numerical Python**, is the most basic package that Python offers. It is fast, versatile, interoperable, and easy to use. It provides support for large, multi-dimensional arrays and matrices, and high-level mathematical functions to operate on these arrays.
- **Pandas** is designed for fast and easy data analysis operations.

  It allows complex operations such as merging, joining, and transforming huge chunks of data, performed using simple, single-line commands.

  Using Pandas, you can prevent common errors that result from misaligned data coming in from different sources.

R

- also offers a series of libraries and packages that are explicitly created for wrangling messy data—such as Dplyr, Data.table, and Jsonlite. Using these libraries, you can investigate, manipulate, and analyze data.
- **Dplyr** is a powerful library for data wrangling. It has a precise and straightforward syntax.
- **Data.table** helps to aggregate large data sets quickly.
- Jsonlite is a robust JSON parsing tool, great for interacting with web APIs.

Tools for data wrangling come with varying capabilities and dimensions. Your decision regarding the best tool for your needs will depend on factors that are specific to your use case, infrastructure, and teams—such as supported data size, data structures, cleaning and transformation capabilities, infrastructure needs, ease of use, and learnability.

Tools for Data Wrangling

Some of the popularly used data wrangling software and tools, such as:

- Excel Power Query / Spreadsheets
- OpenRefine
- Google DataPrep,
- Watson Studio Refinery
- Trifacta Wrangler
- Python
- R

Excel Power Query / Spreadsheets

- such as Microsoft Excel and Google Sheets have a host of features and in-built formulae that can help you identify issues, clean, and transform data.
- Add-ins are available that allow you to import data from several different types of sources and clean and transform data as needed—such as Microsoft Power Query for Excel and Google Sheets Query function for Google Sheets.

OpenRefine

Open-source tool

- Can import and export data in a wide variety of formats, such as TSV, CSV, XLS, XML, and JSON.
- Can clean data, transform it from one format to another, and extend data with web services and external data.

- Easy to learn
- Easy to use.

- Offers menu-based operations, which means you don't need to memorize commands or syntax

Google DataPrep

An intelligent cloud data service

- Can visually explore, clean, and prepare both structured and unstructured data for analysis.
- Fully managed service
- Extremely easy to use.
- Offers suggestions on ideal next steps
- Automatically detects schemas, data types and anomalies.

Watson Studio Refinery

- Available via IBM Watson Studio
- Allows you to discover, cleanse, and transform data with built-in operations.
- Transforms large amounts of raw data into consumable, quality information that's ready for analytics.
- Offers the flexibility of exploring data residing in a spectrum of data sources.
- Detects data types and classifications automatically
- Enforces applicable data governance policies automatically.

Trifacta Wrangler

An interactive cloud-based service for cleaning and transforming data.

- Takes messy, real-world data and cleans and rearranges it into data tables
- Can export tables to Excel, Tableau, and R
- Known for its collaboration features, allowing multiple team members to work simultaneously.

Python

Python has a huge library and set of packages that offer powerful data manipulation capabilities.

Jupyter Notebook is an open-source web application widely used for data cleaning and transformation, statistical modeling, also data visualization.

- The most basic package that Python offers.
- Fast, versatile, interoperable, and easy to use.
- Provides support for large, multi-dimensional arrays and matrices, and high-level mathematical functions to operate on these array

- Designed for fast and easy data analysis operations
- Allows complex operations such as merging, joining, and transforming huge chunks of data, performed using simple, single-line commands.
- Helps prevent common errors that result from misaligned data coming in from different sources.

R

- R, also offers a series of libraries and packages that are explicitly created for wrangling messy data—such as Dplyr, Data.table, and Jsonlite.
- Using these libraries, you can investigate, manipulate, and analyze data.

- Dplyr is a powerful library for data wrangling.
- It has a precise and straightforward syntax.

Data.table helps to aggregate large data sets quickly.

Robust JSON parsing tool, great for interacting with web APIs.

Tools for Data Wrangling

Your decision regarding the best tool for your needs will depend on factors that are specific to your use case, infrastructure, and teams—such as:

- supported data size
- data structures
- cleaning and transformation capabilities
- infrastructure needs
- ease of use
- learnability.

Data Cleaning

According to a Gartner report on data quality, poor quality data weakens an organization's competitive standing and undermines critical business objectives.

Missing, inconsistent, or incorrect data can lead to false conclusions and therefore ineffective decisions. And in the business world, that can be costly.

Data sets picked up from disparate sources could have a number of issues, including missing values, inaccuracies, duplicates, incorrect or missing delimiters, inconsistent records, and insufficient parameters. In some cases, data can be corrected manually or automatically with the help of data wrangling tools and scripts, but if it cannot be repaired, it must be removed from the dataset.

Although the terms Data Cleaning and Data Wrangling are sometimes used interchangeably, it is important to keep in mind that data cleaning is only a subset of the entire Data Wrangling process.

Data Cleaning forms a very significant and integral part of the Transformation phase in a data wrangling workflow. A typical data cleaning workflow includes: **Inspection, Cleaning**, and **Verification.**

The first step in the data cleaning workflow is to detect the different types of issues and errors that your dataset may have. You can use scripts and tools that allow you to define specific rules and constraints and validate your data against these rules and constraints. You can also use data profiling and data visualization tools for inspection.

Data profiling helps you to inspect the source data to understand the structure, content, and interrelationships in your data. It uncovers anomalies and data quality issues. For example, blank or null values, duplicate data, or whether the value of a field falls within the expected range.

Visualizing the data using statistical methods can help you to spot outliers. For example, plotting the average income in a demographic dataset can help you spot outliers.

That brings us to the actual cleaning of the data. The techniques you apply for cleaning your dataset will depend on your use case and the type of issues you encounter. Let’s look at some of the more common data issues.

Let’s start with missing values.

Missing values are very important to deal with as they can cause unexpected or biased results. You can choose to filter out the records with missing values or find a way to source that information in case it is intrinsic to your use case. For example, missing age data from a demographics study. A third option is a method known as imputation, which calculates the missing value based on statistical values. Your decision on the course of action you choose needs to be anchored in what’s best for your use case.

You may also come across duplicate data, data points that are repeated in your dataset. These need to be removed.

Another type of issue you may encounter is that of irrelevant data. Data that does not fit within the context of your use case can be considered irrelevant data. For example, if you are analyzing data about the general health of a segment of the population, their contact numbers may not be relevant for you.

Cleaning can involve data type conversion as well. This is needed to ensure that values in a field are stored as the data type of that field—for example, numbers stored as numerical data type or date stored as a date data type. You may also need to clean your data in order to standardize it.

For example, for strings, you may want all values to be in lower case. Similarly, date formats and units of measurement need to be standardized. Then there are syntax errors. For example, white spaces, or extra spaces at the beginning or end of a string is a syntax error that needs to be rectified. This can also include fixing typos or format, for example, the state name being entered as a full form such as New York versus an abbreviated form such as NY in some records.

Data can also have outliers, or values that are vastly different from other observations in the dataset. Outliers may, or may not, be incorrect. For example, when an age field in a voters database has the value 5, you know it is incorrect data and needs to be corrected.

Now let's consider a group of people where the annual income is in the range of one hundred thousand to two hundred thousand dollars—except for that one person who earns a million dollars a year. While this data point is not incorrect, it is an outlier, and needs to be looked at. Depending on your use case, you may need to decide if including this data will skew the results in a way that does not serve your use case.

This brings us to the next step in the data cleaning workflow—Verification. In this step, you inspect the results to establish effectiveness and accuracy achieved as a result of the data cleaning operation. You need to re-inspect the data to make sure the rules and constraints applicable on the data still hold after the corrections you made.

And in the end, it is important to note that all changes undertaken as part of the data cleaning operation need to be documented. Not just the changes, but also the reasons behind making those changes, and the quality of the currently stored data. Reporting how healthy the data is, is a very crucial step.

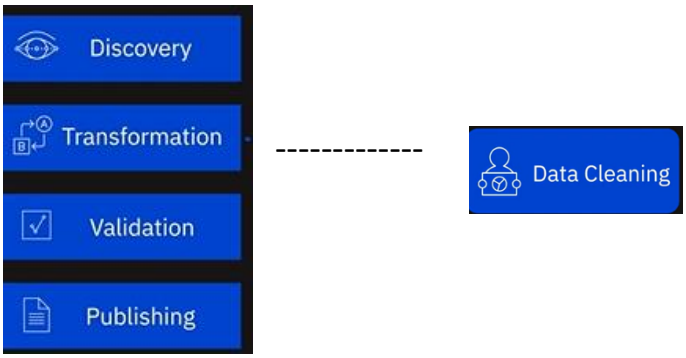
Quality of Data

According to a Gartner report on data quality, poor quality data weakens an organization's competitive standing and undermines critical business objectives.

- Missing data -> False conclusions
- Inconsistent data -> Ineffective decisions *Gartner report on Data Quality*
- Incorrect data ->

Data sets picked up from disparate sources could have a number of issues, including missing values, inaccuracies, duplicates, incorrect or missing delimiters, inconsistent records, and insufficient parameters.

Data Wrangling Process



Data Cleaning Workflow

Data Cleaning Workflow includes:

Inspection

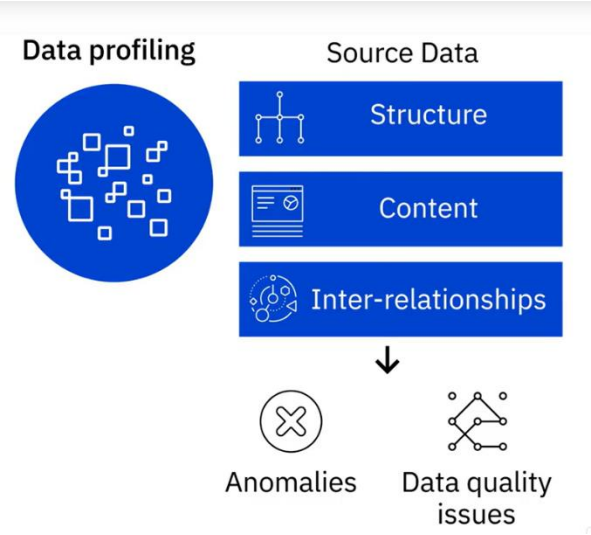
Cleaning

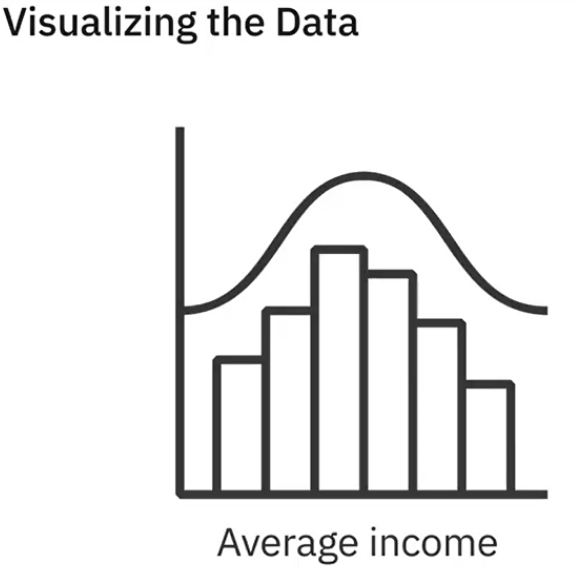
Verification

Inspection

Inspection includes:

- Detecting issues and errors
- Validating against rules and constraints
- Profiling data to inspect source data
- Visualizing data using statistical methods





Cleaning

The techniques you apply for cleaning your dataset will depend on your use case and the type of issues you encounter.

Missing values can cause unexpected or biased results

- Filter out records with missing data
- Source missing information
- Imputate, that is, calculate the missing value based on statistical values

Duplicate data are data points that are repeated in your dataset

- Need to be removed

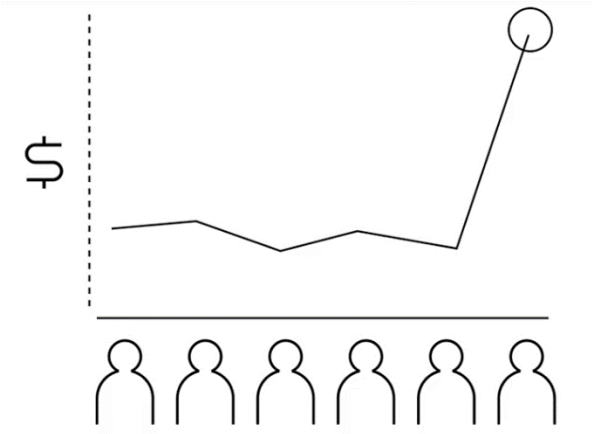
Irrelevant data is data that is not contextual to your use case

Data type conversion is needed to ensure that values in a field are stored as the data type of that field

Standardizing data is needed to ensure date-time formats and units of measurement are standard across the dataset

Syntax errors, such as white spaces, extra spaces, typos, and formats need to be fixed

Outliers need to be examined for accuracy and inclusion in the dataset

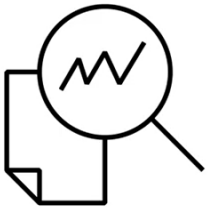


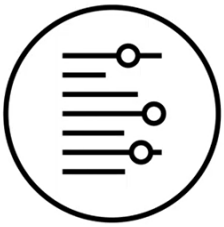
Now let's consider a group of people where the annual income is in the range of one hundred thousand to two hundred thousand dollars—except for that one person who earns a million dollars a year. While this data point is not incorrect, it is an outlier, and needs to be looked at.

Verification

Verification includes:

- Inspecting results to establish effectiveness and accuracy achieved as a result of the data cleaning





It is important to document:

- Changes undertaken as part of the data cleaning operation
- Reasons for undertaking these changes,
- Quality of the currently stored data.

Data Preparation and Reliability

In this segment, data professional share what portion of their job involves gathering, cleaning, and preparing data for analysis. I would say, a relatively big proportion of my job involves gathering, preparing, and cleaning data for analysis. I work at a company with a really great data engineering team. So I don't have to do this kind of work as much as some other data scientists do. But still, any person that is working closely with data, be they're a data scientist, a data analyst, machine learning engineer, really needs to get comfortable understanding where the data comes from. Inevitably, no dataset is perfect. There's always going to be compromises or small errors. So it's really important to spend a significant portion of your time, understanding the underlined data that was used to generate the dataset and what some potential problems might be with that data.

My job as a CPA involves a lot of analysis. Financial statements, account activity, assessing processes, and controls. The gathering piece can be pretty simple as long as, the accounting information resides in a general ledger system or a central repository where the data is easy to gather. Probably, about 30 percent of the job is laying everything out. So when you get into analytics of it, you can just dive right into the meat and potatoes of it. So you need to track the data, make sure it's accurate, make sure things are adding up. Make sure you have all mumps of information. So for example, on financial statements, I need to make sure that people have given me 12 months of [inaudible] statements, I'm not missing any data and that if I am, that I have enough information to be able to project or to forecast or even look back to estimate what was done in the [inaudible] based on what I have. That is definitely helpful.

In this segment, data professionals talk about the steps they take to ensure data is reliable.

One of the essential steps to making sure your data is reliable, is to run summary statistics on individual columns in your data and make sure that they're consistent with reality.

For example, if you have a column somewhere that records visits per month to a website and you run summary statistics on that column, you get the minimum, the mean, the median, the max, and you see something funky like, one month there's negative visits or something like this.

You know, that data isn't reliable. Financial information in particular must be reliable. It must be non-bias. It must be free from error. Those are just a few of the many attributes that are necessary for data to be relied upon. So doing what I call a logic check before you get into the details of a transaction.

Does it make sense at a high level? If you expected top-line revenue to increase, but you see that it has drastically decreased, then figure that part out first.

Is my source correct? Am I running a query in the right period? Am I pulling the right general ledger account? So start there, make sure that basic data integrity questions have been addressed first. Once we know that the data is reliable, then we can start to deep dive into the reviews and form conclusions about the financial performance based on our analysis of the data.

What portion of your job involves gathering, cleaning, and preparing data for analysis?

- A relatively big proportion of my job involves gathering, preparing, and cleaning data for analysis
- no dataset is perfect.
- It's important to spend time to understand the potential problems in a data set

- Gathering data can be simple if the data resides in a central repository
- Preparing data for analysis is about 30% of the job
- It's important to ensure data is accurate and provides all the information you need

What are some essential steps you take to ensure that the data is reliable?

- Run summary statistics on data to make sure it is consistent with reality
- Financial information must be reliable, non-biased and free from error

- Perform a logic check at high level
- Ensure basic data integrity questions are addressed before analysis begins

Reading: Summary and Highlights

In this lesson, you have learned the following information:

Once the data you identified is gathered and imported, your next step is to make it analysis-ready. This is where the process of Data Wrangling, or Data Munging, comes in.
Data Wrangling is an iterative process that involves data
exploration, transformation, and validation.

Transformation of raw data includes the tasks you undertake to:

- Structurally manipulate and combine the data using Joins and Unions.
- Normalize data, that is, clean the database of unused and redundant data.
- Denormalize data, that is, combine data from multiple tables into a single table so that it can be queried faster.
- Clean data, which involves profiling data to uncover quality issues, visualizing data to spot outliers, and fixing issues such as missing values, duplicate data, irrelevant data, inconsistent formats, syntax errors, and outliers.
- Enrich data, which involves considering additional data points that could add value to the existing data set and lead to a more meaningful analysis.

A variety of software and tools are available for the Data Wrangling process. Some of the popularly used ones include Excel Power Query, Spreadsheets, OpenRefine,
Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python, and R, each with their own set of characteristics, strengths, limitations, and applications.

Quiz: Practice Quiz

Bookmarked

Question 1

1/1 point (ungraded)

What is one of the common structural transformations used for combining data from one or more tables?

- ○ Cleaning
- ◉ Joins
- ○ Denormalization
- ○ Normalization

✔

Question 2

1/1 point (ungraded)

What tool allows you to discover, cleanse, and transform data with built-in operations?

- ○ OpenRefine
- ○ Trifacta Wrangler
- ◉ Watson Studio Refinery
- ○ Google DataPrep

✔

Question 3

1/1 point (ungraded)

What is data called that does not fit within the context of the use case?

- ○ Relevant data
- ◉ Irrelevant data
- ○ Missing data
- ○ Duplicate data

✔

Quiz: Graded Quiz

Bookmarked

Graded Quiz due Jul 16, 2022 19:44 +08

Question 1

1/1 point (graded)

What does a typical data wrangling workflow include?

- (●) Validating the quality of the transformed data
- ( ) Using mathematical techniques to identify correlations in data
- ( ) Recognizing patterns
- ( ) Predicting probabilities

✔

Question 2

1/1 point (graded)

OpenRefine is an open-source tool that allows you to:

- (●) Transform data into a variety of formats such as TSV, CSV, XLS, XML, and JSON
- ( ) Enforces applicable data governance policies automatically
- ( ) Automatically detect schemas, data types, and anomalies
- ( ) Use add-ins such as Microsoft Power Query to identify issues and clean data

✔

Question 3

1/1 point (graded)

What is one of the steps in a typical data cleaning workflow?

- (●) Inspecting data to detect issues and errors
- ( ) Establishing relationships between data events
- ( ) Building classification models
- ( ) Clustering data

✔

Question 4

1/1 point (graded)

When you're combining rows of data from multiple source tables into a single table, what kind of data transformation are you performing?

- ( ) Denormalization
- ( ) Normalization
- ( ) Joins
- (●) Unions

✔

Question 5

1/1 point (graded)

When you detect a value in your data set that is vastly different from other observations in the same data set, what would you report that as?

- (●) Outlier
- ( ) Irrelevant data
- ( ) Syntax error
- ( ) Missing value

✔

Module Introduction

In this module, you will learn about the process and tools required for mining and analyzing data. You will learn how statistical methods can be applied to data in order to gain a deeper understanding of the data. You will gain an understanding of what patterns, trends, and correlations are and how they can help you gain deeper insights into your data. You will also learn about the features and characteristics of some of the popular tools used for data mining.

Learning Objectives

After completing this module, you will be able to:

- Explain how Statistical tools and techniques can help create a deeper understanding of what the data means.
- Explain the process and tools used for mining and analyzing data to understand patterns, trends, and correlations that exist in the data.
- List some of the popular data mining tools and describe their features and use cases.

Overview of Statistical Analysis

Statistics

- is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical or quantitative data.

- It's all around us in our day to day lives. Whether we're talking about average income, average age, or highest-paid professions—it's all statistics.

- Today, statistics is being applied across industries for decision-making based on data. For example, researchers using statistics to analyze data from the production of vaccines to ensure safety and efficacy, or companies using statistics to reduce customer churn by gaining greater insight into customer requirements.

Now let's look at what Statistical Analysis is.

Statistical Analysis

- is the application of statistical methods to a sample of data in order to develop an understanding of what that data represents.

- It includes collecting and scrutinizing every data sample in a set of items from which samples can be drawn.

- A **sample,** in Statistics, is a representative selection drawn from a total population, where **population** is a discrete group of people or things that can be identified by at least one common characteristic for purposes of data collection and analysis.

- For example, in a certain use case, population may be all people in a state that have a driving license, and a sample of this population that is a part, or subset, of the population could be men drivers over the age of 50.

Statistical methods are mainly useful to ensure that data is interpreted correctly, and apparent relationships are meaningful and not just happening by chance. Whenever we collect data from a sample,

there are two different types of statistics we can run.

Descriptive statistics

- to summarize information about the sample; and Inferential statistics to make inferences or generalizations about the broader population.
- Descriptive Statistics enables you to present data in a meaningful way allowing simpler interpretation of the data.
- Data is described using summary charts, tables, and graphs without any attempts to draw conclusions about the population from which the sample is taken. T

- The objective is to make it easier to understand and visualize raw data without making conclusions regarding any hypotheses that were made.
- For example, we want to describe the English test scores in a specific class of 25 students. We record the test scores of all students, calculate the summary statistics, and produce a graph.

Some of the common measures of Descriptive Statistical Analysis include **Central Tendency, Dispersion,** and **Skewness**:

Central Tendency,

- or locating the center of a data sample.
- Some of the common measures of central tendency include **mean, median,** and **mode.**
- These measures tell you where most values in your dataset fall.
- So, in the earlier example, the **mean** score, or the mathematical average, of the class of 25 students would be the sum total of the scores of all 25 students, divided by 25, that is, the number of students.
- If you order the above dataset from the smallest score value to the highest score value of the 25 students and pick the middle value— that is the value with 12 values to the left and 12 values to the right of a score value, that score value would be the **median** for this dataset. If 12 students have scored less than 75%, and 12 students have scored greater than 75%, then the median is 75. Median is unique for each dataset and is not affected by outliers.
- **Mode** is the value that occurs most frequently in a set of observations. For example, if the most common score in this group of 25 students is 72%, then that is the mode for this dataset.

So, you can see how looking at your dataset through these values can help you get a clearer understanding of your dataset. Dispersion is the measure of variability in a dataset.

Common measures of statistical dispersion are **Variance, Standard Deviation, and Range**.

Variance

- defines how far away the data points fall from the center, that is, the distribution of values.
- When a distribution has lower variability, the values in a dataset are more consistent. However, when the variability is higher, the data points are more dissimilar, and extreme values become more likely. Understanding variability can help you grasp the likelihood of an event happening.

Standard deviation

- tells you how tightly your data is clustered around the mean.

Range

- gives you the distance between the smallest and largest values in your datasets.

Skewness

- is the measure of whether the distribution of values is symmetrical around a central value or skewed left or right. Skewed data can affect which types of analyses are valid to perform.

These are some of the basic and most commonly used descriptive statistics tools, but there are other tools as well, for example, using correlation and scatterplots to assess the relationships of paired data. The second type of statistical analysis is Inferential Statistics.

Inferential statistics

- takes data from a sample to make inferences about the larger population from which the sample was drawn.
- Using methods of inferential statistics you can draw generalizations that apply the results of the sample to the population as a whole.
- Some common methodologies of Inferential Statistics include **Hypothesis Testing**, **Confidence Intervals**, and **Regression Analysis**:

Hypothesis Testing—

- For example, for studying the effectiveness of a vaccine by comparing outcomes in a control group, hypothesis tests can tell you whether the efficacy of a vaccine observed in a control group is likely to exist in the population as well.

Confidence Intervals

- incorporate the uncertainty and sample error to create a range of values the actual population value is like to fall within.

Regression Analysis

- incorporates hypothesis tests that help determine whether the relationships observed in the sample data actually exist in the population rather than just the sample. T

There are various software packages to perform statistical data analysis, such as **Statistical Analysis System (or SAS), Statistical Package for the Social Sciences (or SPSS), and Stat Soft.**

Statistics form the core of data mining by:

- Providing measures and methodologies necessary for data mining; and
- Identifying patterns that help identify differences between random noise and significant findings.

Both data mining, which we will learn more about in this course, and Statistics, as techniques of data analysis, help in better decision-making.

Overview of Statistical Analysis

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical or quantitative data.

Everyday examples of statistics at work:

- Calculations such as average income, average age, highest-paid professions
- Analyzing vaccine data to ensure safety and efficacy
- Gaining greater insight into customer requirements to reduce customer churn

Statistical Analysis

Statistical Analysis is the application of statistical methods to a sample of data in order to develop an understanding of what that data represents.

Sample - A representative selection drawn from a total population.

Population - A discrete group of people or things that can be identified by at least one common characteristic for purposes of data collection and analysis.

Statistical methods help ensure:

- Data is interpreted correctly
- Apparent relationships are meaningful

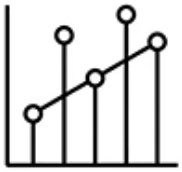
Types of Statistics:

- Descriptive statistics – Summarizing information about the sample
- Inferential statistics – Making inferences or generalizations about the broader population

Descriptive Statistics:

- Enables you to present data in a meaningful way
- Allows for simpler interpretation of data
- Does not attempt to draw conclusions about the population from which the sample is taken

Descriptive Statistics



Common measures of Descriptive Statistical Analysis:

- Central Tendency
- Dispersion
- Skewness

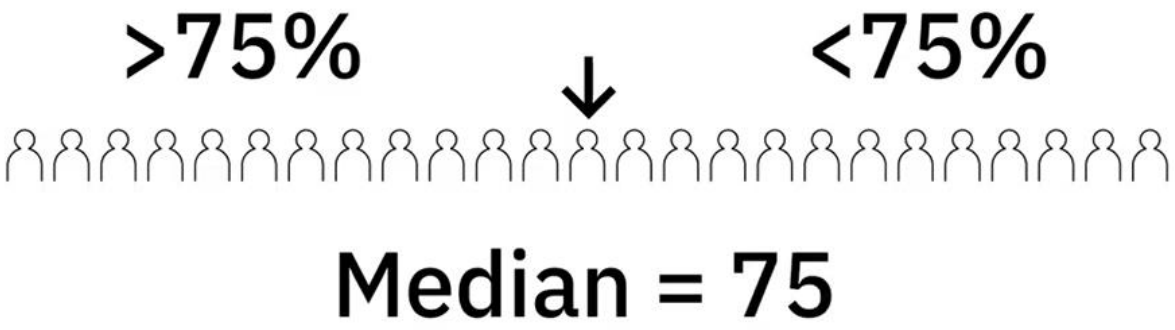
Central Tendency — locating the center of a data sample. Common measures include: Mean, Median, and Mode.

Measures of Central Tendency

1.

Mean =
Sum total
of scores
÷ 25

2.



Median = 75

3.

Mode= 72

Dispersion is the measure of variability in a dataset. Common measures of statistical dispersion are:

- Variance
- Standard Deviation
- Range

- **Variance** defines how far away the data points fall from the center.

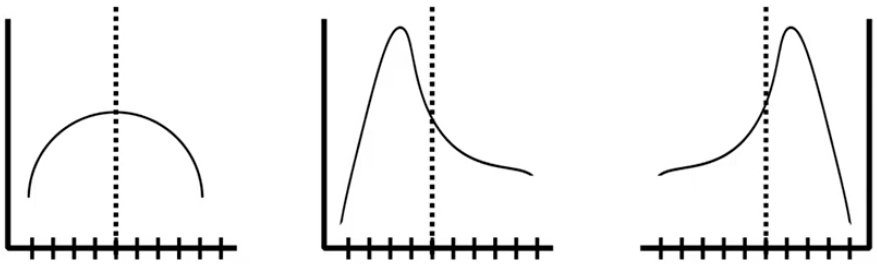
  Lower variability > Consistent values in a dataset

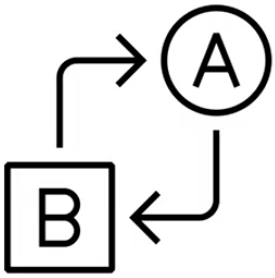
  Higher variability > Dissimilar values with likelihood of extreme values

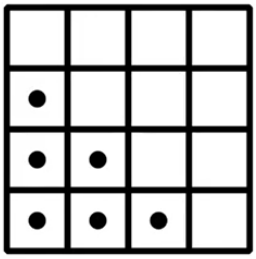
- **Standard Deviation** tells you how tightly your data is clustered around the mean.
- **Range** gives you the distance between the smallest and largest values in your datasets.

Skewness is the measure of whether the distribution of values is symmetrical around a central value or skewed left or right.

Skewed data can affect which types of analyses are valid to perform.







Inferential Statistics

Common methodologies of Inferential Statistics include:

- **Hypothesis Testing**—For example, for studying the effectiveness of a vaccine by comparing outcomes in a control group, hypothesis tests can tell you whether the efficacy of a vaccine observed in a control group is likely to exist in the population as well.

- **Confidence Intervals** incorporate the uncertainty and sample error to create a range of values the actual population value is likely to fall within.
- **Regression Analysis** incorporates hypothesis tests that help determine whether the relationships observed in the sample data actually exist in the population rather than just the sample.

Statistical Software Packages:

Conclusion:

Statistics combined with Data Mining, together help better decision-making

- Providing measures and methodologies necessary for data mining
- Identifying patterns that help identify differences between random noise and significant findings

Data Mining

Data mining or

- the process of extracting knowledge from data, is the heart of the data analysis process.
- It is an interdisciplinary field that involves the use of pattern recognition technologies, statistical analysis and mathematical techniques.
- Its goal is to identify correlations in data, find patterns and variations. Understand trends and predict probabilities.

You'll hear about patterns and trends frequently in the context of data analysis, so let's first understand these concepts.

Pattern recognition

- is the discovery of regularity's or commonality's in data.
- Consider the log data for logins to an application in an organization. I
- It contains information such as the username, login timestamp, time spent in each login session, and activities performed.
- When we analyze this data to gain insights into the habits or behaviors of users, for example, the time of the day when maximum users tend to login or user roles that typically spend the maximum hours logged into the application or modules in the workflow application that are being used where examining the data manually or through tools to uncover patterns hidden in the data.

A trend, on the other hand,

- is the general tendency of a set of data to change overtime.
- For example, global warming in the short term, like a year on year basis temperatures may remain the same or go up or down by a few degrees, but the overall global temperatures continue to increase overtime, making global warming a trend.

Data mining has applications across industries and disciplines

- For example, profiling customer behaviors needs and disposable income in order to offer targeted campaigns, financial institutions, tracking customer transactions for unusual behaviors, and flagging fraudulent transactions using data mining models.
- The use of statistical models to predict a patients likelihood for specific health conditions and prioritizing treatment.
- Accessing performance data of students to predict achievement levels and make a focused effort to provide support where required.
- Helping investigation agencies deploy police force where the likelihood of crime is higher and aligning supply and logistics with demand forecasts.

There are several techniques you can use to detect patterns and build accurate models for discovery, be it descriptive, diagnostic, predictive, or prescriptive modeling.

Let's understand some of the most commonly used techniques.

1. Classification is a technique that classifies attributes into target categories, for example, classifying customers into low, medium, or high spenders based on how much they earn.

2. Clustering is similar to classification, but involves grouping data into clusters so they can be treated as groups. For example, clustering customers based on geographic regions anomaly

3. Outlier detection is a technique that helps find patterns and data that are not normal or unexpected. For example, spikes in the usage of a credit card that can flag possible misuse.

4. Association rule mining is a technique that helps establish our relationship between two data events. For example, the purchase of a laptop being frequently accompanied by the purchase of a cooling pad.

5. Sequential patterns is the technique that traces a series of events that take place in a sequence. For example, tracing a customer shopping trail from the time they log into an online store to the time they log out.

6. Affinity grouping is a technique used to discover Co occurrence in relationships. This technique is widely used in on line stores for cross selling and up selling their products by recommending products to people based on the purchase history of other people who purchased the same item.

7. Decision trees help build classification models in the form of a tree structure with multiple branches, where each branch represents a probable occurrence. This technique helps to build a clear understanding of the relationship between input and output.

8. Regression is a technique that helps identify the nature of the relationship between two variables, which could be causal or correlational. For example, based on factors such as location and covered area, a regression model could be used to predict the value of a house.

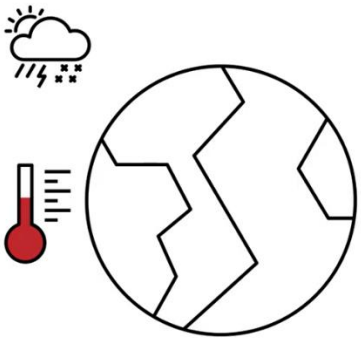
Data mining essentially helps separate the noise from the real information and helps businesses focus their energies on only what is relevant.

Data Mining

- The process of extracting knowledge from data
- An interdisciplinary field that involves the use of pattern recognition technologies, statistical analysis and mathematical techniques
- Aims to identify correlations in data, find patterns and variations, understand trends and predict probabilities

Patterns and Trends

Pattern recognition is the discovery of regularities, or commonalities, in data.



A Trend is the general tendency of a set of data to change over time

Year	Temp
2005	51° F
2010	52° F
2015	53° F
2020	54° F

For example, global warming in the short term, like a year on year basis temperatures may remain the same or go up or down by a few degrees, but the overall global temperatures continue to increase overtime, making global warming a trend.

Applications of Data Mining

Data Mining has applications across industries and disciplines

1.

- Profiling customer behaviors, needs, and disposable income in order to offer targeted campaigns

2.

- Financial institutions tracking customer transactions for unusual behaviors and flagging fraudulent transactions using data mining models

3.

- The use of statistical models to predict a patient's likelihood for specific health conditions and prioritizing treatment

4.

- Accessing performance data of students to predict achievement levels and make a focused effort to provide support where required

5.

- Helping investigation agencies deploy police force where the likelihood of crime is higher

6.

- Aligning supply and logistics with demand forecasts

Data Mining Techniques

Descriptive

Diagnostic

Predictive

Prescriptive modeling

Some commonly used data mining techniques:

- **Classification** - Classifying attributes into target categories

for example, classifying customers into low, medium, or high spenders based on how much they earn.

- **Clustering** - Involves grouping data into clusters so they can be treated as groups

for example, clustering customers based on geographic regions

- **Anomaly or Outlier Detection** - Finding patterns in data that are not normal or unexpected

for example, spikes in the usage of a credit card that can flag possible misuse.

- **Association Rule Mining** - Establishing a relationship between two data events

for example, the purchase of a laptop being frequently accompanied by the purchase of a cooling pad.

- **Sequential Patterns** - Tracing a series of events that take place in a sequence

for example, tracing a customer shopping trail from the time they log into an online store to the time they log out

- **Affinity Grouping** - Discovering co-occurrence in relationships

This technique is widely used in on line stores for cross selling and up selling their products by recommending products to people based on the purchase history of other people who purchased the same item.

- **Decision trees** - Building classification models in the form of a tree structure with multiple branches, where each branch represents a probable occurrence

This technique helps to build a clear understanding of the relationship between input and output.

- **Regression** - Identifying the nature of the relationship between two variables, which could be causal or correlational

For example, based on factors such as location and covered area, a regression model could be used to predict the value of a house.

Data mining essentially helps separate the noise from the real information and helps businesses focus their energies on only what is relevant.

Tools for Data Mining

In this video, we will learn about some of the commonly used software and tools for data mining, such as: Spreadsheets, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio; and SAS.

Spreadsheets, such as Microsoft Excel and Google Sheets, are commonly used for performing basic data mining tasks.

- Spreadsheets can be used to host data that has been exported from other systems in an easily accessible and easy-to-read format.
- You can pivot tables to showcase specific aspects of your data, which is vital when you have huge amounts of data to sort through and analyze.
- They also make it relatively easier to make comparisons between different sets of data. Add-ins available for Excel, such as the Data Mining Client for Excel, XLMiner, and KnowledgeMiner for Excel, allow you to perform common mining tasks, such as classification, regression, association rules, clustering, and model building.

GoogleSheets also has an array of add-ons that can be used for analysis and mining, such as Text Analysis, Text Mining, Google Analytics.

R

- is one of the most widely used languages for performing statistical modeling and computations by statisticians and data miners.
- R is packaged with hundreds of libraries explicitly built for data mining operations such as regression, classification, data clustering, association rule mining, text mining, outlier detection, and social network analysis.
- Some of the popular R packages include tm and twitteR. tm, a framework for text mining applications within R, provides functions for text mining.
- twitteR provides a framework for mining tweets. R Studio is a popularly used open-source Integrated Development Envionrment (or IDE) for working with the R programming language.

Python libraries like Pandas and NumPy are commonly used for Data Mining.
Pandas

- is an open-source module for working with data structures and analysis.
- It is possibly one of the most popular libraries for data analysis in Python.
- It allows you to upload data in any format and provides a simple platform to organize, sort, and manipulate that data.
- Using Pandas, you can: perform basic numerical computations such as mean, median, mode, and range; calculate statistics and answer questions regarding correlation between data and distribution of data; explore data visually and quantitatively; visualize data with help from other Python libraries.

NumPy is a tool for mathematical computing and data preparation in Python. NumPy offers a host of built-in functions and capabilities for data mining.

Jupyter Notebooks have become the tool of choice for Data Scientists and Data Analysts when working with Python to perform data mining and statistical analysis.

SPSS stands for Statistical Process for Social Sciences.

- While the name suggests its original usage in the field of Social Sciences, it is popularly used for advanced analytics, text analytics, trend analysis, validation of assumptions, and translation of business problems into data science solutions.
- SPSS is closed source and requires a license for use. SPSS has an easy to use interface that requires minimal coding for complex tasks.
- It comprises of efficient data management tools and is popular because of its in-depth analysis capabilities and accurate data results.

IBM Watson Studio,

- included in the IBM Cloud Pak for Data, leverages a collection of open source tools such as J**upyter notebooks,** and extends them with closed source IBM tools that make it a powerful environment for data analysis and data science.
- It is available through a web browser on the public cloud, private cloud, and as a desktop app.

Watson Studio

- enables team members to collaborate on projects, that can range from simple exploratory analysis to building machine learning and AI models.
- It also includes SPSS Modeller flows that enable you to quickly develop predictive models for your business data.

SAS Enterprise Miner

- is a comprehensive, graphical workbench for data mining. It provides powerful capabilities for interactive data exploration, which enables users to identify relationships within data.
- SAS can manage information from various sources, mine and transform data, and analyze statistics. It offers a graphical user interface for non-technical users.
- With SAS, you can: identify patterns in the data using a range of available modeling techniques; explore relationships and anomalies in data; analyze big data; validate the reliability of findings from the data analysis process. SAS is very easy to use because of its syntax and is also easy to debug.
- It has the ability to handle large databases and offers high security to its users.

In this video, we have learned about just a few of the data mining tools available today. Your decision regarding the best tool for your needs will be driven by the data size and structures the tool supports, the features it offers, its data visualization capabilities, infrastructure needs, ease of use, and learnability. It's fairly common to use a combination of data mining tools to meet all your needs.

Tools for Data Mining

Some of the commonly used software and tools for data mining

- Spreadsheets
- R-Language
- Python
- IBM SPSS Statistics
- IBM Watson Studio
- SAS

Spreadsheets

Spreadsheets, such as Microsoft Excel and Google Sheets, are commonly used for performing basic data mining tasks.

Spreadsheets can be used for:
hosting data that has been exported from other systems in an easily accessible and easy-to-read format.

Creating pivot tables to showcase specific aspects of your data

Drawing comparisons between sets of data

Excel add-ins, such as:

Data Mining Client XLMiner KnowledgeMiner

Allow you to perform common mining tasks such as classification, regression, association rules, clustering and model building.

Google sheets also has an array of add-ons that can be used for analysis and mining, such as Text Analysis, text Mining, Google Analytics.

R Language

R is one of the most widely used languages for performing statistical modeling and computations by statisticians and data miners.

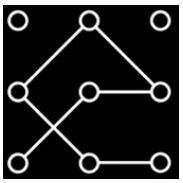
Using R libraries you can perform data mining operations such as:

tm A framework for text mining applications within R

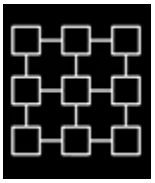
twitteR A framework for mining tweets

RStudio is a popularly used open-source Integrated Development Environment (or IDE) for working with the R programming language.

Python



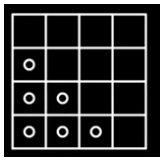
Open-source module for working with data structures and analysis



Allows you to upload data in any format and provides a simple platform to organize, sort, and manipulate that data.

Using Pandas, you can:

- Perform basic numerical computations such as mean, median, mode, and range
- Calculate statistics and answer questions regarding correlation between data and distribution of data
- Explore data visually and quantitatively
- Visualize data with help from other Python libraries



•A tool for mathematical computing and data preparation in Python

- Offers a host of built-in functions and capabilities for data mining.

Jupyter Notebooks have become the tool of choice for Data Scientists and Data Analysts when working with Python to perform data mining and statistical analysis.

IBM SPSS Statistics

SPSS stands for Statistical Process for Social Sciences

- Popularly used for advanced analytics, text analytics, trend analysis, validation of assumptions, and translation of business problems into data science solutions
- Is closed-source
- Requires a license for use
- Has an easy to use interface

IBM Watson Studio

IBM Watson Studio, included in the IBM Cloud Pak for Data, leverages a collection of open source tools such as Jupyter notebooks, and extends them with closed source IBM tools that make it a powerful environment for data analysis and data science.

Is available through a web browser on the public cloud, private cloud, and as a desktop app. Watson Studio

Enables team members to collaborate on projects, that can range from simple exploratory analysis to building machine learning and AI models.

It also includes SPSS Modeller flows that enable you to quickly develop predictive models for your business data.

SAS

SAS Enterprise Miner is a comprehensive, graphical workbench for data mining

Provides powerful capabilities for interactive data exploration

Can manage information from various sources, mine and transform data, and analyze statistics

Offers a graphical user interface for non-technical users.

With SAS, you can

- Identify patterns in the data using a range of available modeling techniques
- Explore relationships and anomalies in data
- Analyze big data
- Validate the reliability of findings from the data analysis process

Conclusion:

Key considerations for selecting the right data mining tool:

Data size and structures supported by the tool;
Key features;
Data Visualization capabilities;
Infrastructure needs;
Easy of use and
Learnability

Reading: Summary and Highlights

In this lesson, you have learned the following information:

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical or quantitative data.

Statistical Analysis involves the use of statistical methods in order to develop an understanding of what the data represents.

Statistical Analysis can be:

Descriptive; that which provides a summary of what the data represents. Common measures include Central Tendency, Dispersion, and Skewness.

Inferential; that which involves making inferences, or generalizations, about data. Common measures include Hypothesis Testing, Confidence Intervals, and Regression Analysis.

Data Mining, simply put, is the process of extracting knowledge from data. It involves the use of pattern recognition technologies, statistical analysis, and mathematical techniques, in order to identify correlations, patterns, variations, and trends in data.

There are several techniques that can help mine data, such as, classifying attributes of data, clustering data into groups, establishing relationships between events, variables, and input and output.

A variety of software and tools are available for analyzing and mining data. Some of the popularly used ones include Spreadsheets, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio, and SAS, each with their own set of characteristics, strengths, limitations, and applications.

Quiz: Practice Quiz

Bookmarked

Question 1

1/1 point (ungraded)

What is one of the common measures of Central Tendency?

- ○ Classification
- ○ Variance
- ○ Regression
- ◉ Mean

✔

Question 2

1/1 point (ungraded)

What technique is used to help identify the nature of the relationship between two variables?

- ○ Classification
- ○ Clustering
- ◉ Regression
- ○ Anomaly Detection

✔

Question 3

1/1 point (ungraded)

What Python libraries are commonly used for data mining?

- ☑ Pandas
- ☐ Tm
- ☑ NumPy
- ☐ twitter

✔

Quiz: Graded Quiz

Bookmarked

Graded Quiz due Jul 19, 2022 03:44 +08

Question 1

1/1 point (graded)

What is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical or quantitative data?

- ○ Pie
- ○ Calculus
- ○ Algebra
- ◉ Statistics

✔

Question 2

1/1 point (graded)

Data Mining is defined as the process of:

- ○ Filtering data based on pre-defined criteria
- ○ Identifying errors in data
- ○ Preparing raw data for analysis
- ◉ Extracting knowledge from data

✔

Question 3

1/1 point (graded)

What type of data mining operations was R specifically built to handle?

- ◉ Classification of data
- ○ Calculating mean, median, and mode
- ○ Sorting
- ○ Filtering

✔

Question 4

1/1 point (graded)

When you're calculating the middle value of a data field in a data set, what are you really calculating?

- ○ Mode
- ◉ Median
- ○ Mean
- ○ Average

✔

Question 5

1/1 point (graded)

What is the general tendency of a set of data to change over time called?

- ○ Anomaly
- ○ Variation
- ○ Pattern
- ◉ Trend

✔

Module Introduction

In this module, you will learn how to effectively visualize data in order to communicate your findings to stakeholders in ways that impact decision-making.You will also learn about the features and characteristics of the various data visualization tools and how best you can leverage them for telling a compelling story with data.

Learning Objectives

After completing this module, you will be able to:

Describe the process that can help you share your insights with your stakeholders in a way that it impacts decision-making.
Explain how to choose the best visualization for your data and the possibilities offered by some of the most popular data visualization and dashboarding tools.
Understand how you can a compelling and convincing story with your data.

Overview of Communicating and Sharing Data Analysis Findings

The data analysis process begins with understanding the problem that needs to be solved and the desired outcome that needs to be achieved. And it ends with communicating the findings in ways that impact decision making.

Data projects are the result of a collaborative effort spread across business functions involving people with multi-disciplinary skills, with the findings being incorporated into a larger business initiative. The success of your communication depends on how well others can understand and trust your insights to take further action. So, as data analysts, you need to tell the story with your data by visualizing the insights clearly and creating a structured narrative explicitly targeted at your audience.

Before you begin to create the communication, you need to reconnect with your audience. Begin by asking yourself these questions

Who is my audience?
What is important to them?
What will help them trust me?

Your audience is mostly going to be a diverse group—in terms of the business functions they represent, whether they play an operational or strategic role in the organization, how impacted are they by the problem, and other such factors.

Your presentation needs to be framed around the level of information your audience already has.
Based on your understanding of the audience, you will decide what, and how much, information is essential to enable a better understanding of your findings. It's tempting to bring out all the data that you've been working with, but you have to consider what pieces are more important to your audience than others.

A presentation is not a data dump. Facts and figures alone do not influence decisions and move people to action. You have to tell a compelling story. Include only that information as is needed to address the business problem. Too much information will have your audience struggling to understand the point you're making.

Begin your presentation by demonstrating you runderstanding of the business problem to your audience.
It's easy to fall back on the assumption that we all know what we're here for, but reflecting your understanding of the problem that needs to be solved, and the outcome that needs to be achieved, is a great first step in winning their attention and starting with trust. Speaking in the language of the organization's business domain is another important factor in building a connect between you and your audience.

The next step in designing your communication is to structure and organize your presentation for maximum impact. Reference the data you have collected. Remember that the data, the very basis of everything that you are communicating, is like a black box for the audience. If you're unable to establish the credibility of your data, people don't know that they can trust your findings. Share your data sources, hypotheses, and validations.

Work towards establishing credibility of your findings along the way – don't gloss over any key assumptions made during the analysis.
Organize information into logical categories based on the information you have—do you have both qualitative and quantitative information, for example?
Be deliberate in taking a top-down or bottom-up approach in your narrative. Both can be effective—depends on your audience and use case.
Be consistent in your approach. It's important to determine what communication formats will be most useful to your audience.
Do they need to take away an executive summary, a fact sheet, or a report?
How is your audience going to use the information you have presented, that should determine the formats you choose. Insights must be explained in a way that inspires action.

If your audience doesn't grasp the significance of your insight or are unconvinced of its utility, the insight will not drive any value. A thousand-word essay will not have the same impact as a visual in creating a clear mental image in the minds of your audience.

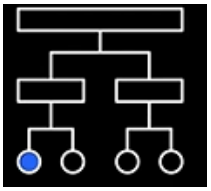
A powerful visualization tells a story through the graphical depiction of facts and figures. Data visualizations—graphs, charts, diagrams—are a great way to bring data to life. Whether you're showing a comparison, a relationship, distribution, or composition, you have tools that can help you show patterns and conclusions about hypotheses.

Data has value through the stories that it tells. Your audience must be able to trust you, understand you, and relate to your findings and insights. Establishing credibility of your findings, presenting the data within a narrative, and supporting it through visual impressions, you can help your audience drive valuable insights.

Overview of Communicating and Sharing Data Analysis Findings

Data Analysis Process

Understanding the problem

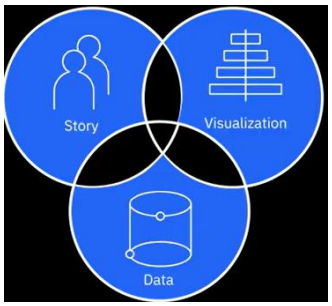


Communicating the findings

Data projects involve

- ✓ A collaborative effort spread across business functions
- ✓ People with multi-disciplinary skills
- ✓ Findings being incorporated into a larger business initiative.

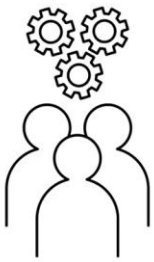
The success of your communication depends on how well others can understand and trust your insights to take further action.



So, as data analysts, you need to tell the story with your data by visualizing the insights clearly and creating a structured narrative explicitly targeted at your audience. Before you begin to create the communication, you need to reconnect with your audience.

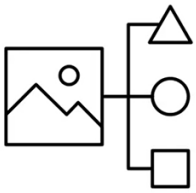
Who is my audience?

A diverse group of people representing different business functions and roles



What is important to them?

Understanding the information needs of your audience will help you decide what, and how much, information is essential to enable a better understanding of your findings.

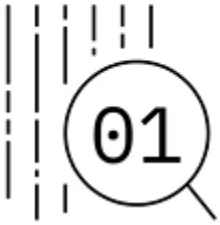


Your presentation needs to be framed around the level of information your audience already has. Based on your understanding of the audience, you will decide what, and how much, information is essential to enable a better understanding of your findings. It's tempting to bring out all the data that you've been working with, but you have to consider what pieces are more important to your audience than others.

A presentation is not a data dump. Facts and figures alone do not influence decisions and move people to action. You have to tell a compelling story. Include only that information as is needed to address the business problem. Too much information will have your audience struggling to understand the point you're making.

What will help them trust me?

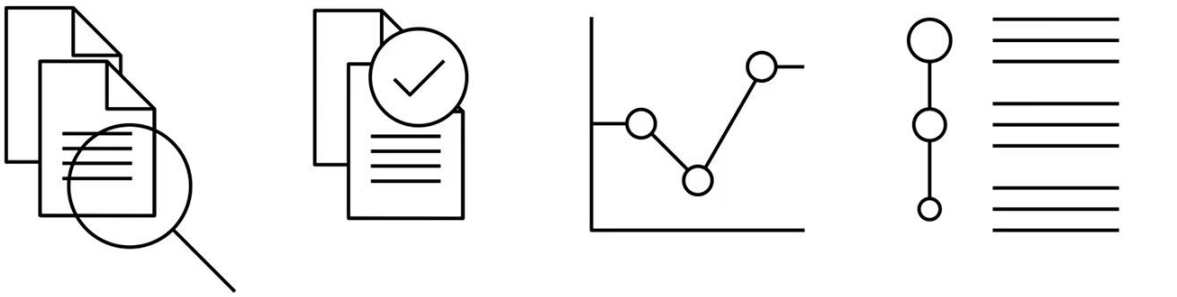
Begin your presentation by demonstrating your understanding of the business problem to your audience.
Speak in the language of the organization's business domain.



Begin your presentation by demonstrating yourunderstanding of the business problem to your audience. It's easy to fall back on the assumption that we all know what we're here for, but reflecting your understanding of the problem that needs to be solved, and the outcome that needs to be achieved, is a great first step in winning their attention and starting with trust. Speaking in the language of the organization's business domain is another important factor in building a connect between you and your audience.

Structure your presentation

The next step in designing your communication is to structure and organize your presentation for maximum impact.



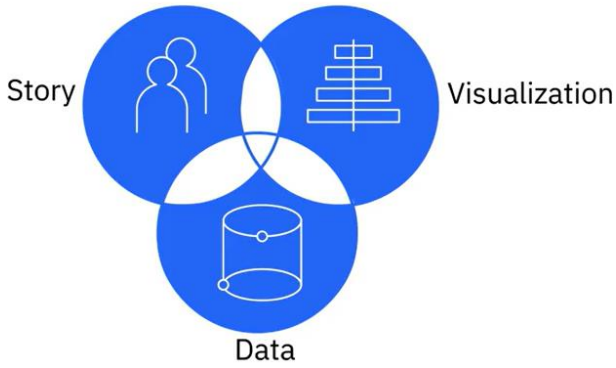
A thousand-word essay will not have the same impact as a visual in creating a clear mental image in the minds of your audience.

The Role of Visuals

A powerful visualization tells a story through the graphical depiction of facts and figures.

Trust, Understanding, Relatability

- Establish credibility of your findings
- Present data within a narrative
- Support the narrative with visual



Viewpoints: Storytelling in Data Analysis

What role does storytelling play in Data Analysis?

- ✓ Storytelling with data is a critical skill for data analysts
- ✓ It's important to tell a clear, concie and compelling story to convince people to take action
- ✓ Develop a story for your data set to understand your data better
- ✓ Find the balance between telling a simple story and conveying the complexities of the data
- ✓ It doesn't matter what information you have if you can't communicate it effectively to your audience
- ✓ The best way to communicate your information is through visuals and telling a story
- ✓ Storytelling is an essential skill set – the last mile in delivery
- ✓ The ability to extract value from data and to tell a compelling story with data is critical
- ✓ Storytelling is crucial to data analytics
- ✓ Stories is how you convey your message
- ✓ A compelling story helps your audience resonate with your findings
- ✓ People remember stories
- ✓ Stories help build an emotional connect and drive people into action

Introduction to Data Visualization

Data visualization is the discipline of communicating information through the use of visual elements such as graphs, charts, and maps.

- Its goal is to make information easy to comprehend, interpret, and retain. Imagine having to look through thousands of rows of data to draw interpretations and compare that to a visual representation of that same data summarizing the findings.
- Using data visualization, you can provide a summary of the relationships, trends, and patterns hidden in the data, which, if not impossible, would be very hard to decipher from a data dump.
- For data visualization to be of value, you have to choose the visualization that most effectively delivers your findings to your audience. And for that, you need to begin by asking yourself some questions.

What is the relationship that I am trying to establish?

Do I want to compare the relative proportion of the sub-parts of a whole, for example, the contribution of different product lines in the total revenue of the company?

Do I want to compare multiple values, such as the number of products sold, and revenues generated over the last three years?

Or, do I want to analyze a single value over time, which in this example could mean how the sale of one specific product has changed over the last three years.

Do I need my audience to see the correlation between two variables? The correlation between weather conditions and bookings in a ski resort, for example.

Do I want to detect anomalies in data—for example, finding values in data that could potentially skew the findings?

What is the question I'm trying to answer is not just an overarching question in the data visualization design and process—you need to be able to answer this question for your audience with every dataset and information that you visualize.

You also need to consider whether the visualization needs to be static or interactive.

An interactive visualization, for example, can allow you to change values and see the effects on a related variable in real-time. So, think about the key takeaway for your audience, anticipate their information needs and the questions they may have, and then plan the visualization that delivers your message clearly and impactfully.

Let's look at some basic examples of the types of graphs you can create for visualizing your data.

Bar Charts are great for comparing related data sets or parts of a whole. For example, in this bar chart, you can see the population numbers of 10 different countries and how they compare to one another.

Column Charts compare values side-by-side. You can use them quite effectively to show change over time. For example, showing how page views and user sessions time on your website is changing on a month-to-month basis. Although alike, except for the orientation, bar charts and column charts cannot always be used interchangeably. For example, a column chart may be better suited for showing negative and positive values.

Pie Charts show the breakdown of an entity into its sub-parts and the proportion of the sub-parts in relation to one another. Each portion of the pie represents a static value or category, and the sum of all categories is equal to hundred percent. In this example, in a marketing campaign with four marketing channels—social sites, native advertising, paid influencers, and live events—you can see the total number of leads generated per channel.

Line Charts display trends. They're great for showing how a data value is changing in relation to a continuous variable. For example, how has the sale of your product, or multiple products, changed over time, where time is the continuous variable. Line charts can be used for understanding trends, patterns, and variations in data; also, for comparing different but related data sets with multiple series.

Data visualization can also be used to build dashboards.

- Dashboards organize and display reports and visualizations coming from multiple data sources into a single graphical interface. You can use dashboards to monitor daily progress or the overall health of a business function or even a specific process.

Dashboards

- can present both operational and analytical data. For example, you could have a marketing dashboard using which you monitor your current marketing campaign for reach-outs, queries generated, and sales conversions, in real-time.
- As part of the same dashboard, you could also be seeing how the conversion rate of this campaign compares to the conversion rate of some of the successfully run campaigns in the past.
- Dashboards are a great tool to present a bird's eye view of the complete picture while also allowing you to drill down into the next level of information for each parameter.
- Dashboards: are easy to comprehend by an average user make collaboration easy between teams; and allow you to generate reports on the go.
- Using dashboards, you can see the result of variations in data and metrics almost instantly—and this can help you evaluate a situation from multiple perspectives, on the go, without having to go back to the drawing board.

Introduction to Data Visualization

Overview

Data visualization is the discipline of communicating information through the use of visual elements such as graphs, charts. And maps its goal is to make information easy to comprehend, interpret, and retain.

Choosing Appropriate Visualizations

For data visualization to be of value, choose the visualization that effectively delivers your findings to your audience.

- What is the relationship that I am trying to establish?
- Do I want to compare multiple values, such as the number of products sold, and revenues generated over the last three years?
- Do I need my audience to see the correlation between two variables?
- Do I want to detect anomalies in data?

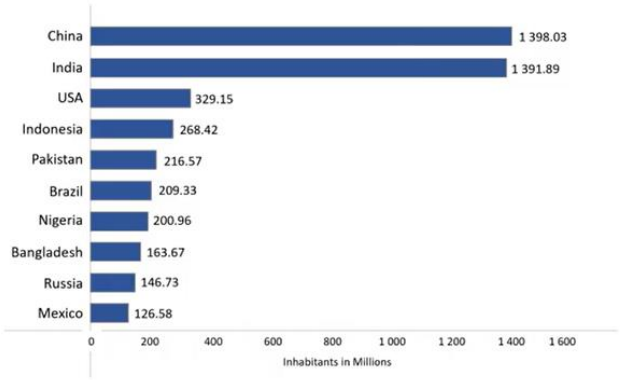
What is the question I’m trying to answer?

You need to be able to answer this question for your audience with every dataset and information that you visualize.

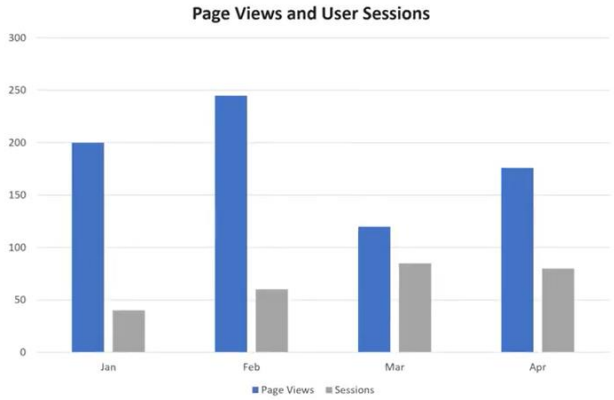
- What should be the key takeaway for my audience?
- What does my audience need to know?
- What are the questions they have?

Common types of graphs

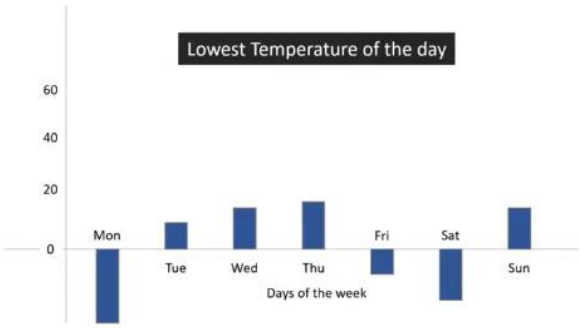
1) Bar Charts are great for comparing related data sets or parts of a whole.



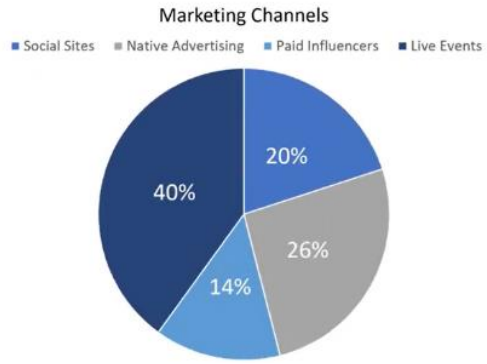
2) Column Charts compare values side-by-side. You can use them quite effectively to show change over time.



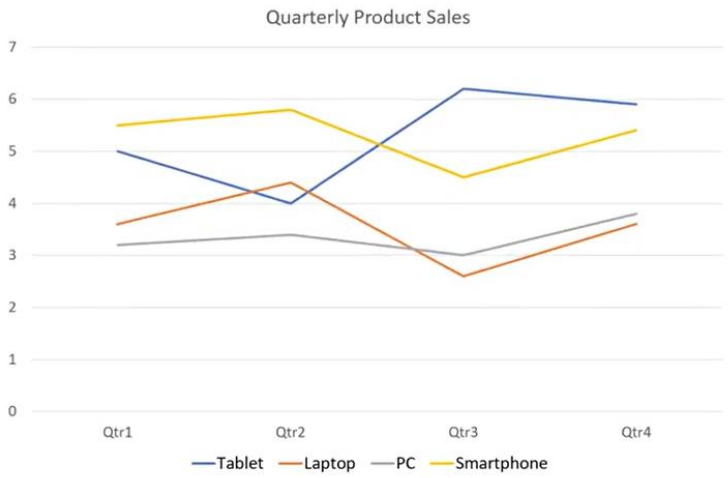
2) Column Charts compare values side-by-side. You can use them quite effectively to show change over time.



3) Pie Charts show the breakdown of an entity into its sub-parts and the proportion of the sub-parts in relation to one another. Each portion of the pie represents a static value or category, and the sum of all categories is equal to hundred percent.



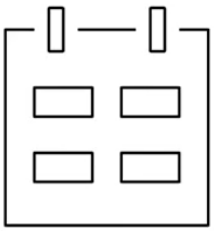
4) Line Charts display trends. They're great for showing how a data value is changing in relation to a continuous variable.

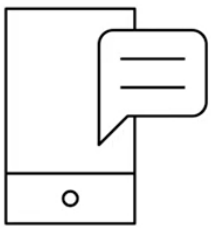


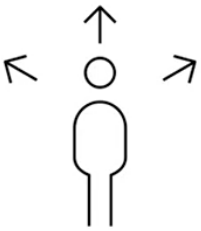
Dashboards

Dashboards organize and display reports and visualizations coming from multiple data sources into a single graphical interface.

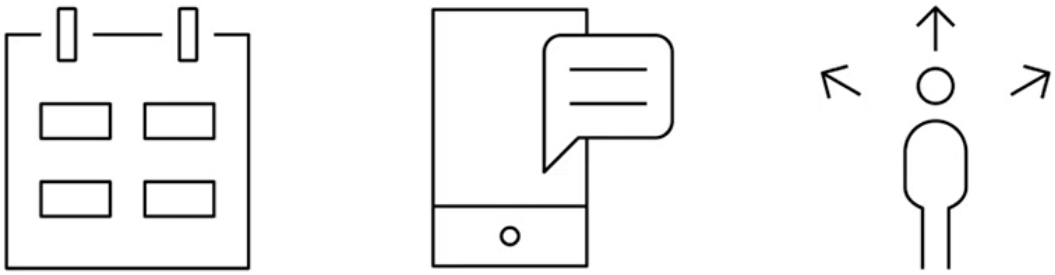
Dashboards can present both operational and analytical data.







Dashboards can present both operational and analytical data.



Dashboards present a bird's eye view of the complete picture while also allowing you to drill down into the next level of information for each parameter.

Dashboards:

- are easy to comprehend by an average user
- make collaboration easy between teams
- allow you to generate reports on the go

Introduction to Visualization and Dashboarding

In this video, we will look at some of the most commonly used data visualization software and tools. These include: Spreadsheets, Jupyter Notebook and Python libraries, R-Studio and R-Shiny, IBM Cognos Analytics, Tableau and Microsoft Power BI.

Some of these are end-to-end data analytics solutions, while others are specifically for data visualization—ranging from free, open-source tools to commercially available solutions.

Spreadsheets, such as Microsoft Excel and Google Sheets, are possibly the most commonly used software to make graphical representations of data sets. Spreadsheets are easy to learn and have a ton of documentation and video tutorials available
online for ready reference

Excel

- provides several chart types ranging from the basic bar, line, pie, and pivot charts, to the more advanced options such as scatter charts, trendlines, Gantt charts, waterfall charts, and combination charts (using which you can combine more than one type of charts).

- Excel also provides recommendations on the best visual representation for your data set. To make the charts more presentable, you can add a chart title, change colors of the elements, and add labels to data.

Google Sheets

- also offers similar chart types for visualization, though Excel does have more inbuilt formula-based options than Google Sheets.
- Like Excel, Google Sheets can help you choose the right visualization. All you have to do is highlight the data you wish to visualize and click the chart button—and you get a list of suggested charts best suited for your data.

- Charts and reports automatically update, in Excel as well as in Google Sheets, as the underlying data is changed.
- Google Sheets is preferred over Excel, where multiple users need to collaborate.

Jupyter Notebook is an open-source web application that provides a great way to explore data and create visualizations. You don't have to be a Python expert to use Jupyter Notebook.

Python provides a host of libraries that are used for data visualization. Let's look at a few of those libraries.

- Matplotlib is a widely used Python data visualization library.
- It provides different kinds of 2D and 3D plots and the flexibility to create plots in several different ways.
- Using Matplotlib, you can create high-quality interactive graphs and plots with just a few lines of code.

- It has large community support and cross-platform support as it is an open-source tool.

- Bokeh provides interactive charts and plots and is known for delivering high-performance interactivity over large or streaming datasets.

- Bokeh offers flexibility for applying interaction, layouts, and different styling options to visualization.
- It can also transform visualizations written in some of the other Python libraries, such as Matplotlib, Seaborn, and Ggplot.

Dash is a Python framework for creating interactive web-based visualizations.

- Using Dash, you can build highly interactive web applications using Python code. While knowledge of HTML and javascript is useful, but it is not a requirement.
- Dash is easily maintainable, cross-platform, and mobile-ready.

R-Studio,

- you can create basic visualizations such as histograms, bar charts, line charts, box plots, and scatter plots; and advanced visualizations such as heat maps, mosaic maps, 3D graphs, and correlograms.

- Shiny is an R package that helps build interactive web apps that you can host as standalone apps on a webpage.

- These web apps seamlessly display R objects, such as plots and tables, and can be made live to allow access to anyone. You can also build dashboards using Shiny.

- The ease of working with Shiny is what popularized it among data professionals.

IBM Cognos Analytics is an end-to-end analytics solution.

- Some of the visualization features provided by Cognos include: Importing custom visualizations;
- A forecasting feature that provides time-series data modeling and forecasts based on data presented in corresponding visualizations;
- Recommendation for visualizations based on your data;
- Conditional formatting which allows you to see the distribution of your data and highlight exceptional data points, for example, highlighting high and low sales numbers over a certain threshold;
- Cognos is known for its superior visualizations and overlaying data on the physical world using its geospatial capabilities.

Tableau

- is a software company that produces interactive data visualization products.
- Using tableau products, you can create interactive graphs and charts in the form of dashboards

and worksheets, with drag and drop gestures.

- Tableau also offers the option to publish results in the form of stories.
- You can import R and Python scripts in Tableau and take advantage of its visualization features that are far more superior to that of other languages.
- Tableau's visualization capabilities are easy and intuitive to use.
- Tableau is compatible with excel files, text files, relational databases, and cloud database sources such as Google Analytics and Amazon Redshift.

Power BI

- is a cloud-based business analytics service from Microsoft that enables you to create reports and dashboards.
- It is a powerful and flexible tool known for its speed and efficiency, and an easy to use drag and drop interface.
- Power BI is compatible with multiple sources, including Excel, SQL Server, and cloud-based data repositories, which makes it an excellent choice for data professionals.
- Power BI provides the ability to collaborate and share customized dashboards and interactive reports securely, even on mobiles.
- Power BI's dashboard consists of many visualizations on a single page that help you tell your story. These visualizations, called tiles, are pinned to the dashboard.

The dashboard is interactive, which means a change in one tile affects the other. When deciding which tools to use, you need to consider the ease-of-use and purpose of the visualization.
In terms of the tools that are available and the visualization capabilities they offer —if you can visualize it, you can create it.

Overview:

Commonly used data visualization software and tools include:

- Spreadsheets
- Jupyter notebook and Python libraries
- R-Studio and R-Shiny
- IBM Cognos Analytics
- Tableau
- Microsoft Power BI

Spreadsheets

- Most commonly used software for graphical representations of data sets
- Easy to learn
- Documentation and video tutorials for ready reference

Excel

- Provides several chart types—bar charts, line charts, pie charts, pivot charts, scatter charts, trendlines, Gantt charts, Waterfall charts, and combination charts
- Provides recommendations on visual representation
- Can add chart title, change colors of elements, and add labels to data

Google Sheets:

- Offer a wide range of charts
- Suggests visualization best suited for your data set
- Preferred over Excel for its collaboration features

Jupyter Notebook and Python Libraries

Jupyter Notebook is an open-source web application that provides a great way to explore data and create visualizations.

Python provides a host of libraries that are used for data visualization.

Python provides a host of libraries that are used for data visualization.

matplotlib **Matplotlib:**

- Widely used Python data visualization library
- Provides different kinds of 2D and 3D plots and the flexibility to create plots in several different ways
- Helps create high-quality interactive graphs and plots with just a few lines of code
- Has a large community support and cross-platform support

bokeh **Bokeh:**

- Provides interactive charts and plots
- Delivers high-performance interactivity over large or streaming datasets
- Offers flexibility for applying interaction, layouts, and different styling options to visualization
- Can transform visualizations written in other Python libraries, such as Matplotlib, Seaborn, and Ggplot

plotly | Dash **Dash:**

- A Python framework for creating interactive web-based visualizations
- Helps build highly interactive web interactive web applications using Python code
- Does not require knowledge of HTML and javascript
- Is easily maintainable, cross-platform, and mobile-ready

R Studio and R - Shiny

Using R-Studio, you can create

- Basic visualizations such as histograms, bar charts, line charts, box plots, and scatter plots
- Advanced visualizations such as heat maps, mosaic maps, 3D graphs, and correlograms

Shiny is an R package that helps build interactive web apps that can be hosted as standalone apps on a webpage.

You can also build dashboards using Shiny.

The ease of working with Shiny is what popularized it among data professionals.

IBM Cognos and Analytics

IBM Cognos Analytics is an end-to-end analytics solution.

Some of the visualization features provided by Cognos include:

- Importing custom visualizations
- A forecasting feature that provides time-series data modeling and forecasts
- Recommendation for visualizations based on your data
- Conditional formatting which allows you to see the distribution of your data and highlight exceptional data points

Cognos is known for its superior visualizations and overlaying data on the physical world using its geospatial capabilities.

Tableau

Tableau is a software company that produces interactive data visualization products.

Tableau products allow you to:

- Create interactive graphs and charts in the form of dashboards and worksheets, with drag and drop gestures
- Publish results in the form of stories
- Import R and Python scripts

Tableau is compatible with:

- Excel files
- Text files
- Relational databases
- Cloud database sources such as Google Analytics and Amazon Redshift

Microsoft Power BI

Power BI is a cloud-based business analytics service from Microsoft that enables you to create reports and dashboards.

- A powerful and flexible tool known for its speed and efficiency
- Has a drag and drop interface
- Is compatible with multiple sources, including Excel, SQL Server, and cloud-based data repositories
- Provides the ability to collaborate and share dashboards and reports securely

Considerations

Considerations for selecting the right tool:

Ease of use

Purpose of the visualization

Reading: Summary and Highlights

In this lesson, you have learned the following information:

Data has value through the stories that it tells. In order to communicate your findings impactfully, you need to:

- Ensure that your audience is able to trust you, understand you, and relate to your findings and insights.
- Establish the credibility of your findings.
- Present the data within a structured narrative.
- Support your communication with strong visualizations so that the message is clear and concise, and drives your audience to take action.

Data visualization is the discipline of communicating information through the use of visual elements such as graphs, charts, and maps. The goal of visualizing data is to make information easy to comprehend, interpret, and retain.

For data visualization to be of value, you need to:

- Think about the key takeaway for your audience.
- Anticipate their information needs and questions, and then plan the visualization that delivers your message clearly and impactfully.

There are several types of graphs and charts available for you to be able to plot any kind of data, such as bar charts, column charts, pie charts, and line charts.

You can also use data visualization to build dashboards. Dashboards organize and display reports and visualizations coming from multiple data sources into a single graphical interface. They are easy to comprehend and allow you to generate reports on the go.

When deciding which tools to use for data visualization, you need to consider the ease-of-use and purpose of the visualization. Some of the popularly used tools include Spreadsheets, Jupyter Notebook, Python libraries, R-Studio and R-Shiny, IBM Cognos Analytics, Tableau, and Power BI.

Quiz: Practice Quiz

Bookmarked

Question 1

1/1 point (ungraded)

Data visualizations such as graphs and charts are a great way to bring data to life.

- (●) True
- ( ) False

✔

Question 2

1/1 point (ungraded)

You can use dashboards to present operational data such as daily progress data, as well as analytical data, such as the overall health of a business function.

- (●) True
- ( ) False

✔

Question 3

1/1 point (ungraded)

What spreadsheet software is preferred when multiple users need to collaborate?

- ( ) Microsoft Excel
- ( ) Tableau
- ( ) R-Studio
- (●) Google Sheets

✔

Quiz: Graded Quiz

Bookmarked

Graded Quiz due Jul 21, 2022 11:44 +08

Question 1

1/1 point (graded)

"A presentation is not a data dump". What is the one thing you would do to ensure your presentation is not a data dump?

- ○ Not use visuals in the presentation
- ○ Not include facts and figures in the presentation
- ○ Deliver the findings in a single slide
- ◉ Include only that information as is needed to address the business problem

✔

Question 2

1/1 point (graded)

What is the discipline of communicating information through the use of visual elements?

- ○ Data profiling
- ○ Data regression
- ○ Data type conversion
- ◉ Data visualization

✔

Question 3

1/1 point (graded)

Matplotlib is a widely used Python data visualization library.

- ◉ True
- ○ False

✔

Question 4

1/1 point (graded)

What is the goal of Data Visualization?

- ○ Establish trust in the audience
- ○ Make collaboration easy
- ○ Make the presentation look attractive
- ◉ Make information easy to comprehend, interpret, and retain

✔

Question 5

1/1 point (graded)

What can you do to help your audience trust you?

- ◉ Share your data sources, hypotheses, and validations
- ○ Hand them copies of the data sets you have used for analysis
- ○ Share the detailed documentation of every aspect of your project so they can verify all details
- ○ Make your presentation look good

✔

Module Introduction

In this module, you will learn about the different career opportunities in the field of Data Analytics and the different paths that you can take for getting skilled as a Data Analyst. Experienced data professionals share how they got into the field of data analytics, how you too can become a Data Analyst, and what are some of the opportunities available to you.

Learning Objectives

After completing this module, you will be able to:

- List the different career opportunities in Data Analytics.
- Describe the various learning paths you can consider for getting skilled as a Data Analyst.
- Describe some of the opportunities available in the field of Data Analytics.

Career Opportunities in Data Analysis

Data analyst job openings exist across industry, government and academia. Every industry, be it banking and finance, insurance, healthcare, retail or information technology has space for skilled data analysts. These roles are a sought after in large businesses as they are in startup San new ventures.

According to Forbes, the global big data analytics market that stood at 37.34 billion US dollars in 2018 is expected to grow at a compound annual growth rate of 12.3% from 2019 to 2027 to reach 105.08 billion US dollars by the year 2027. Currently, the demand for skilled data analysts far outweighs the supply, which means companies are willing to pay a premium to hire skilled data analysts.

There's a wide variety of job roles available for data analysts to understand. The career path is open to you, we will broadly classify the rolls into data analyst, specialist roles and domain specialist roles.

Data analyst specialist roles are for data analysts who want to stay focused and grow in the technical and functional aspects of their role. On this path. You could be starting your career as an **associate or junior data analyst** and work your way up through **analyst, senior analyst, lead analyst and principle analyst roles.**

The boundaries between these roles, the years of experience that qualify you for the next level and the nature of experience you need to gain to move up could vary depending on the industry, the size of the organization, and how big your team is. In smaller teams, for example, you could be gaining experience in all facets of data analysis from gathering data all the way through to visualizing and presenting your findings to stakeholders, and this may happen within a short span of time in larger teams and organizations, roles may typically be bifurcate it based on activity, which means you could be gaining experience in one specific phase of the process before you move to the next. This helps you hone your skills in one part of the process before you move to the next.

On your journey from an associate data analyst to a lead or principle data analyst, you will be continually advancing your technical, statistical and analytical skills from a foundational level to an expert level. You will be demonstrating your ability to work with a wide ranging set of tools and platforms.

Different aspects of the data analysis process and a wide variety of use cases in terms of technical skills, you may start off knowing just one querying tool and programming language. Anyone type of data repository or a limited set of visualization tools. As you gather more experience, you're expected to learn and demonstrate your ability to work with more and more tools, languages, data, repository's and newer technologies, your communication skills, presentation skills, stakeholder management skills and project management skills all need to be honed and taken up A notch progressively.

As **a lead or principle analyst**, you may also be responsible for establishing processes in your team, making recommendations for software and tools. The team should work on upskilling the team and expanding the team to include more profiles.

In some organizations, these responsibilities could be aligned with the manager level person who has risen through the ranks to manage a team of data analysts.

Domain specialists, also known as **functional analysts**, are analysts who require specialization in a specific domain and are seen as an authority in their domain such as our healthcare, sales, finance, social media or digital marketing. They may not be the most technically skilled people. These roles carry titles such as our analyst, marketing analyst, sales analyst, Healthcare analyst or social media analyst.

And then there are the analytics enabled job roles. These include roles such as project managers, marketing managers and HR managers. These are jobs where analytics skills lead to greater efficiency and effectiveness. Affair amount of the data analyst job openings are analytics enabled. As more and more organizations rely on data for decision making.

As a data analyst you also have options for exploring and learning new skills to gain entry into other data professions such as data engineering or data science. For example, if you're starting off as a junior data analyst and really like working with data lakes and big data repository's, you can acquire further expertise in these technologies and evolve your career into becoming a big data engineer.

If the business side of things excite you more, you could similarly explore the skills required for making a lateral move into business analytics or business intelligence Analytics. While the data analyst career landscape is very vast, the good thing is that you have a plethora of resources available to help you grow to be successful in your journey as a data analyst, all you need to do is grab the opportunities you want to pursue, or the ones that present themselves to you and learn along the way.

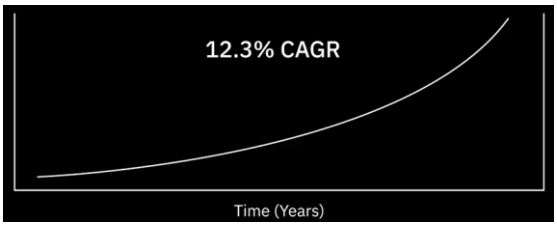
Overview

Data Analyst job openings exist across industry, government and academia

Banking and finance
Insurance
Healthcare
Retail
Information technology

According to Forbes, the global big data analytics market that stood at 37.34 billion US dollars in 2018 is expected to grow at a compound annual growth rate of 12.3% from 2019 to 2027

105.08 billion



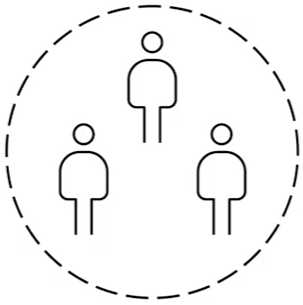
37.34 billion

to reach 105.08 billion US dollars by the year 2027. Currently, the demand for skilled data analysts far outweighs the supply, which means companies are willing to pay a premium to hire skilled data analysts.

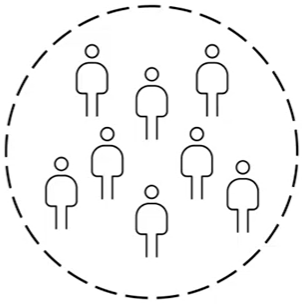
Roles and Responsibilities

Data Analyst Specialist Roles

The boundaries between these roles, the years of experience that qualify you for the next level, and the nature of experience you need to gain to move up, could vary depending on the **industry**, the **size of the** organization,



Gaining experience in all facets of data analysis from data gathering to data visualization and presentation.



Gaining specialization in one area of the data analysis process at a time.

Associate Data Analyst

Advancing your technical, statistical, and analytical skills from foundational to expert level.

- Tools and platforms
- Data analysis process
- Use cases
- Communication skills
- Presentation skills
- Stakeholder management skills
- Project management skills

Principal Analyst

Principal Analyst

As a Lead or Principal Analyst, you may also be responsible for

- Establishing processes in your team
- Making recommendations for software and tools the team should work on
- Upskilling the team
- Expanding the team to include more profiles

Domain Specialist Roles

- Seen as an authority in their domain, such as, HR, Healthcare, Sales, Finance, Social Media, or Digital Marketing
- May or may not be technically skilled
- Carry titles such as **HR Analyst**, **Marketing Analyst**, Sales

Analytics –enabled Job Roles

Project Managers | Marketing Managers | HR Managers

- Analytics-enabled job roles include roles such as **Project Managers**, **Marketing Managers**, and **HR Managers**
- Analytics skills in these job roles lead to greater efficiency and effectiveness

Other Data Professions

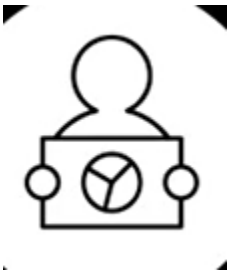
Data Analysts can evolve into other data professions such as

Data Engineers | Data Scientists

For example, if you're starting off as a **Junior Data Analyst** and really like working with data lakes and big data repositories, you can acquire further expertise in these technologies and evolve your career into becoming a **Big Data Engineer**.

Data Analysts can evolve into other data professions such as

Data Analysts Roles



Data analyst Specialist Roles

Domain Specialist Roles

Analytics-enabled Roles

Other Data Professionals

Get into Data Profession

In this video, we will listen to data professionals talk about how they got into this profession. My current role as a data professional did not exist before I took the position. I realized that there was a need in our company to provide data in a faster, more efficient manner, than going to the IS department who would have a meeting to discuss the meeting, to have requirements, and then they would have an end product that people weren't satisfied with. But you had to get at the end of the line and go through the whole process again, to get what you were looking for.

Through filling a need at the company to provide reports in two weeks, I put together a company database that has access to more information. We have analysts that are now able to meet that unmet need in the company. I got into the data professional role by chance. I was actually working on my PhD in Economics at University of Illinois, Urbana-Champaign, when a colleague of mine suggested that a master's in statistics would also be an excellent value add. That's how I got into the statistics program as well in Illinois. But once I started that, I was pretty hooked and there was no going back, so to speak.

In other words, my original goal of becoming an economist actually evolved into a career filled with data, modeling, analytics, insight gathering, communication, visualization, and of course, underlying all of that data-driven problem-solving.

I got into a data analyst role in a financial data company, actually by accident. Back then, my company started to hire equity data analyst in [inaudible] , China, and I was very lucky to join the team, because they were looking for someone who has financial analysis skill sets, which I can bring to the table. After that, my team started to hire someone, with technical skill sets like Python, R, and Sickle. I've always had a love of numbers.

One of the things that happens is when you work with numbers so much, they start to tell a story, and the ability to look at those numbers and tell that story is what speaks to me. Having always had that level of numbers, either just that always attracted to data analytics and whether it's Excel spreadsheets, or whether it is QuickBooks, or any sort of datasets that can help drive the information that we're looking for, especially in the financial industry where we're looking at profit, and loss, and balance sheet, and what happens when one company buys another company.

We're always looking at that data to talk to, and speak about the company's history, and in future. I got my current role as a data scientist straight out of my grad program, which was a Masters in Data Science. Before my grad program, I worked as both a data analyst and an analytics manager.

Get into Data Profession

How did you get into your current role as a Data Professional?

- ✓ My current role as a data professional did not exist before I took the position
- ✓ I identified the need to make data and reports available faster and more efficiently

- ✓ I got into a data profession role by chance
- ✓ While working on my PhD in Economics, I also studied Statistics
- ✓ I loved Statistics so much that my original goal of becoming a economist evolved into a data role

- ✓ I got into the Data Analyst role by accident
- ✓ I was selected to work in the Data Analyst team because of my financial analysis skillset

- ✓ I always had a love of numbers
- ✓ The ability to look at numbers and be able to tell the story is what speaks to me

- ✓ I got my current role as a Data Scientist straight out of my grad program- Masters in Data Science
- ✓ Before my Data Science program, I worked as Data Analyst and an Analytics Manager

Viewpoints: What do Employers look for in a Data Analyst?

In this video, we will listen to data professionals talk about what employers look for in a Data Analyst.
Employers look for Data Analysts with integrity.
During the hiring process, I will ask, if you had to choose just one, would you rather meet a deadline or get a right answer? I'm always looking for someone who would say, I want to make sure that the information is right. Missing a deadline isn't as detrimental as a company making a multi-million dollar decision on wrong information, or someone losing their job because it wasn't pulled or it wasn't reported correctly. It's much more important to have integrity. I think the number one thing employers look for in

Data Analysts is someone who can communicate clearly. If you do the most brilliant analysis in the world, but you can't communicate it to external stakeholders, then it's really not worth anything. I think that skill is really sought after. I think another thing that companies obviously look for when they look for a Data Analyst is fluency with numbers, ability to understand complex analysis, ability to understand AB tests and what the results of AB tests are saying, and the implication of those results. I also think, increasingly, employers are looking for Data Analysts with really strong SQL skills. Another thing employers are looking for in Data Analysts is a growth mindset and willingness to learn, because the industry is changing at a really fast pace.

I think they are looking for the programming skills, including Python, R, SQL. At the same time, they're looking for some personalities. Whether you are detail-oriented, whether you like working with data, and whether you are a problem solver, so on and so forth.

As an employer, I hire people all the time. What am I looking for? We're looking for people who are detail-oriented and who are somewhat overachievers. They don't just want to do what's in front of them, they want to go further. We're looking for people who have higher aspirations, and who also are able to think outside the box. If I say, do ABC, they're not just going to do that, they're going to do it plus [inaudible] and give me some alternatives. People who are able to trouble-shoot. If something goes wrong, they're not just going to stop and say, my goodness, I need to go talk to my supervisor. They're going to say, here's a problem, here's my thoughts. Here are two possible solutions on how you can resolve this so that the job and the company can keep moving forward. That's what you want. Not just detail-oriented and not just good with numbers. You also have to be someone who can think outside the box, and be able to problem solve, and trouble-shoot. That's what employers are going to be looking for now more than ever.

They look for the ability to know data, and by know data we mean several things. Be comfortable with it in various formats, be able to think about it. By that we mean, know what data you want to solve the problems that are at hand. Knowing the data skill is very important. Problem-solving is another very key skill. Meaning, if there is a problem presented to a Data Analyst, they should be able to know how to tackle that problem using data in whatever format it may be sitting in, and being able to analyze it and present the insights that will then solve the problem. They also need to be very dynamic in that, if they are presented with a very different data set suddenly, which looks nothing like it did before, they need to be able to adapt to that change. That's why the quality of being dynamic and adaptable is also important. They also need to be able to pick up technical skills quickly. By that we mean, if there is one SQL DIAdem being used in one setting, they need to be able to operate under a different paradigm. If there is a place that's using RStudio, but they know Python, they need to be able to pick up RStudio quickly, and that thing. Being able to learn fast, being dynamic, and knowing data, those are the few things that employers do look for in a good Data Analyst.

Viewpoints: What do Employers look for in a Data Analyst?

- ✓ Clear communication
- ✓ Fluency with numbers
- ✓ Ability to understand complex analysis
- ✓ Ability to understand AB tests
- ✓ Strong SQL skills
- ✓ A growth mindset

- ✓ Programming skills including Python, R, SQL
- ✓ Detail-orientation
- ✓ Ability to work with data
- ✓ Problem solving skills

- ✓ Detail oriented
- ✓ Over-achievers
- ✓ Think outside the box
- ✓ Trouble-shooters
- ✓ Good with numbers
- ✓ Problem-solvers

- ✓ The ability to know data
- ✓ To know what kind of data will solve the given problem
- ✓ Problem-solving skills
- ✓ Being dynamic skills
- ✓ Being dynamic and adaptable
- ✓ The ability to pick up technical skills quickly

The Many Paths to Data Analysis

There are various paths you can take for gaining entry into the data analyst field. While some employers may ask for an academic degree as a pre-requisite, even if you don't have a degree, you still have several options available to you that can help you gain an entry, or even make a lateral move, into the field of data analysis.

Let's start with the most obvious path.
An academic degree in Data Analytics, Statistics, Computer Science, Management Information Systems, or Information Technology Management can start you off with a strong advantage. You could alternately enroll in online training programs that can equip you with the required
knowledge.

Comprehensive online programs for data analysis are multi-course specializations offered by learning platforms such as Coursera, edX, and Udacity. These courses are designed and delivered by some of the world's best domain experts.

Since you have a fair idea, by now, of the technical, functional, and soft skills you need in order to be a data analyst, choosing the right learning path should be fairly straightforward. As you gather more work experience, you can keep advancing your knowledge and skills in specific areas, for example, Statistics, Spreadsheets, SQL, Python, Data Visualization, Problem-Solving, Storytelling, or making impactful presentations. These courses also give you hands-on assignments and projects which give you a feel for the real-world application of your knowledge and skills. You can even add these projects to your portfolio.
So, if you don't have an academic qualification, these courses can help you gain opportunities
at an entry-level and work your way up as your experience grows.

Now let's look at a scenario where you have a couple of years of experience in a different line of work and want to make a switch into the data analysis field. There's a very good chance that you can do that successfully if you plan well. Since data analysis is a vast field, it would be useful for you to first research the knowledge and skills you need, the various job opportunities that are available, and the growth opportunities available on the path you may be considering.

You can tap into online resources, forums, and your network of friends and colleagues to connect with people in this field and gain insights into real-world scenarios. If you're currently working in a non-technical role, you may consider exploring the Domain Specialist, or Functional Analyst path. If you're in Sales, you could consider starting your journey by positioning and skilling yourself for a Sales Analyst position. You begin with the advantage of industry experience and skill yourself in other areas such as Statistics and programming, for example.

If you're currently working in a technical role, you have the ability to quickly pick up the tools and software you need for the data analyst role. You're also probably stepping in with the advantage of having a good understanding of the domain or industry you're from. For some of the other skills, such as problem-solving, project management, communication, and storytelling—you may already be using these in some capacity in your existing job. You can always enhance these skills through trainings, online courses, communities of practice, and forums. Data Analysis is a fast-moving field. If you're curious, open to learning new things, and

excited about the field, you will be able to forge a path forward, regardless of the formal qualifications you think you may be missing.

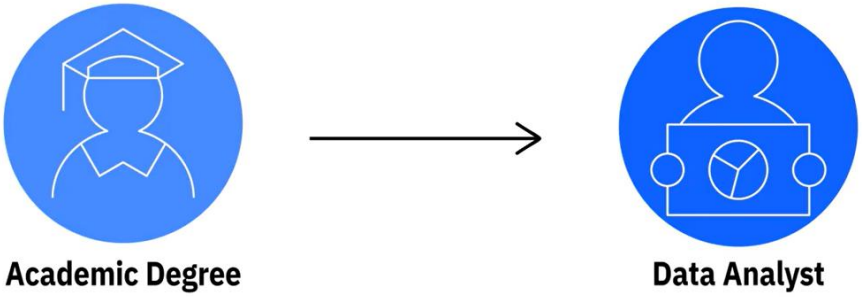
The Many Paths to Data Analysis

Overview

There are various paths you can take for gaining entry into the data analyst field

How to Become a Data Analyst

An academic degree in Data Analytics, Statistics, Computer Science, Management Information Systems, or Information Technology Management can start you off with a strong advantage.



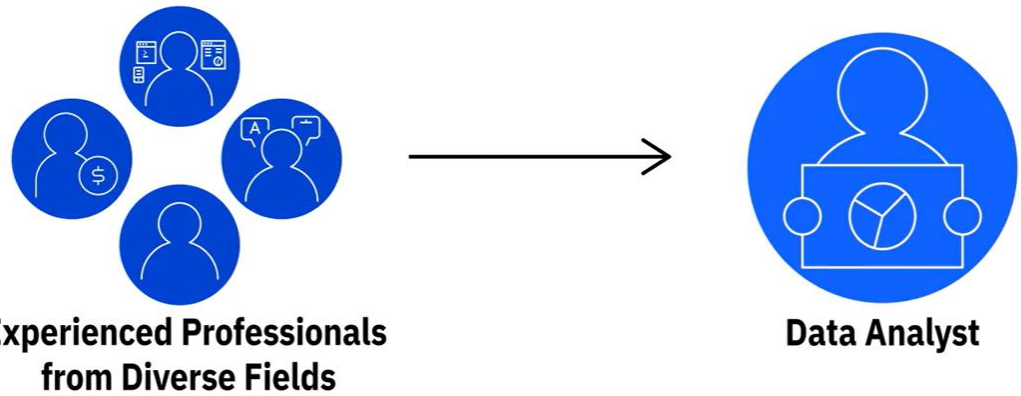
Comprehensive online programs offering multi-course specializations

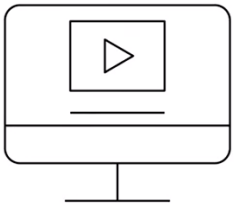
Comprehensive online programs for data analysis are multi-course specializations offered by learning platforms such as Coursera, edX, and Udacity.

Advancing specific skills as you gain work experience, such as, Statistics, Spreadsheets, SQL, Python, Data Visualization, Problem-Solving, Storytelling, or making impactful presentations

Mid-Career Transition to a Data Analyst Role

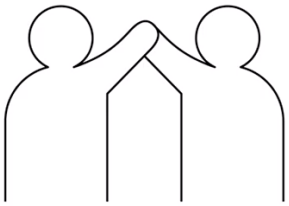
Now let's look at a scenario where you have a couple of years of experience in a different line of work and want to make a switch into the data analysis field.





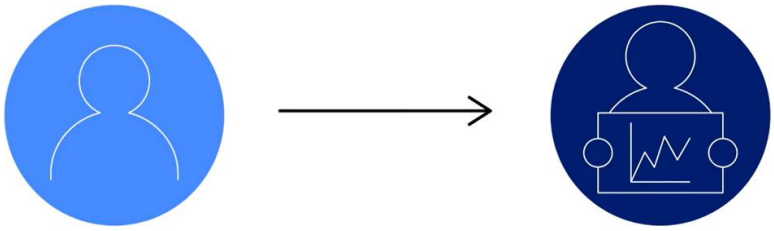
Online Resources

Forums



Network of Friends and Colleagues

There's a very good chance that you can do that successfully if you plan well. Since data analysis is a vast field, it would be useful for you to first research the knowledge and skills you need, the various job opportunities that are available, and the growth opportunities available on the path you may be considering.
You can tap into online resources, forums, and your network of friends and colleagues to connect with people in this field and gain insights into real-world scenarios.

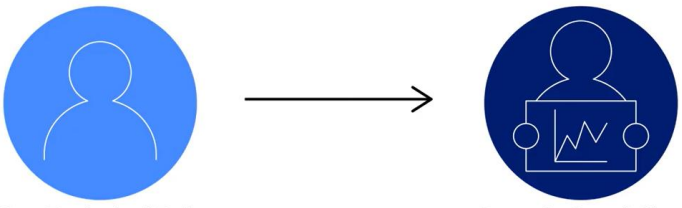


Non-Technical Role

Domain Specialist
Functional Analyst

If you're currently working in a non-technical role, you may consider exploring the Domain Specialist, or Functional Analyst path. If you're in Sales, you could consider starting your journey by positioning and skilling yourself for a Sales Analyst position.

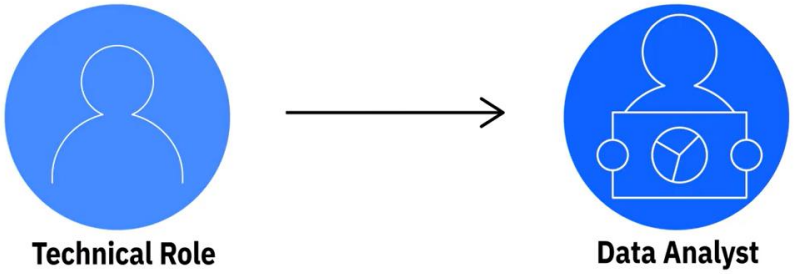
You begin with the advantage of industry experience and skill yourself in other areas such as Statistics and programming.



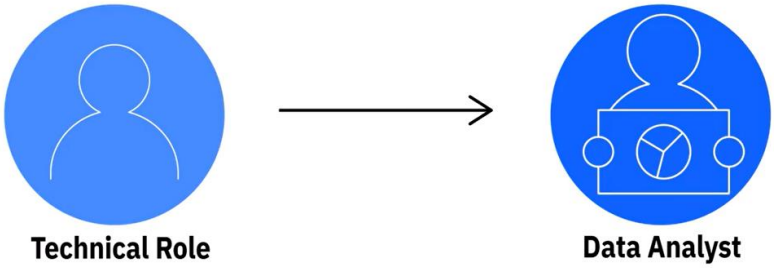
Non-Technical Role

Domain Specialist
Functional Analyst

You begin with the advantage of technical and domain experience and skill yourself in tools and platforms specific to data analysis field.



Problem-Solving **Project Management** **Communication** **Storytelling**



You're also probably stepping in with the advantage of having a good understanding of the domain or industry you're from. For some of the other skills, such as problem-solving, project management, communication, and storytelling—you may already be using these in some capacity in your existing job. You can always enhance these skills through trainings, online courses, communities of practice, and forums.

The Many Paths to Data Analysis

Data Analyst Specialist Roles	Domain Specialist Roles	Analytics-enabled Roles	Other data Professionals

Viewpoints: Career Options for Data Professionals

In this video, we will listen to practicing data professionals talk about the various career options available in this field.

The whole data related profession today has also become very colorful, very dynamic, evolving all the time, and it also presents a lot of range of options to anyone who wants to enter the field of being a data professional. It ranges from, if you were to think of various circles as options, starting with a Data Analyst. From there you can upscale a lot more become a data scientist. You can also become a statistician, which is what I was when I first started off. You can then further specialize yourself in a specific direction of data in order to become a data engineer. Or you can start by being a BI analyst or a specialist and then don't go to become a data engineer. In other words, either you can do a track of Data Analysts and data scientists, or you can do a track of a BI analyst and a data engineer. Those are parallel tracks within the data profession. You can then also go to the other extreme where you can become a Machine Learning Engineer, an AI Engineer and so on. There are many such roles that anyone interested in the field of data can really take on.

A few of the most common career options available to Data Analyst is to get deeper into the weeds with Machine Learning and Engineering, and become a Data Scientist or a Machine Learning Engineer that focus more on Machine Learning modeling. Other career option available to Data Analyst is to dive deeper into the business they're in and to inform top-level company strategy. I think that role is really important and interesting and has really evolved in recent years. Another path for a data analyst is to start to become a people manager and manage other Data Analysts and work to triage what gets worked on. Because there's always going to be more questions in the organization that can be answered with data than there are people to answer them. A Data Manager role can be really interesting and critical in terms of making sure the most important pieces of work actually do get worked on.

You can be a Bookkeeper. You can be an Accountant. You could be a CPA. You can be a Stockbroker or a Financial Analyst for the government or a lot of large companies. You could be a Real Estate Broker. Lots of people are great Data Analysts, but to do that you do have to really like numbers and you have to be really detail oriented. If that's not you and numbers don't jump off the page at you, Data Analyst might not be the right thing for you.

Viewpoints: Career Options for Data Professionals

- ✓ Data professions are dynamic, evolving and present range of options:
    - Data Analyst
    - Data Scientist
    - Statistician
    - Data Engineer
    - BI Analyst
- ✓ Data Analyst can become Machine Learning or AI Engineers

- ✓ Machine Learning Engineer
- ✓ Data Engineer
- ✓ Dive deeper into business and inform company strategy
- ✓ Manage other Data Analysts

- ✓ Bookkeeper
- ✓ Accountant
- ✓ Chartered Professional Accountant
- ✓ Stockbroker
- ✓ Financial Analyst
- ✓ Real Estate Broker

Viewpoints: Advice for aspiring Data Analysts

In this video, we will listen to data professionals giving advice to aspiring data analysts.

One piece of advice I'd give to aspiring data analyst is keep learning and don't get discouraged. There is more that's been written about analytics than you could ever learn in a lifetime. Don't try to learn everything at once but take your time and make sure every week, every month, every year you are constantly learning something new. I think that'll serve you well. One piece of advice I've been given in my career that I found to be really helpful is to consider your career like an uppercase T, and you should have broad knowledge. The top of the T represents that you should have broad knowledge in a number of different areas. Although it doesn't have to be deep, you should know a little bit, at least about A/B testing, about machine learning, bout data visualization, about SQL, about Python, about R. Then the bottom part of the T is you should go really deep on at least one area. There should be one area among the ones I just mentioned, where you have a really deep rigorous understanding of it.

It is, use every job that you have to your advantage meaning something can be found from everything. Whether it is looking at your parents budget or asking your parents if you can see the checkbooks or if you work at a fast food restaurant, looking at the numbers. How many people are coming in? How many dollars are being turned over? Talk to the manager about what's next, what the numbers actually mean. When you're talking to potential employers, have your examples ready. It doesn't have to necessarily be just word experienced but your life experience, how have analytics, how have you used analytics even in your personal life. If you can tell me and talk to me about what you've done, personally or professionally, and how it relates to what we're doing. That will take you a very long way.

Piece of advice I'd give to aspiring data scientists is to build out a professional portfolio that showcases your data science or data analytics skills. You can do this by looking up fun data sets online and analyzing those data sets. You can also do that within your job. Even if your current job isn't to be a data analyst, look for opportunities where you can crunch numbers, and then that'll just naturally lead you to a nice portfolio or nice wins in terms of data analyst projects.

My advice to an aspiring data analyst is to follow your passion. Find a job that meet your needs and gives you joy doing it. There's nothing worse than waking up every morning and hating to go to your place of employment. There are so many data analyst jobs in various industries, departments. There's just so many options that there's no need to take a job, just to have a job. Find something that really fuels your passion and gives you something to get up every morning for.

Viewpoints: What advice would you give to aspiring data analysts?

- ✓ Keep learning and don't get discouraged
- ✓ Don't try to learn everything at once
- ✓ Develop broad knowledge in several areas
- ✓ Develop a deep and rigorous understanding in at least one area

- ✓ Use every job to your advantage
- ✓ Keep your examples ready while talking to potential employers
- ✓ How have you used analytics in your personal life?

- ✓ Build a professional portfolio
- ✓ Use data sets available online
- ✓ Look for opportunities within your existing job

- ✓ Follow your passion
- ✓ Find a job that meets your needs and gives you joy
- ✓ There are several options available – find a job that fuels your passion

Viewpoints: Women in Data Professions

In this video, we will listen to women share their experience of being a data professional,
and their advice to women aspiring to enter this field.

As a woman in Data Science, I still run up against the stereotype that this is a man's job. I've walked into meetings and had people looked disappointed or confused. I take that as an opportunity to prove them wrong. This isn't a job just for men, it's for a person who has the insight, the ability, and the drive to get the job done. As long as you possess those skills, then there's no reason why anyone can't do anything that they put their mind to. Whether you're male or female, whether you are white or black, you have the opportunity to prove people wrong by the work that you produce.

I would say it can be tough, but you have to find your voice and don't be afraid to use it. A lot of times, as women, were not able to find our voice or speak up, or we're afraid of how people will want to treat us if we speak up. But it's more important that you be heard and seen, not just being loud or wrong, but if you have the data to back it up, if you have good content and things you want to say, don't be afraid to raise your hand and let people know that you are a thinker and a you can get this done, because that's going to be important as you progress. The only real way to get ahead is drive, and people don't know you have drive if you're too quiet. If you're just quietly working away in a corner, a lot of times people can't see it. Speak up, make sure your voices may heard, make sure you are being seen as a woman who knows how to grow and how to help in the Data Science field.

When I started, it was mostly men in my class, especially back in grad school. But now, I'm seeing that data teams, both data science and data engineering teams, are filled with a lot of women as well. I would advise women do continue upskilling. If they are fond and if they like a career filled with programming, data and problem-solving, then they should continue building their technical skill set, so that they can represent themselves in the landscape of a data professional as strongly as possible?

Don't allow your gender to be a crutch. Still go hard, put in the work and show the world your amazing talents. There are no roles that are set aside for specific genders. If you're fortunate enough to work in a profession that you thoroughly enjoy, then go for it.

Viewpoints: Women in Data Professions

- ✓ I still run up against the stereotype that this is a man's job
- ✓ You can be a data professional if you have the insight, the ability and the drive to get the job done

- ✓ It can be tough
- ✓ Find your voice
- ✓ Speak up
- ✓ Get things done
- ✓ Be seen

- ✓ Nowadays, you can see a lot of women in both Data Science and Data Engineering teams
- ✓ Women need to continue upskilling and building their technical skills

- ✓ Don't allow your gender to be a crutch – put in the work to show your gender
- ✓ If you thoroughly enjoy the profession – go for it!

Reading: Summary and Highlights

In this lesson, you have learned the following information:
Data Analyst roles are sought after in every industry, be it Banking and Finance, Insurance, Healthcare, Retail, or Information Technology.

Currently, the demand for skilled data analysts far outweighs the supply, which means companies are willing to pay a premium to hire skilled data analysts.
Data Analyst job roles can be broadly classified as follows:

- Data Analyst Specialist roles - On this path, you start as a Junior Data Analyst and move up to the level of a Principal Analyst by continually advancing your technical, statistical, and analytical skills from a foundational level to an expert level.
- Domain Specialist roles - These roles are for you if you have acquired specialization in a specific domain and want to work your way up to be seen as an authority in your domain.
- Analytics-enabled job roles - These roles include jobs where having analytic skills can up-level your performance and differentiate you from your peers.
- Other Data Professions - There are several other roles in a modern data ecosystem, such as Data Engineer, Big Data Engineer, Data Scientist, Business Analyst, or Business Intelligence Analyst. If you upskill yourself based on the required skills, you can transition into these roles.

There are several paths you can consider in order to gain entry into the Data Analyst field. These include:

- An academic degree in Data Analytics or disciplines such as Statistics and Computer Science.
- Online multi-course specializations offered by learning platforms such as Coursera, edX, and Udacity.
- Mid-career transition into Data Analysis by upskilling yourself. If you have a technical background, for example, you can focus on developing the technical skills specific to Data Analysis. If you do not have a technical background, you can plan to skill your self in some basic technologies and then work your way up from an entry-level position.

Quiz: Practice Quiz

Bookmarked

Question 1

1/1 point (ungraded)

On the Data Analyst Specialist path, you could be starting your career as an Associate or Junior Data Analyst and working your way up to a Principal Analyst role. What are some of the factors that influence your growth on this path?

- ○ Specialization in at least one domain area
- ○ At least five to six years of experience at each level
- ◉ The experience and exposure you gain in the different areas within Data Analysis
- ○ A Master's degree in either Mathematics or Statistics

✔

Question 2

0/1 point (ungraded)

Skills such as problem-solving, communication, and storytelling are critical to the role of a Data Analyst. And like most soft skills, you're either good at them, or you're not; these skills cannot be acquired over time.

- ◉ True
- ○ False

✖

Quiz: Graded Quiz

Bookmarked

Graded Quiz due Jul 23, 2022 19:44 +08

Question 1

1/1 point (graded)

Which of the following statement describes Data Analyst Specialist Roles?

- ○ Analysts who specialize in specific fields like HR, Sales, and Finance
- ○ Analysts who can work with Machine and Deep Learning models
- ◉ Analysts who advance technical, statistical, and analytical skills, over time, to expert levels
- ○ Analysts who specialize in data lakes and data repositories

✔

Question 2

1/1 point (graded)

A Principal Data Analyst is responsible for:

- ○ Being a domain specialist
- ○ Being well-versed in Big Data processing tools
- ○ Having expertise in all tools and technologies used in data analytics
- ◉ Establishing processes in the team

✔

Question 3

1/1 point (graded)

Job roles such as Project Managers, Marketing Managers, and HR Managers, can achieve greater efficiency and effectiveness in their current roles by acquiring data analysis skills, and are therefore known as analytics-enabled job roles.

- ◉ True
- ○ False

✔

Question 4

1/1 point (graded)

Which of these is essential for getting started and growing as a Data Analyst?

- ○ Domain specialization
- ○ A degree in Computer Science
- ○ A degree in Statistics
- ◉ Love for numbers, a curious mind, and openness to learn

✔

Question 5

1/1 point (graded)

What Data Analysis role may be best suited for people with little or no technical training?

- ○ Data Scientist
- ◉ Functional Analyst
- ○ Big Data Engineer
- ○ Data Analyst

✔

Using Data Analytics for Forecasting and Planning Inventory

One of the retail industry's operational challenges is maintaining an optimal supply of products while reducing unused inventory. More and more retailers today employ analytical techniques to track sales and inventory data and make projections based on historical data.

Imagine you are a Data Analyst in the planning division of a leading retailer in the USA. Your assignment is to forecast the demand for three key product categories for the current year so that you can be adequately stocked to meet the year-round demand.

Before you can forecast the inventory level for these product categories, you need to look at the sales history for these product categories. You need to analyze trends and patterns hidden in historical data and understand some of the influences that drive these trends.

Here is a sample dataset that captures the sales data for these three product categories over the years 2018 and 2019. The dataset also captures whether the retailer conducted targeted marketing campaigns during a specific quarter and the periods during which the product category had discounts. In a real-life scenario, the dataset would capture many more details, but we are presenting a more simplified dataset for our purposes.

Month-Year of Purchase	Product Category	Units Sold	Discount	Targeted Campaign	Customer Ratings	Defects Reported
Jan-Mar 2018	Designer Clothes	1275	0%	N	9.6	12
Jan-Mar 2018	Fitness Gadgets	4250	15%	N	9.5	12
Jan-Mar 2018	Travel Accessories	1670	0%	N	9.3	8
Apr-Jun 2018	Designer Clothes	1825	0%	N	8.9	23
Apr-Jun 2018	Fitness Gadgets	3760	0%	N	7.7	32
Apr-Jun 2018	Travel Accessories	1720	0%	N	9.1	7
Jul-Sep 2018	Designer Clothes	3150	10%	Y	9.3	15
Jul-Sep 2018	Fitness Gadgets	1330	0%	N	8.5	8
Jul-Sep 2018	Travel Accessories	3550	0%	Y	9.1	12
Oct-Dec 2018	Designer Clothes	4715	20%	Y	7.1	48
Oct-Dec 2018	Fitness Gadgets	6450	20%	Y	8.7	22
Oct-Dec 2018	Fitness Gadgets	6450	20%	Y	8.7	22
Oct-Dec 2018	Travel Accessories	5430	20%	N	9.2	9
Jan-Mar 2019	Designer Clothes	1375	0%	N	8.6	6
Jan-Mar 2019	Fitness Gadgets	1765	0%	N	8.9	5
Jan-Mar 2019	Travel Accessories	1475	0%	N	7.9	23
Apr-Jun 2019	Designer Clothes	2175	0%	Y	8.8	8
Apr-Jun 2019	Fitness Gadgets	1925	0%	N	8.6	8
Apr-Jun 2019	Travel Accessories	1215	0%	N	8.2	6
Apr-Jun 2018	Flat Screen	4750	0%	Y	7.8	34
Jul-Sep 2019	Designer Clothes	3100	15%	N	9.2	14
Jul-Sep 2019	Fitness Gadgets	2530	0%	N	8.7	12
Jul-Sep 2019	Travel Accessories	3275	0%	Y	8.7	13
Oct-Dec 2019	Designer Clothes	6425	25%	Y	9.3	12
Oct-Dec 2019	Fitness Gadgets	7125	20%	Y	9.6	16
Oct-Dec 2019	Travel Accessories	6510	30%	Y	9.3	8

Descriptive techniques of analysis, that is, techniques that help you understand what happened, include identifying trends, patterns, and correlations in data. Some of the common events that you may need to watch out for in this dataset include:

- Quarterly sale numbers of a product category over multiple years.
- Demand fluctuations over holidays.

New Year's Eve (Jan), Valentine's Day (Feb), Mother's Day (May), Father's Day (June), Independence Day (July), Thanksgiving (Nov), and Christmas (Dec).

- Change in product sales numbers in the weeks following a targeted marketing campaign.
- Change in product sales numbers when a product is available at a discounted price.
- A correlation between increase or decrease of sales of one product leading to a corresponding increase or decrease in sales of another product.
- A correlation between customer ratings accrued for a product in a quarter and its impact on sales numbers in the next quarter.
- A correlation between complaints received for defective products in a quarter and its impact on sales numbers in the subsequent quarter.

Before you can analyze the data for patterns and anomalies, you need to:

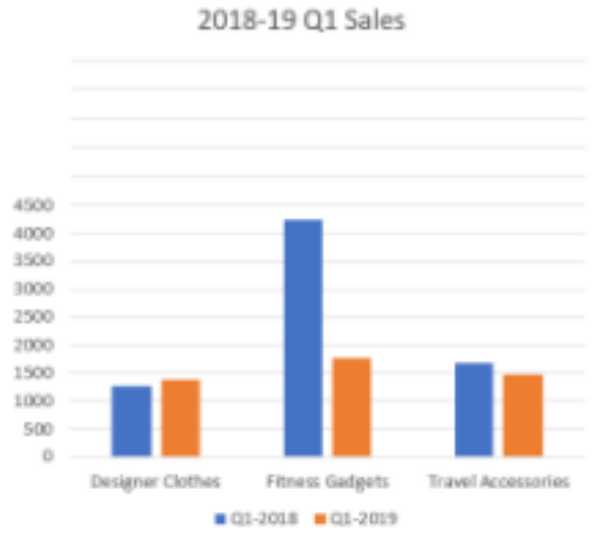
- Identify and gather all data points that can be of relevance to your use case.

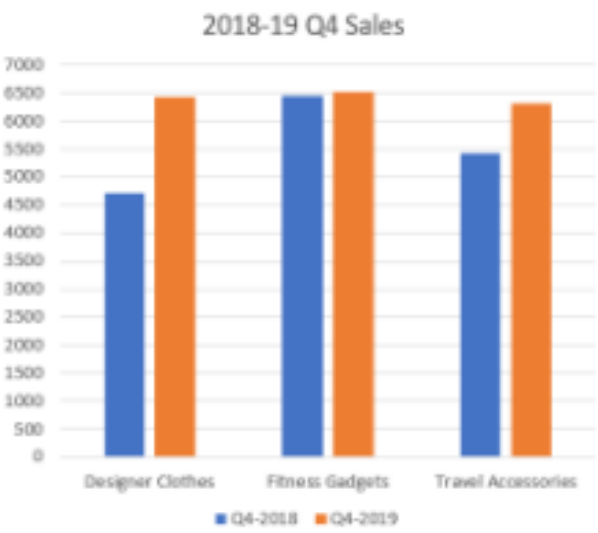
For example, product category, month and year of sale, product discounts, targeted marketing campaigns, product ratings, and defects reported.

- Clean the data.

You need to identify and fix issues in the data that can lead to false or incomplete findings, such as missing data, redundant data, and incorrect data.

Finally, when you arrive at the findings, you create appropriate visualizations that communicate your findings to your audience. The graphs below show visualizations you may use to capture hidden trends in data.





OPEN RESPONSE ASSESSMENT

This assignment has several steps. In the first step, you'll provide a response to the prompt. The other steps appear below the **Your Response** field.

IN PROGRESS

1 | **Your Response** due Dec 31, 2099 08:00 SGT (in 77 years, 5 months)

Enter your response to the prompt. You can save your progress and return to complete your response at any time before the due date (Thursday, Dec 31, 2099 08:00 SGT). **After you submit your response, you cannot edit it.**

The prompt for this section

> Task 1: Introduce yourself.
>
> (1 Point)

Your Response (Optional)

My name is Em. I am a senior lecturer in one of the universities here in the Philippines. I'm thinking of treading a new career path as a data professional.

The prompt for this section

> Task 2: Briefly explain why you want to learn about Data Analytics.
>
> (1 Point)

Your Response (Optional)

I want to learn about data analytics so I can clearly understand, specifically, the difference and the similarities between data analytics and data science. Same for the the functions, roles and responsibilities of being a data analyst and a data scientist.

> Task 3: List at least 2 data points that are required for analyzing trends and patterns that will help you forecast product inventory for the current year. (2 points)

Your Response (Optional)

Product Category and Units Sold

> Task 4:Refer to the data table in the reading and identify 2 errors/issues that could impact the accuracy of your findings. (2 Points)

Your Response (Optional)

1.The row, April-June 2018 with product category of Flat Screen and so on is under 2019.
2. This is a different product category - an appliance - perhaps TV, and not for a personal use category.

> Task 5: Refer to the data table in the reading and identify 2 correlations that you notice in the dataset. (2 Points)

Your Response (Optional)

Customer ratings and Defects reported - Negative correlation. The lowest customer rating is 7.1 with 48 as the no. of defects reported. The highest customer rating is 9.6 with 16 as the no. of defects reported.

Task 6: Briefly explain 2 observations or insights that you can take away from the provided data visualization charts. (2 Point)

Your Response (Optional)

The Targeted Campaign and the Discount categories can be taken away for now . These 2 are the least that can be used to forecast the inventory level for the three product categories.

Task 7: "Data has value through the stories that it tells. Your audience must be able to trust you, understand you, and relate to your findings and insights." Refer to Module 8 video, "Overview of Communicating and Sharing Data Analysis Findings" and list two ways in which you can help your audience trust, understand, and relate to you

(2 Points)

Your Response (Optional)

For your audience to trust, understand and relate to you, they must see the credibility of your data by sharing your data sources, hypotheses and validations.