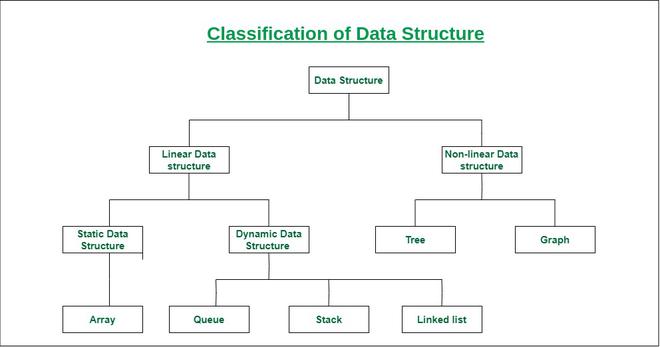
**UNIT 1**

**What is Data Structure?**

*A data structure is a storage that is used to store and organize data. It is a way of arranging data on a computer so that it can be accessed and updated efficiently.*

A data structure is not only used for organizing the data. It is also used for processing, retrieving, and storing data. There are different basic and advanced types of data structures that are used in almost every program or software system that has been developed. So we must have good knowledge about data structures.

**Classification of Data Structure:**



***Classification of Data Structure***

* **Linear data structure:** Data structure in which data elements are arranged sequentially or linearly, where each element is attached to its previous and next adjacent elements, is called a linear data structure.   
  *Examples of linear data structures are array, stack, queue, linked list, etc.*
  + **Static data structure:**Static data structure has a fixed memory size. It is easier to access the elements in a static data structure.   
    *An example of this data structure is an array.*
  + **Dynamic data structure:**In dynamic data structure, the size is not fixed. It can be randomly updated during the runtime which may be considered efficient concerning the memory (space) complexity of the code.   
    *Examples of this data structure are queue, stack, etc.*
* **Non-linear data structure:**Data structures where data elements are not placed sequentially or linearly are called non-linear data structures. In a non-linear data structure, we can’t traverse all the elements in a single run only.   
  *Examples of non-linear data structures are trees and graphs.*

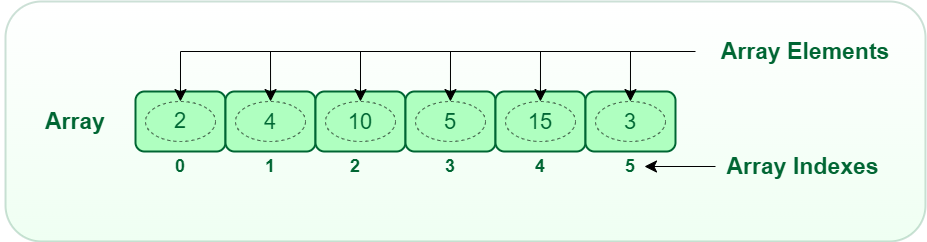
**For example,** we can store a list of items having the same data-type using the *array* data structure.



Array Data Structure

## What is Array?

*An array is a collection of items stored at contiguous memory locations. The idea is to store multiple items of the same type together. This makes it easier to calculate the position of each element by simply adding an offset to a base value, i.e., the memory location of the first element of the array (generally denoted by the name of the array).*

[](https://media.geeksforgeeks.org/wp-content/uploads/20220721080308/array.png)

*Array Data Structure*

The above image can be looked as a top-level view of a staircase where you are at the base of the staircase. Each element can be uniquely identified by their index in the array

**Applications of Array Data Structure:**

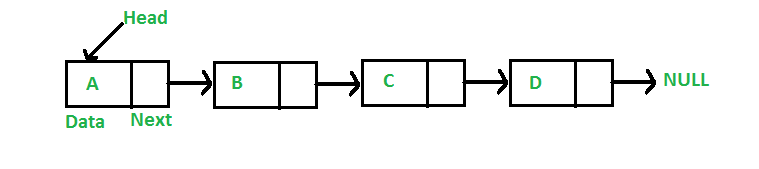
Below are some applications of arrays.

* **Storing and accessing data**: Arrays are used to store and retrieve data in a specific order. For example, an array can be used to store the scores of a group of students, or the temperatures recorded by a weather station.
* **Sorting:** Arrays can be used to sort data in ascending or descending order. Sorting algorithms such as bubble sort, merge sort, and quicksort rely heavily on arrays.
* **Searching**: Arrays can be searched for specific elements using algorithms such as linear search and binary search.
* **Matrices**: Arrays are used to represent matrices in mathematical computations such as matrix multiplication, linear algebra, and image processing.
* **Stacks and queues:** Arrays are used as the underlying data structure for implementing stacks and queues, which are commonly used in algorithms and data structures.
* **Graphs**: Arrays can be used to represent graphs in computer science. Each element in the array represents a node in the graph, and the relationships between the nodes are represented by the values stored in the array.
* **Dynamic programming**: Dynamic programming algorithms often use arrays to store intermediate results of sub problems in order to solve a larger problem.

Linked List Data Structure

## What is Linked List

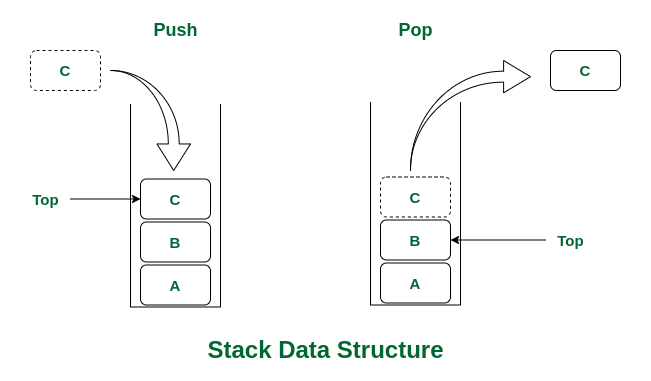
*A linked list is a linear data structure, in which the elements are not stored at contiguous memory locations. The elements in a linked list are linked using pointers as shown in the below image:*



In simple words, a linked list consists of nodes where each node contains a data field and a reference (link) to the next node in the list.

Stack Data Structure

Stack is a linear data structure that follows a particular order in which the operations are performed. The order may be LIFO(Last In First Out) or FILO(First In Last Out). LIFO implies that the element that is inserted last, comes out first and FILO implies that the element that is inserted first, comes out last.



*Stack Data Structure*

There are many real-life examples of a stack. Consider an example of plates stacked over one another in the canteen. The plate which is at the top is the first one to be removed, i.e. the plate which has been placed at the bottommost position remains in the stack for the longest period of time. So, it can be simply seen to follow LIFO (Last in First Out)/FILO (First

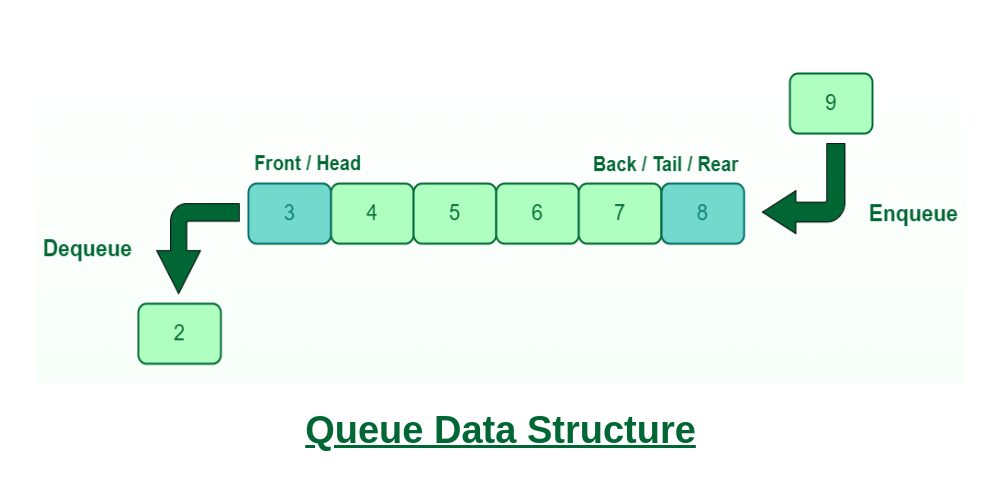
in Last Out) order.

Queue Data Structure

**What is Queue Data Structure?**

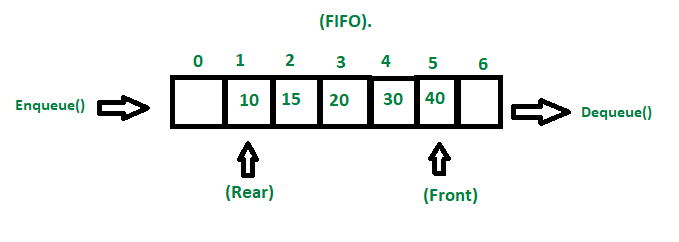
*A****Queue****is defined as a linear data structure that is open at both ends and the operations are performed in First In First Out (FIFO) order.*

We define a queue to be a list in which all additions to the list are made at one end, and all deletions from the list are made at the other end.  The element which is first pushed into the order, the operation is first performed on that.

[](https://media.geeksforgeeks.org/wp-content/cdn-uploads/20221213113312/Queue-Data-Structures.png)

**FIFO Principle of Queue:**

* A Queue is like a line waiting to purchase tickets, where the first person in line is the first person served. (i.e. First come first serve).
* Position of the entry in a queue ready to be served, that is, the first entry that will be removed from the queue, is called the **front** of the queue(sometimes, **head** of the queue), similarly, the position of the last entry in the queue, that is, the one most recently added, is called the **rear** (or the **tail**) of the queue. See the below figure.

[](https://media.geeksforgeeks.org/wp-content/cdn-uploads/20221213111946/fifo-property-in-Queue.png)

*Fifo Property in Queue*

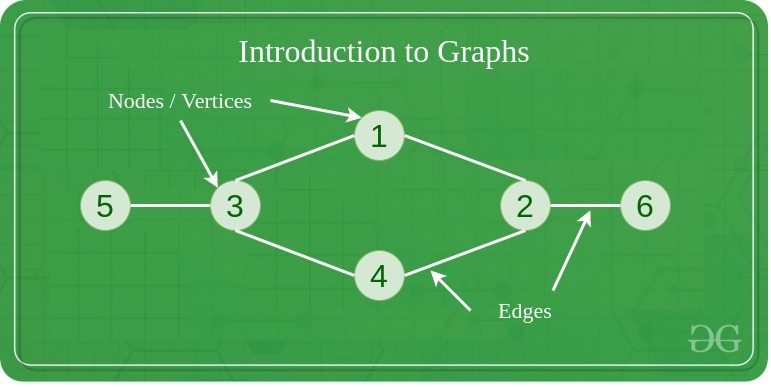
**Characteristics of Queue:**

* Queue can handle multiple data.
* We can access both ends.
* They are fast and flexible.
* **Queue:** the name of the array storing queue elements.
* **Front**: the index where the first element is stored in the array representing the queue.
* **Rear:** the index where the last element is stored in an array representing the queue.

Graph Data Structure And Algorithms

## [What is Graph Data Structure?](https://www.geeksforgeeks.org/introduction-to-graphs-data-structure-and-algorithm-tutorials/)

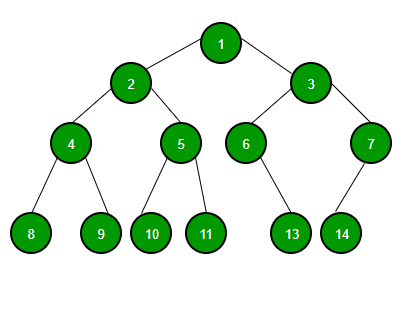
A Graph is a non-linear data structure consisting of vertices and edges. The vertices are sometimes also referred to as nodes and the edges are lines or arcs that connect any two nodes in the graph. More formally a Graph is composed of a set of vertices( V ) and a set of edges( E ). The graph is denoted by G(E, V).



Graphs are used to solve many real-life problems. Graphs are used to represent networks. The networks may include paths in a city or telephone network or circuit network. Graphs are also used in social networks like linkedIn, Facebook. For example, in Facebook, each person is represented with a vertex(or node). Each node is a structure and contains information like person id, name, gender, locale etc.

Binary Tree Data Structure

***Binary Tree is defined as a tree data structure where each node has at most 2 children. Since each element in a binary tree can have only 2 children, we typically name them the left and right child.***

****

**Binary Tree Representation**

**A Binary tree is represented by a pointer to the topmost node (commonly known as the “root”) of the tree. If the tree is empty, then the value of the root is NULL. Each node of a Binary Tree contains the following parts:**

1. **Data**
2. **Pointer to left child**
3. **Pointer to right child**

**Basic Operation On Binary Tree:**

* **Inserting an element.**
* **Removing an element.**
* **Searching for an element.**
* **Traversing the tree.**

**Auxiliary Operation On Binary Tree:**

* **Finding the height of the tree**
* **Find the level of a node of the tree**
* **Finding the size of the entire tree.**

## What is a Data Repository?



A data repository, often called a data archive or library, is a generic terminology that refers to a segmented data set used for reporting or analysis.

It’s a vast [database](https://en.wikipedia.org/wiki/Database) infrastructure that gathers, manages, and stores varying data sets for analysis, distribution, and reporting.

## Types of Data Repositories

Some common types of data repositories include:

### Data Warehouse

A [data warehouse](https://www.astera.com/type/blog/data-warehouse-strategy/) is a large central data repository that gathers data from several sources or business segments. The stored data is generally used for [reporting and analysis](https://www.astera.com/type/blog/data-warehouse-and-business-intelligence/) to help users make critical business decisions.

In a broader perspective, a data warehouse offers a consolidated view of either a physical or logical data repository gathered from numerous systems. The main objective of a data warehouse is to establish a connection between data from current systems, such as product catalogue data stored in one system and procurement orders for a client stored in another one.

### Data Lake

A data lake is a unified data repository that allows you to store structured, semi-structured, and [unstructured](https://www.astera.com/type/blog/automated-data-extraction-tools-for-faster-insights/) enterprise data at any scale. Data can be in raw form and used for different tasks like reporting, visualizations, advanced analytics, and machine learning.

### Data Mart

A data mart is a subject-oriented data repository often a segregated section of a [data warehouse](https://www.astera.com/solutions/technology-solutions/data-warehousing/). It holds a subset of data usually aligned with a specific business department, such as marketing, finance, or support.

Due to its smaller size, a data mart can fast-track business procedures as you can easily access relevant data within days instead of months. As it only includes the data pertinent to a specific area, a data mart is an economical way to acquire actionable insights swiftly.

### Metadata Repositories

While metadata incorporates information about the structures that include the actual data, metadata repositories contain information about the data model that store and share this data. They describe where the data source is, how it was collected, and what it signifies. It may define the arrangement of any data or subject deposited in any format.

For businesses, metadata repositories are essential in helping people understand administrative changes, as they contain detailed information about the data.

### Data Cubes

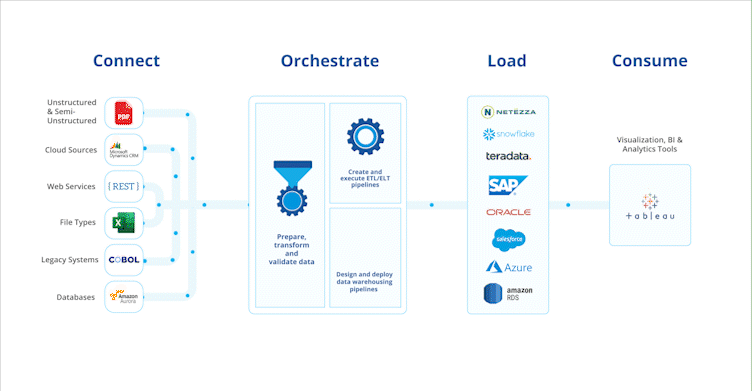
Data cubes are data lists with multidimensions (usually three or more dimensions) stored as a table. They are used to describe the time sequence of an image’s data and help assess gathered data from a range of standpoints.

Each dimension of a data cube signifies specific database characteristics such as day-to-day, monthly or annual sales. The data within a data cube allows you to analyze all the information for almost any client, sales representative, products, and more. Consequently, a data cube can help you identify trends and scrutinize business performance.

## Why Do You Need A Data Repository?

A data repository can help businesses fast-track decision-making by offering a consolidated space to store data critical to your operations. This segmentation enables easier data access and troubleshooting and streamlines reporting and analysis.

For instance, if you want to find out which of your workplaces incur the most cost, you can create an information repository for leases, energy expenses, amenities, security, and utilities, excluding employees or business function information. Storing this data in one place can make it easier for you to come to a decision.



## Challenges Associated with a Data Repository

Although an information repository offers many benefits, it also includes several challenges that you must manage efficiently to alleviate possible data security risks.

Some challenges of maintaining data repositories include:

* An increase in data sets can reduce your system’s speed. To rectify this problem, ensure that the database management system can scale with data expansion.
* In case a system crashes, it can negatively impact your data. It’s best to maintain a backup of all the [databases](https://www.astera.com/type/blog/a-quick-overview-of-different-types-of-databases/) and restrict access to control the system risk.
* Unauthorized operators can [access sensitive data](https://www.astera.com/type/blog/security-and-access-control-in-centerprise-8-0/) more quickly if stored in a single location than if it’s dispersed across numerous sources. On the contrary, implementing security protocols on a single data storage location is more accessible than multiple ones.

## +Best Practices to Create and Manage Data Repositories

When creating and maintaining software repositories, you have to make several hardware and software decisions. Therefore, it is best to involve all stakeholders during the development and usage phase of the data repositories. For example, in case of building a clinical data repository architecture, it is a good idea to involve doctors, data experts, analysts and data pipeline engineers in the initial planning stages.

Here are some of the best practices to help you make the most of this storage solution:

### 1. Select the Right Tool

Using [ETL](https://www.astera.com/solutions/technology-solutions/etl/) tools to create a data repository and transfer data can help ensure data quality is maintained during the process. But keep in mind that different data repository tools offer additional features to create, maintain, and control the repository. So, find a tool that provides the features that support your business requirements.

### 2. Limit the Scope Initially

It’s best to narrow down the scope of your information repository in the initial days. Accumulate smaller data sets and limit the number of subject areas. Gradually increase the complexity as the data operators get familiar with the system.

### 3. Automate as Much as Possible

Automating the process for loading and maintaining the data repository saves the user from manual efforts and reduces the chances of errors.

### 4. Prioritize Flexibility

The data repository should be scalable enough to accommodate evolving data types and increase volumes. So, make flexible plans that make allowance for alterations in technology.

## What is Business Intelligence vs. Data Science?

Business intelligence (BI) and data science are both data-focused processes, but there are some key differences between the two. In general, business intelligence focuses on analyzing past events, while data science aims to predict future trends. Data science requires a more [technical skill set](https://corporatefinanceinstitute.com/resources/career/technical-skills/) compared to business intelligence.

### **Key Highlights**

* **Business intelligence converts data into information that can support business leaders in decision-making.**
* **Data science involves creating forecasts by analyzing the patterns behind the raw data.**
* **Business intelligence is backward-looking that discovers the previous and current trends, while data science is forward-looking and forecasts future trends.**
* **Compared to business intelligence, data science is able to manage more dynamic and less organized data. Yet, it also requires more technical skills and resources.**

## What is Business Intelligence?

Business intelligence is based on the concept of using data to drive actions. It aims to provide business leaders with actionable insights through data processing and analysis. For example, a business analyzes its[KPIs (key performance indicators)](https://corporatefinanceinstitute.com/resources/management/key-performance-indicators-kpis/) to identify its strengths and weaknesses. Thus, the management team can decide in which area the company can improve its operating efficiency.

It is not a new practice to support decision-making with data. However, dramatic improvements in BI technology also mean significant improvements in speed, efficiency, and effectiveness. Automation and data visualization are two examples, which both are transforming the process of business intelligence.

**What is Data Science?**

Data science involves extracting information from datasets and creating forecasts. It uses [machine learning](https://corporatefinanceinstitute.com/resources/data-science/machine-learning-in-finance/), descriptive analytics, and other sophisticated analytics tools. The process of data science starts from collecting and maintaining data. The second step is to process data through data mining, modelling, and summarization.

The next step is data analysis, which can be conducted through text mining, regression, descriptive and predictive analytics, and so on. By analysing the data, the patterns behind the raw data can be discovered to forecast future trends.

Data science is broadly used in many industries. Businesses can use such an approach to develop new products, study customer preferences, and predict market trends. For example, auto-driving developers collect extensive amounts of data for statistical analysis. The developers work to improve the auto-driving system so that it can be responsive to different situations through machine learning.

Data science is also an essential tool in the healthcare industry. High volumes of data can be collected from electronic medical records and individuals’ fitness trackers. Professionals can better understand diseases and develop more effective treatments by applying data science tools to the collected data.

**How is Business Intelligence Different from Data Science?**

Both business intelligence and data science turn data into information that supports business decision-making. However, there are nuances between the two approaches.

**How is Business Intelligence Different from Data Science?**

Both business intelligence and data science turn data into information that supports business decision-making. However, there are nuances between the two approaches.

|  | **Business Intelligence** | **Data Science** |
| --- | --- | --- |
| **Objectives** | Focuses on identifying historical trends; answers questions such as what happened during the last period and what trends are developing | Extracts information from datasets and creating forecasts; answers the question of what will happen or which is the most likely outcome |
| **Skills requirements** | Basic statistics and business knowledge, as well as data transformation and visualization skills | More technical skillset like coding, data mining, as well as more advanced statistics and domain knowledge |
| **Data collection and management** | Designed to manage well-organized data | Designed to manage a large volume of dynamic and less structured data |
| **Complexity** | More practical in daily business management; less costly and requires fewer resources | More complex in terms of capacity for forecasting, ability to manage dynamic data, and requirements for more advanced skills |

# Current Analytical Architecture

Analytics architecture refers to the systems, protocols, and technology used to collect, store, and analyze data. The concept is an umbrella term for a variety of technical layers that allow organizations to more effectively collect, organize, and parse the multiple data streams they utilize.

When building analytics architecture, organizations need to consider both the hardware — how data will be physically stored — as well as the software that will be used to manage and process it.

Analytics architecture also focuses on multiple layers, starting with data warehouse architecture, which defines how users in an organization can access and interact with data. Storage is a key aspect of creating a reliable analytics process, as it will establish both how your data is organized, who can access it, and how quickly it can be referenced.

Structures like data marts, data lakes, and more standard warehouses are all popular foundations for modern analytics architecture. On the user side, creating easier processes for access means including tools like natural language processing and ad-hoc analytics capabilities to reduce the need for specialized workers and wasted resources. When seen as a whole, analytics architecture is a key aspect of business intelligence.

Drivers of BIG DATA

## ****Volume****

Volume refers to how much data is actually collected. An analyst must determine what data and how much of it needs to be collected for a given purpose. To imagine the possibilities, consider a social media site where people write updates, like photos, review business, watch videos, search for new items and interact in some way with just about everything they see on their screens. Each of these interactions generates data about that person that can be fed into algorithms.

## ****Veracity****

Veracity relates to how reliable data is. An analyst wants to ensure that the data they look at is valid and comes from a trusted source. This is determined by where the data comes from and how it is collected. Data collected from native sites rather than third-parties is necessary for reliable results. Additionally, testing measures must be properly designed to ensure that data results in the desired information and is not extraneous.

## ****Velocity****

Velocity in big data refers to how fast data can be generated, gathered and analyzed. Big data does not always have to be used imminently, but in some fields, there is a great advantage to receiving up to the second information about rates and being able to act accordingly. In other businesses, the data trend over time is more important to help make predictions or solve lingering problems.

## ****Variety****

Variety refers to how many points of reference are used to collect data. If data is collected from a single source, that information may be skewed in some ways. It will not represent a broad population or wide trend. In some cases, like with velocity, that is fine. A pet microchipping service, for example, may only want to target data from a neighborhood social networking site. A movie company, on the other hand, may want to target several social media sites and people of various age groups. So they would need more points of reference to decide on the best places to do business.

### Emerging Big Data Ecosystem and a New Approach to Analytics

Evolve, the market sees the introduction of data vendors and data cleaners that use crowdsourcing (such as Mechanical Turk and GalaxyZoo) to test the outcomes of machine learning techniques. Other vendors oﬀer added value by repackaging open source tools in a simpler way and bringing the tools to market. Vendors such as Cloudera, Hortonworks, and Pivotal have provided this value-add for the open source framework Hadoop.

As the new ecosystem takes shape, there are four main groups of players within this interconnected web. These are shown in Figure 1-11.

●**Data devices** the “Sensornet” gather data from multiple locations and continuously generate new data about this data. For each gigabyte of new data

created, an additional petabyte of data is created about that data.

●For example, consider someone playing an online video game through a PC, game console, or smartphone. In this case, the video game provider captures data about the skill and levels attained by the player. Intelligent systems monitor and log how and when the user plays the game. As a consequence, the game provider can ﬁne-tune the diﬃculty of the game, suggest other related games that would most likely interest the user, and oﬀer additional equipment and enhancements for the character based on the user’s age, gender, and interests. This information may get stored locally or uploaded to the game provider’s cloud to analyze the gaming habits and opportunities for upsell and cross-sell, and identify archetypical proﬁles of speciﬁc kinds of users.

●Smartphones provide another rich source of data. In addition to messaging and basic phone usage, they store and transmit data about Internet usage, SMS usage, and real-time location. This metadata can be used for analyzing traﬃc patterns by scanning the density of smart- phones in locations to track the speed of cars or the relative traﬃc congestion on busy roads. In this way, GPS devices in cars can give drivers real-time updates and oﬀer alternative routes to avoid traﬃc delays.

●Retail shopping loyalty cards record not just the amount an individual spends, but the locations of stores that person visits, the kinds of products purchased, the stores where goods are purchased most often, and the combinations of products purchased together. Collecting this data provides insights into shopping and travel habits and the likelihood of successful advertisement targeting for certain types of retail promotions.

●**Data collectors** ●Data results from a cable TV provider tracking the shows a person watches, which TV channels someone will and will not pay for to watch on demand, and the prices someone is willing to pay for premium TV content

●Retail stores tracking the path a customer takes through their store while pushing a shop- ping cart with an RFID chip so they can gauge which products get the most foot traﬃc using geospatial data collected from the RFID chips(Radio Frequency Identification)

●**Data aggregators** make sense of the data collected from the various entities from the “SensorNet” or the “Internet of Things.” These organizations compile data from the devices and usage patterns collected by government agencies, retail stores,

and websites. In turn, they can choose to transform and package the data as products to sell to list brokers, who may want to generate marketing lists of people who may be good targets for speciﬁc ad campaigns.

## WHAT IS A DATA ECOSYSTEM?

The term **data ecosystem** refers to the programming languages, packages, algorithms, cloud-computing services, and general infrastructure an organization uses to collect, store, analyze, and leverage data.

No two organizations leverage the same data in the same way. As such, each organization has a unique data ecosystem. These ecosystems may overlap in some cases, particularly when data is pulled or scraped from a public source, or when third-party providers are leveraged (for example, cloud storage providers).

In the Harvard Online course [Data Science Principles](https://www.harvardonline.harvard.edu/course/data-science-principles), the concept of the data ecosystem is explored through the lens of key stages in the[data project life cycle](https://online.hbs.edu/blog/post/data-life-cycle): sensing, collection, wrangling, analysis, and storage.



## COMPONENTS OF A DATA ECOSYSTEM

### **1. Sensing**

**Sensing** refers to the process of identifying data sources for your project. It involves evaluating the quality of data so you can better understand whether it’s valuable. This evaluation includes asking such questions as:

* Is the data accurate?
* Is the data recent and up to date?
* Is the data complete?
* Is the data valid? Can it be trusted?

Data can be sourced from internal sources, such as databases, spreadsheets, CRMs, and other software. It can also be sourced from external sources, such as websites or third-party data aggregators.

Key pieces of the data ecosystem leveraged in this stage include:

* **Internal data sources:** Proprietary databases, spreadsheets, and other resources that originate from within your organization
* **External data sources:** Databases, spreadsheets, websites, and other data sources that originate from outside your organization
* **Software:** Custom software that exists for the sole purpose of data sensing
* **Algorithms:** A set of steps or rules that automates the process of evaluating data for accuracy and completion before it’s used

### **2. Collection**

Once a potential data source has been identified, data must be **collected**.Data collection can be completed through manual or automated processes. That being said, it generally isn’t feasible to manually perform large-scale data collection. That’s why data scientists use programming languages to write software designed to automate the data collection process.

For example, it’s possible to write a piece of code designed to “scrape” relevant information from a website (aptly named a **web scraper**). It’s also possible to design and code an **application programming interface**, or **API**, to directly extract specific information from a database or interact with a web application.

Key pieces of the data ecosystem leveraged in this stage include:

* **Various programming languages**: These include R, Python, SQL, and JavaScript
* **Code packages and libraries:** Existing code that’s been written and tested and allows data scientists to generate programs more quickly and efficiently
* **APIs:** Software programs designed to interact with other applications and extract data

### **3. Wrangling**

[**Data wrangling**](https://online.hbs.edu/blog/post/data-wrangling) is a set of processes designed to transform raw data into a more usable format. Depending on the quality of the data in question, it may involve merging multiple datasets, identifying and filling gaps in data, deleting unnecessary or incorrect data, and “cleaning” and structuring data for future analysis.

As with data collection, data wrangling can be performed manually or in an automated fashion. If a dataset is small enough, manual processes can work well. For most larger data projects, the amount of data is too vast and requires automation.

Key pieces of the data ecosystem leveraged in this stage include:

* **Algorithms:**A series of steps or rules to be followed to solve a problem (in this case, the evaluation and manipulation of data)
* **Various programming languages:** These include R, Python, SQL, and JavaScript, and can be used to write algorithms
* **Data wrangling tools:**A variety of data wrangling tools can be purchased or sourced for free to perform parts of the data wrangling process. OpenRefine, DataWrangler, and CSVKit are all examples.

### **4. Analysis**

After raw data has been inspected and transformed into a readily usable state, it can be **analyzed.** Depending on the specific challenge your data project seeks to address, this analysis can be diagnostic, descriptive, predictive, or prescriptive. While each of these forms of analysis is unique, they rely on the same processes and tools.

Typically, your analysis begins with some form of automation, especially when datasets are exceptionally large. After automated processes have been completed, data analysts use their expertise to glean additional insights.

Key pieces of the data ecosystem leveraged in this stage include:

* **Algorithms:**A series of steps or rules to be followed to solve a problem (in this case, the analysis of various data points)
* **Statistical models:** Mathematical models used to investigate and interpret data
* [**Data visualization tools**](https://online.hbs.edu/blog/post/data-visualization-tools)**:** These include Tableau, Microsoft BI, and Google Charts, which can generate graphical representations of data. Data visualization software may also have other functionality you can leverage.

### **5. Storage**

Throughout all of the data life cycle stages, data must be **stored** in a way that’s both secure and accessible. The exact medium used for storage is dictated by your organization’s [data governance](https://online.hbs.edu/blog/post/data-governance) procedures.

Key pieces of the data ecosystem leveraged in this stage include:

* **Cloud-based storage solutions:** These allow an organization to store data off-site and access it remotely
* **On-site servers:** These give organizations a greater sense of control over how data is stored and used
* **Other storage media:**These include hard drives, USB devices, CD-ROMs, and floppy disks

# **Big Data Examples**

1. Transportation

Big Data powers the GPS smartphone applications most of us depend on to get from place to place in the least amount of time. GPS data sources include satellite images and government agencies.

Airplanes generate enormous volumes of data, on the order of 1,000 gigabytes for transatlantic flights. Aviation analytics systems ingest all of this to analyze fuel efficiency, passenger and cargo weights, and weather conditions, with a view toward optimizing safety and energy consumption.

Big Data simplifies and streamlines transportation through:

* Congestion management and traffic control  
  Thanks to Big Data analytics, Google Maps can now tell you the least traffic-prone route to any destination.
* Route planning  
  Different itineraries can be compared in terms of user needs, fuel consumption, and other factors to plan for maximize efficiency.
* Traffic safety  
  Real-time processing and predictive analytics are used to pinpoint accident-prone areas.

* + Advertising and Marketing

Ads have always been targeted towards specific consumer segments. In the past, marketers have employed TV and radio preferences, survey responses, and focus groups to try to ascertain people’s likely responses to campaigns. At best, these methods amounted to educated guesswork.

Today, advertisers buy or gather huge quantities of data to identify what consumers actually click on, search for, and “like.” Marketing campaigns are also monitored for effectiveness using click-through rates, views, and other precise metrics.

For example, Amazon accumulates massive data stories on the purchases, delivery methods, and payment preferences of its millions of customers. The company then sells ad placements that can be highly targeted to very specific segments and subgroups.

1. Banking and Financial Services

The financial industry puts Big Data and analytics to highly productive use, for:

* Fraud detection  
  Banks monitor credit cardholders’ purchasing patterns and other activity to flag atypical movements and anomalies that may signal fraudulent transactions.
* Risk management  
  Big Data analytics enable banks to monitor and report on operational processes, KPIs, and employee activities.
* Customer relationship optimization  
  Financial institutions analyze data from website usage and transactions to better understand how to convert prospects to customers and incentivize greater use of various financial products.
* Personalized marketing  
  Banks use Big Data to construct rich profiles of individual customer lifestyles, preferences, and goals, which are then utilized for micro-targeted marketing initiatives.

1. Government

Government agencies collect voluminous quantities of data, but many, especially at the local level, don’t employ modern data mining and analytics techniques to extract real value from it.

Examples of agencies that do include the IRS and the Social Security Administration, which use data analysis to identify tax fraud and fraudulent disability claims. The FBI and SEC apply Big Data strategies to monitor markets in their quest to detect criminal business activities. For years now, the Federal Housing Authority has been using Big Data analytics to forecast mortgage default and repayment rates.

The Centers for Disease Control tracks the spread of infectious illnesses using data from social media, and the FDA deploys Big Data techniques across testing labs to investigate patterns of foodborne illness. The U.S. Department of Agriculture supports agribusiness and ranching by developing Big Data-driven technologies.

Military agencies, with expert assistance from a sizable ecosystem of defense contractors, make sophisticated and extensive use of data-driven insights for domestic intelligence, foreign surveillance, and cybersecurity.

1. Media and Entertainment

The entertainment industry harnesses Big Data to glean insights from customer reviews, predict audience interests and preferences, optimize programming schedules, and target marketing campaigns.

Two conspicuous examples are Amazon Prime, which uses Big Data analytics to recommend programming for individual users, and Spotify, which does the same to offer personalized music suggestions.

1. Meteorology

Weather satellites and sensors all over the world collect large amounts of data for tracking environmental conditions. Meteorologists use Big Data to:

* Study natural disaster patterns
* Prepare weather forecasts
* Understand the impact of global warming
* Predict the availability of drinking water in various world regions
* Provide early warning of impending crises such as hurricanes and tsunamis

1. Healthcare

Big Data is slowly but surely making a major impact on the huge healthcare industry. Wearable devices and sensors collect patient data which is then fed in real-time to individuals’ electronic health records. Providers and practice organizations are now using Big Data for a number of purposes, including these:

* Prediction of epidemic outbreaks
* Early symptom detection to avoid preventable diseases
* Electronic health records
* Real-time alerting
* Enhancing patient engagement
* Prediction and prevention of serious medical conditions
* Strategic planning
* Research acceleration
* Telemedicine
* Enhanced analysis of medical images

1. Cybersecurity

While Big Data can expose businesses to a greater risk of cyber attacks, the same data stores can be used to prevent and counteract online crime through the power of machine learning and analytics. Historical data analysis can yield intelligence to create more effective threat controls. And machine learning can warn businesses when deviations from normal patterns and sequences occur, so that effective countermeasures can be taken against threats such as ransom ware attacks, malicious insider programs, and attempts at unauthorized access.

After a company has suffered an intrusion or data theft, post-attack analysis can uncover the methods used, and machine learning can then be deployed to devise safeguards that will foil similar attempts in the future.

1. Education

Administrators, faculty, and stakeholders are embracing Big Data to help improve their curricula, attract the best talent, and optimize the student experience. Examples include:

* Customizing curricula  
  Big Data enables academic programs to be tailored to the needs of individual students, often drawing on a combination of online learning, traditional on-site classes, and independent study.
* Reducing dropout rates  
  Predictive analytics give educational institutions insights on student results, responses to proposed programs of study, and input on how students fare in the job market after graduation.
* Improving student outcomes  
  Analyzing students’ personal “data trails” can provide a better understanding of their learning styles and behaviors, and be used to create an optimal learning environment.
* Targeted international recruiting  
  Big Data analysis helps institutions more accurately predict applicants’ likely success. Conversely, it aids international students in pinpointing the schools best matched to their academic goals and most likely to admit them.

# Life Cycle Phases of Data Analytics

**Data Analytics Lifecycle:**

The [Data analytic](https://www.geeksforgeeks.org/data-analytics-and-its-type/) lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project. To address the distinct requirements for performing analysis on Big Data, step – by – step methodology is needed to organize the activities and tasks involved with acquiring, processing, analysing, and repurposing data.

**Phase 1: Discovery –**

* The data science team learn and investigate the problem.
* Develop context and understanding.
* Come to know about data sources needed and available for the project.
* The team formulates initial hypothesis that can be later tested with data.

**Phase 2: Data Preparation –**

* Steps to explore, preprocess, and condition data prior to modeling and analysis.
* It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.
* Data preparation tasks are likely to be performed multiple times and not in predefined order.
* Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.

**Phase 3: Model Planning –**

* Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
* In this phase, data science team develop data sets for training, testing, and production purposes.
* Team builds and executes models based on the work done in the model planning phase.
* Several tools commonly used for this phase are – Matlab, STASTICA.

**Phase 4: Model Building –**

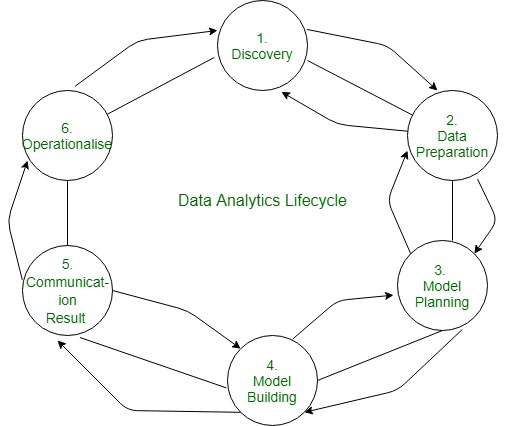
* Team develops datasets for testing, training, and production purposes.
* Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
* Free or open-source tools – Rand PL/R, Octave, WEKA.
* Commercial tools – Matlab , STASTICA.

**Phase 5: Communication Results –**

* After executing model team need to compare outcomes of modeling to criteria established for success and failure.
* Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
* Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

**Phase 6: Operationalize –**

* The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
* This approach enables team to learn about performance and related constraints of the model in production environment on small scale  , and make adjustments before full deployment.
* The team delivers final reports, briefings, codes.
* Free or open source tools – Octave, WEKA, SQL, MADlib.



UNIT 2

# What is Business Intelligence?

## What is Business Intelligence?

BI(Business Intelligence) is a set of processes, architectures, and technologies that convert raw data into meaningful information that drives profitable business actions. It is a suite of software and services to transform data into actionable intelligence and knowledge.

BI has a direct impact on organization’s strategic, tactical and operational business decisions. BI supports fact-based decision making using historical data rather than assumptions and gut feeling.

BI tools perform data analysis and create reports, summaries, dashboards, maps, graphs, and charts to provide users with detailed intelligence about the nature of the business.

## Why is BI important?

* Measurement: creating KPI (Key Performance Indicators) based on historic data
* Identify and set benchmarks for varied processes.
* With BI systems organizations can identify market trends and spot business problems that need to be addressed.
* BI helps on data visualization that enhances the data quality and thereby the quality of decision making.
* BI systems can be used not just by enterprises but SME (Small and Medium Enterprises)

**How Business Intelligence systems are implemented?**

Here are the steps:

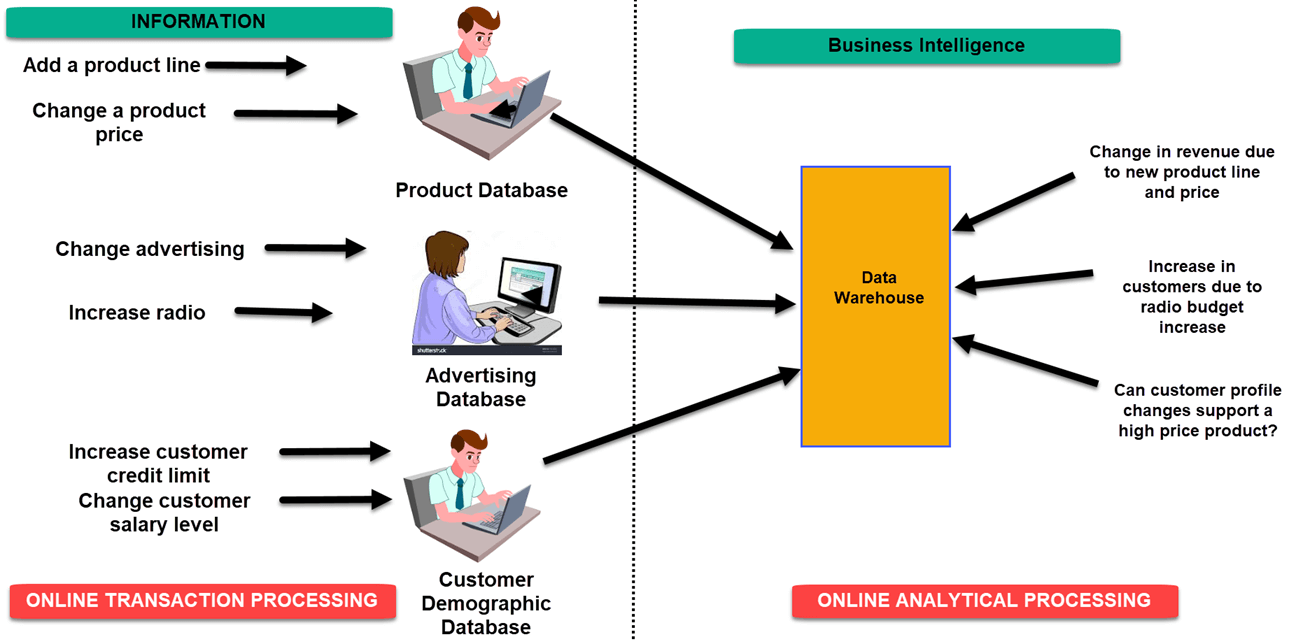
**Step 1**) Raw Data from corporate databases is extracted. The data could be spread across multiple systems heterogeneous systems.

**Step 2)** The data is cleaned and transformed into the data warehouse. The table can be linked, and data cubes are formed.

**Step 3)** Using BI system the user can ask quires, request ad-hoc reports or conduct any other analysis.

## Examples of Business Intelligence System used in Practice

**Example 1:**



In an Online Transaction Processing ([OLTP](https://www.guru99.com/what-is-oltp.html)) system information that could be fed into product database could be

* add a product line
* change a product price

Correspondingly, in a Business Intelligence system query that would beexecuted for the product subject area could be did the addition of new product line or change in product price increase revenues

In an advertising database of OLTP system query that could be executed

* Changed in advertisement options
* Increase radio budget

Correspondigly, in BI system query that could be executed would be how many new clients added due to change in radio budget

In OLTP system dealing with customer demographic data bases data that could be fed would be

* increase customer credit limit
* change in customer salary level

Correspondingly in the [OLAP](https://www.guru99.com/online-analytical-processing.html) system query that could be executed would be can customer profile changes support support higher product price.

# Pattern Recognition

**Pattern** is everything around in this digital world. A pattern can either be seen physically or it can be observed mathematically by applying algorithms.

**Example:** The colors on the clothes, speech pattern, etc. In computer science, a pattern is represented using vector feature values.   
**What is Pattern Recognition?**   
**Pattern recognition** is the process of recognizing patterns by using a machine learning algorithm. Pattern recognition can be defined as the classification of data based on knowledge already gained or on statistical information extracted from patterns and/or their representation. One of the important aspects of pattern recognition is its application potential.

**Examples:** Speech recognition, speaker identification, multimedia document recognition (MDR), automatic medical diagnosis.   
In a typical pattern recognition application, the raw data is processed and converted into a form that is amenable for a machine to use. Pattern recognition involves the classification and cluster of patterns.

* In classification, an appropriate class label is assigned to a pattern based on an abstraction that is generated using a set of training patterns or domain knowledge. Classification is used in supervised learning.
* Clustering generated a partition of the data which helps decision making, the specific decision-making activity of interest to us. Clustering is used in unsupervised learning.

**Features** may be represented as continuous, discrete, or discrete binary variables. A feature is a function of one or more measurements, computed so that it quantifies some significant characteristics of the object.

**Example:** consider our face then eyes, ears, nose, etc are features of the face.   
A set of features that are taken together, forms the **features vector**.

**Example:** In the above example of a face, if all the features (eyes, ears, nose, etc) are taken together then the sequence is a feature vector([eyes, ears, nose]). The feature vector is the sequence of a feature represented as a d-dimensional column vector. In the case of speech, MFCC (Mel-frequency Cepstral Coefficient) is the spectral feature of the speech. The sequence of the first 13 features forms a feature vector.   
**Pattern recognition possesses the following features:**

* Pattern recognition system should recognize familiar patterns quickly and accurate
* Recognize and classify unfamiliar objects
* Accurately recognize shapes and objects from different angles
* Identify patterns and objects even when partly hidden
* Recognize patterns quickly with ease, and with automaticity.

**Training and Learning in Pattern Recognition**   
[**Learning**](https://www.geeksforgeeks.org/getting-started-machine-learning/) is a phenomenon through which a system gets trained and becomes adaptable to give results in an accurate manner. Learning is the most important phase as to how well the system performs on the data provided to the system depends on which algorithms are used on the data. The entire dataset is divided into two categories, one which is used in training the model i.e. Training set, and the other that is used in testing the model after training, i.e. Testing set.

* **Training set:**   
  The training set is used to build a model. It consists of the set of images that are used to train the system. Training rules and algorithms are used to give relevant information on how to associate input data with output decisions. The system is trained by applying these algorithms to the dataset, all the relevant information is extracted from the data, and results are obtained. Generally, 80% of the data of the dataset is taken for training data.
* **Testing set:**   
  Testing data is used to test the system. It is the set of data that is used to verify whether the system is producing the correct output after being trained or not. Generally, 20% of the data of the dataset is used for testing. Testing data is used to measure the accuracy of the system. For example, a system that identifies which category a particular flower belongs to is able to identify seven categories of flowers correctly out of ten and the rest of others wrong, then the accuracy is 70 %

https://media.geeksforgeeks.org/wp-content/uploads/patt_intro.png

**Advantages:**

* Pattern recognition solves classification problems
* Pattern recognition solves the problem of fake biometric detection.
* It is useful for cloth pattern recognition for visually impaired blind people.
* It helps in speaker diarization.
* We can recognize particular objects from different angles.

**Disadvantages:**

* The syntactic pattern recognition approach is complex to implement and it is a very slow process.
* Sometimes to get better accuracy, a larger dataset is required.
* It cannot explain why a particular object is recognized.   
  Example: my face vs my friend’s face.

**Applications:**

* **Image processing, segmentation, and analysis**   
  Pattern recognition is used to give human recognition intelligence to machines that are required in image processing.
* **Computer vision**   
  Pattern recognition is used to extract meaningful features from given image/video samples and is used in computer vision for various applications like biological and biomedical imaging.
* **Seismic analysis**   
  The pattern recognition approach is used for the discovery, imaging, and interpretation of temporal patterns in seismic array recordings. Statistical pattern recognition is implemented and used in different types of seismic analysis models.
* **Radar signal classification/analysis**   
  Pattern recognition and signal processing methods are used in various applications of radar signal classifications like AP mine detection and identification.
* **Speech recognition**   
  The greatest success in speech recognition has been obtained using pattern recognition paradigms. It is used in various algorithms of speech recognition which tries to avoid the problems of using a phoneme level of description and treats larger units such as words as pattern
* **Fingerprint identification**   
  Fingerprint recognition technology is a dominant technology in the biometric market. A number of recognition methods have been used to perform fingerprint matching out of which pattern recognition approaches are widely used.

## Data Processing Cycle

The data processing cycle consists of a series of steps where raw data (input) is fed into a system to produce actionable insights (output). Each step is taken in a specific order, but the entire process is repeated in a cyclic manner. The first data processing cycle's output can be stored and fed as the input for the next cycle, as the illustration below shows us.

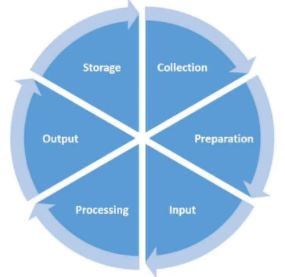


Fig: Data processing cycle ([source](https://planningtank.com/computer-applications/data-processing-cycle))

Generally, there are six main steps in the data processing cycle:

### **Step 1: Collection**

The [collection of raw data](https://www.simplilearn.com/what-is-data-collection-article) is the first step of the data processing cycle. The type of raw data collected has a huge impact on the output produced. Hence, raw data should be gathered from defined and accurate sources so that the subsequent findings are valid and usable. Raw data can include monetary figures, website cookies, profit/loss statements of a company, user behavior, etc.

### **Step 2: Preparation**

Data preparation or [data cleaning](https://www.simplilearn.com/data-cleaning-why-and-how-to-get-started-article) is the process of sorting and filtering the raw data to remove unnecessary and inaccurate data. Raw data is checked for errors, duplication, miscalculations or missing data, and transformed into a suitable form for further analysis and processing. This is done to ensure that only the highest quality data is fed into the processing unit.

The purpose of this step to remove bad data (redundant, incomplete, or incorrect data) so as to begin assembling high-quality information so that it can be used in the best possible way for [business intelligence](https://www.simplilearn.com/what-is-business-intelligence-article).

### **Step 3: Input**

In this step, the raw data is converted into machine readable form and fed into the processing unit. This can be in the form of data entry through a keyboard, scanner or any other input source.

### **Step 4: Data Processing**

In this step, the raw data is subjected to various data processing methods using machine learning and artificial intelligence algorithms to generate a desirable output. This step may vary slightly from process to process depending on the source of data being processed ([data lakes](https://www.simplilearn.com/data-lakes-and-data-analytics-article), online databases, connected devices, etc.) and the intended use of the output.

### **Step 5: Output**

The data is finally transmitted and displayed to the user in a readable form like graphs, tables, vector files, audio, video, documents, etc. This output can be stored and further processed in the next data processing cycle.

### **Step 6: Storage**

The last step of the data processing cycle is storage, where data and metadata are stored for further use. This allows for quick access and retrieval of information whenever needed, and also allows it to be used as input in the next data processing cycle directly.

UNIT 3

BI FOR BETTER DECISION

Business intelligence helps extract crucial facts from a vast amount of unstructured data and transform them into actionable information that enables companies to make informed strategic decisions, improving operational efficiency and business productivity. This actionable information provides crucial insights that reveal the underlying currents of customer behavior, their likes and dislikes, online shopping experience etc.

Everyday contact centers handle thousands of customer interactions that hold virtually untapped vital insights which can be efficiently leveraged to provide immense value to businesses. This information remains largely under utilized except when evaluated for Quality Assurance purposes.

Business intelligence can arm companies with rich data resources that can help them achieve their business goals and targets by guiding timely strategic decisions. Interactions with customers in the form of voice calls, chat interactions, emails etc can be thoroughly analyzed by business intelligence gathering teams to shed light on aspects such as customer preferences; likes and dislikes of customers; technical difficulties faced by customers; customer reactions to promotions; and the online shopping customer experience on a website. This analysis can lead to improving conversion rates and much more.

Apart from the above, other advantages companies can realize due to business intelligence include:

### **1. Improve Business Productivity**

Businesses can channel their vital resources and workforce to foster business productivity as they need not divert their resources to business intelligence (BI) gathering work. Since all BI gathering work is handled by the BI team, it results in cost saving, time saving and efficient reporting supporting improved business productivity. The BI gathering team can pull up vital information from customer interactions and present it in a manner that is easy to understand, communicate and execute. BI gathering teams ensure that detailed reports are provided to companies ensuring that all the crucial information is presented in a well-documented and presentable manner so that information can be used to efficiently manage the business.

### **2. Crucial Information Easily Accessed**

Business intelligence provides crucial information to companies, improving their ability to make quick decisions and generating a competitive advantage.

### **3. Good Return on Investment (ROI)**

BI ensures stronger ROI as companies can significantly reduce costs, enhance revenue, improve margin, drive cost avoidance and much more. Also, BI ensure qualitative ROI such as improvements in business and process efficiency, employee productivity, better strategic decision making, customer satisfaction and much more. Companies relying on business intelligence can reduce operational costs due to improved operational efficiency and optimized business processes.

### **4. Informed Decision Making**

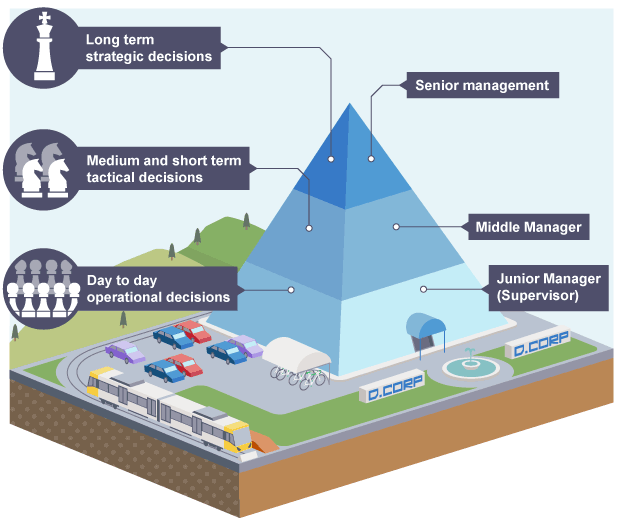
Reporting based on accurate and timely information helps companies measure the performance of their processes. Business Intelligence helps companies make informed decisions on strategic issues by providing crucial information on current and historical performance of the company along with future trends, expected demands, customer behavior etc.

Business Intelligence teams ensure that the company receives real-time advanced reports to ensure that the company can efficiently utilize the information at hand to better manage the business.

# Types of decisions

Decisions are part of the manager's remit. Difficult choices may have to be made for the common good of the organisation. There are three types of decision in business:

* strategic
* tactical
* operational



| **strategic** | **tactical** | **operational** |
| --- | --- | --- |
| long term | medium term | day-to-day |
| complex | less complex | simple and routine |
| made by senior managers | made by middle managers | made by junior managers |
| eg aiming to be market leader | eg launching new product/opening new branches | eg regular ordering of supplies/creating staff rota |

## Tips for making a business decision

Here are a few tips for making good business decisions that benefit your company:

* **Consider the end-goal.** When making an important business decision, take time to consider what you want the end goal to be. For example. A department manager identifies an end-goal as being able to keep every employee in their department even during an economic downturn. Because of this end-goal, they make the decision to reduce their department's budget and streamline department activities.
* **Weigh each potential outcome.** As you figure out which business decision to choose, be sure to consider each potential outcome of a business decision. This allows you to determine the potential advantages and disadvantages of a potential business decision and whether advantages outweigh the disadvantages in the long-term.
* **Seek out advice from qualified professionals or company employees.** If you aren't sure which business decision to make, you can benefit from seeking out advice from your coworkers, department employees, external consultants or upper-management personnel. If the majority of those you speak with choose one decision over another, that could help you determine which decision you need to make that works best for your business.

#### 1. Strategic decisions:

Strategic decisions are major choices of actions and influence whole or a major part of business enterprise. They contribute directly to the achievement of common goals of the enterprise. They have long-term implications on the business en­terprise.

They may involve major departures from practices and procedures being followed earlier. Generally, strategic decision is unstructured and thus, a manager has to apply his business judge­ment, evaluation and intuition into the definition of the problem. These decisions are based on partial knowledge of the environmen­tal factors which are uncertain and dynamic. Such decisions are taken at the higher level of management.

#### Tactical decisions:

* These decisions relate to the implementation of strategic decisions. They are directed towards developing divi­sional plans, structuring workflows, establishing distribution chan­nels, acquisition of resources such as men, materials and money. These decisions are taken at the middle level of management.

#### 3. Operational decisions:

* These decisions relate to day-to-day op­erations of the enterprise. They have a short-term horizon as they are taken repetitively. These decisions are based on facts regarding the events and do not require much of business judge­ment. Operational decisions are taken at lower levels of man­agement. As the information is needed for helping the manager to take rational, well informed decisions, information systems need to fo­cus on the process of managerial decision making.

## What is a business intelligence (BI) tool?

Business intelligence tools collect, process, and analyze large amounts of structured and unstructured data from both internal and external systems. Data sources might include documents, images, email, videos, journals, books, social media posts, files, and more. BI tools find this information through queries, which can present the data in user-friendly formats such as reports, dashboards, charts, and graphs.

The tools can perform functions such as data mining, data visualization, performance management, analytics, reporting, text mining, predictive analytics, and much more. As a result, employees can harness this information to make better decisions based on predictions, market trends, and key performance indicators (KPIs).

## Benefits of BI tools

BI tools can help your business take smart, agile steps toward accomplishing bigger goals.

**Centralized data:**All of your data, in one place. Companies collect data from numerous databases, portals, customer relationship management (CRM) systems, enterprise resource planning (ERP) systems, and more. To make sense of it all, you’ll need business intelligence tools to pool the data together and deliver certain types of views (issues, trends, analytics) based on your queries or what you want to know.

**Self-sufficiency:**Data no longer just belongs to your company’s IT team. Now, any employee without technical expertise, even interns, can access and analyze data necessary for their respective teams. They can generate reports on all electronics sales in 2020 or visualize patient records to identify the number of COVID-19 cases in 2022.

**Make predictions:**With access to so much data from the past and present, employees can make evidence-based decisions. Predictive analytics and forecasting enable users to generate insights based on a product or service’s performance history. If a business condition changes, the intelligent tools can automatically figure out the anomalies and you’ll be able to react to disruptions as they arise.

**Automatic reports:**Instead of inputting data manually into Excel spreadsheets or toggling between different tools, many BI tools are automated. If you need a report on a product over a specific period of time, the tool will generate that for you. If that information is significant and you want to add it to a presentation, you can create interactive visualizations and download any charts or graphs you may need.

**Reduces business costs:**BI tools can do so much, from analyzing consumer behavior and sales forecasting to real-time process monitoring, that analyzing, planning, and reporting processes are way more efficient and accurate than ever before. In fact, more than 50 percent of BI tools users reported that these benefits helped them reduce costs and increase revenue [[1](https://bi-survey.com/benefits-business-intelligence)].

## Top 5 business intelligence tools

Today, there are dozens of BI tools available. To determine the top five, these were the most commonly cited among the websites that ranked the best BI tools. These five BI tools are not ranked in any particular order.

### 1. Microsoft Power BI

One of the most popular BI tools is [Power BI](https://powerbi.microsoft.com/en-us/what-is-power-bi/), offered by leading software giant Microsoft. This tool is downloadable software, so you can choose to run analytics either on the cloud or in a reporting server. Syncing with sources such as Facebook, Oracle, and more, generate reports and dashboards in minutes with this interactive tool. It comes with built-in AI capabilities, Excel integration, and data connectors, and offers end-to-end data encryption and real-time access monitoring.

**Learn Power BI with Coursera:** In just two hours, you can learn the basics of [Power BI Desktop](https://www.coursera.org/projects/power-bi-desktop) with this guided project. You’ll load and transform data to create interactive reports and dashboards.

### 2. Tableau

[Tableau](https://www.tableau.com/products/desktop) is known for its user-friendly data visualization capabilities, but it can do more than make pretty charts. Their offering includes live visual analytics, an interface that allows users to drag and drop buttons to spot trends in data quickly. The tool supports data sources such as Microsoft Excel, Box, PDF files, Google Analytics, and more. Its versatility extends to being able to connect with most databases.

**Learn Tableau with Coursera:**There are several options for learning how to use Tableau.

* [Data Visualization with Tableau](https://www.coursera.org/specializations/data-visualization) specialization from the University of California Davis
* [Use Tableau for Your Data Science Workflow](https://www.coursera.org/specializations/tableau-for-data-science-workflows) specialization from the University of California Irvine
* [Data Visualization and Communication](https://www.coursera.org/learn/analytics-tableau) class with Tableau from Duke University
* Guided Project on [Visualizing Citibike Trips with Tableau](https://www.coursera.org/projects/visualizing-citibike-trips-tableau)

### 3. QlikSense

[QlikSense](https://www.qlik.com/us/products/qlik-sense) is a BI tool that emphasizes a self-service approach, meaning that it supports a wide range of analytics use cases, from guided apps and dashboards to custom and embedded analytics. It offers a user-friendly interface optimized for touchscreens, sophisticated AI, and high-performance cloud platforms. Its associative exploration capability, Search & Conversational Analytics, allows users to ask questions and uncover actionable insights, which helps increase data literacy for those new to using BI tools.

### 4. Dundas BI

[Dundas BI](https://www.dundas.com/) is a browser-based BI tool that’s been around for 25 years. Like Tableau, Dundas BI features a drag-and-drop function that allows users to analyze data on their own, without involving their IT team. The tool is known for its simplicity and flexibility through interactive dashboards, reports, and visual analytics. Since its inception as a data visualization tool in 1992, it has evolved into an end-to-end analytics platform that is able to compete with the new BI tools available today.

### 5. Sisense

[Sisense](https://www.sisense.com/) is a user-friendly BI tool that focuses on being simplified and streamlined. With this tool, you can export data from sources like Google Analytics, Salesforce, and more. Its in-chip technology allows for faster data processing compared to other tools. Key features include the ability to embed white-label analytics, meaning a company can fully customize the services to its needs. Like others, it has a drag-and-drop feature. Sisense allows you to share reports and dashboards with your team members as well as externally.

# BI Skills

## 1. Data Mining

A core responsibility of a business intelligence analyst is data visualization, which is only possible if you understand how to extract and mine data. A data mining skill would help you understand where to extract data relevant to the company you're working for.

Normally, data is always available in large data sets, but with data mining skills, you can extract useful information, clean it up and make it available. The three basic methods of data mining are classification, clustering, and association.

## 2. Data Preparation

Once you understand how to mine data, learning how to prepare this data is next. Even though data mining requires that you put similar data together, you still need to organize and structure the extracted data. You want to turn that raw data into analytics-ready data because you can only use organized and structured data in business intelligence.

You'll also need to be proficient in the [data structures that all programmers should know](http://www.makeuseof.com/data-structures-every-programmer-should-know/) because business intelligence requires some programming knowledge. As part of data preparation skills, you need to learn how to filter, clean, and transform data.

## 3. Business Knowledge

Having business knowledge and understanding of your company's business model gives direction on how to channel company resources in the right direction. Business intelligence analysis goes beyond gathering information. You'll need to understand how to use cleaned data to produce the results a company needs. These should be actionable information for making the right decisions.

With vast business knowledge, you can support a company's vision by providing consultancy on leveraging market trends to improve revenue. These skills also help you interpret data in a way that aligns with a company's vision and mission. So, you're not compiling chunks of data. You visualize them and arrange them in a way consistent with business growth.

## 4. Analytical and Critical Thinking

You need to have critical thinking incorporated into the responsibility of a business intelligence analyst. Also, collecting data, analyzing, and deducing conclusions that would be useful to your company places you on a pedestal for success as a business intelligence analyst.

Analytical and critical thinking ability helps you properly scrutinize data and build something useful from your gathered data. Critical thinking also helps you recognize the weaknesses of a company's system based on the data you've gathered.

## 5. Programming Languages

Programming stands at the core of business intelligence analysis. You'll need basic knowledge of SQL programming to help you pre-aggregate, filter, and only select the data necessary for your analysis. With knowledge of SQL queries, you can create views (rows and columns of data) of a few thousand, lessening the amount of data your BI tool has to work with. This process makes analysis easier and faster.

Therefore, learning [basic SQL commands and queries](https://www.makeuseof.com/tag/important-sql-commands-programmer-know/) is an important skill for business intelligence analysts. SQL programming language is also useful for data extraction, data quality, data validation, and creating analytical reports. In addition to the SQL programming language, you also need basic knowledge of Python and R programming for data visualization.

## 6. Statistical Analysis

Through statistical analysis, you can discern market trends and forecast business patterns. It involves organizing and analyzing raw data to draw up reliable information. During statistical analysis, you remove all bias from the data set, leaving you with structured data.

A business intelligence analyst should have statistical analytical skills to describe patterns that can predict future changes in market demands, services, or prices, depending on what company you're working for. You'll need this skill to coin useful information from messy data, reducing uncertainty during decision-making.

## 7. Creation of Reports and Dashboard

Learning how to create reports and dashboards will help you organize and manage your company's data. It helps you store data in patterns that make it easy to calculate the growth and performance of your company. You can also project future company trends with the report you create to present to stakeholders willing to invest in the company.

As a business intelligence analyst, conveying data in the simplest of ways is crucial. Thus, proficiency in creating reports and dashboards that simplify data is important.

## 8. Data Visualization

Data visualization skills include understanding data purpose, navigating data visualization tools, and telling a compelling story using graphs, web maps, charts, and tables. In addition, learning how to present data to a layperson using charts and graphs is essential. Combining statistical data with art and having strong visual design skills are pluses.

Once you can gather cleaned data and run them on different databases, upgrade your data visualization skills to make yourself a better business intelligence analyst. Also, learning when to use what chart type is important. You want to avoid using the wrong chart types, which may confuse people.

## 9. Organization and Time Management

Your ability to efficiently manage your time and interpret data promptly is important for your company's growth. If you work for a healthcare company looking to scale its medical services and get data on the number of individuals they've treated in the past, you'll need to sort these analyses in an organized manner.

## 10. Industry Knowledge

Finally, having industry knowledge about your company is necessary. If you work in the health and fitness or digital marketing niche, you have to be familiar with the latest trends and trajectories. Read the numbers and watch what other competitors are changing. If they were rethinking their strategy, they could have made their analysis.

## Business Intelligence Applications

### Google

The world’s largest and perhaps the greatest tech company Google is known for its data obsession. However, the company doesn’t only use data solutions to surveil customers, rather it also leverages its power in [business intelligence solutions](https://dynamics.folio3.com/blog/business-intelligence-and-analytics/) to improve its in-house processes and improve procedural efficiencies.

The tech giant maintains a dedicated team “People Analytics”, which is meant to collect employees’ opinions regarding common employment queries. For instance, if there is an actual need for managers or what exactly should be the role of managers, or what are the characteristics of best managers, and so on.

Such opinionated queries help the tech giant to improve the company culture inside the company and design more efficient workflow processes and promote innovation within the company.

### Twitter

From the very beginning, the character limit for Tweets has remained a hot topic of debate among users. For some, the 140 character limit for tweets was too restrictive and made it hard for users to express themselves adequately. However, for others (read majority), the character limit actually made Twitter a better platform for rational users. For them, the character limit saved twitted from wordy screeds.

Nonetheless, when Twitter doubled its character limit to 280, there was yet again an uproar with many users calling it the end of twitter’s distinction amongst the other social media platforms. However, what users didn’t know was that the platform had performed its due diligence using business intelligence applications. In reality, the increased character limit just makes it easier and faster for users to express themselves without adjusting and condensing to the count limit.

### Facebook

Facebook with over 2 billion active monthly users is the largest and biggest social media platform. However, the social media giant has also been the root of various controversies from creepy data usage to monopoly and various other controversies.

However, recently the platform has been accused of causing a dip in users’ mental health. The cause for this mental dip is users’ comparison of actual lives to the carefully-curated lives of other people online. Also, the number of “likes” has been impacting the mental health of users; only compounding the issue.

To counter this negative trend, Facebook is already experimenting with making the number of likes private to the original poster only. While this experiment isn’t started in the USA yet, the company is already performing these tests in certain countries including Canada, and Japan.

### TIKTOK

TikTok takes data analytics to another level. The platform’s data analytics capabilities are so efficient that it has made browsing for videos almost obsolete for users. The app immediately starts showing users the related videos, as soon as they open its “For You” tab. With time, the users are able to train the app’s algorithms for their liking by reacting to the videos. As the users indicate, within a short time the number of “Hard-likes” diminishes for users, indicating the powerful algorithms that are able to identify users’ taste correctly every time. This reflects the strong utilization of business intelligence by the platform to train its algorithms for users’ liking.

### FOOD & DRINK

American diet can be bizarre for many people outside of the USA. The people here eat just way too many calories, whereas, they also get less time to eat proper meals during the day.

For food companies, the eating habits and routine of people are just business opportunities, as they collect business intelligence for people’s eating habits to come up with personalized offerings.

### Starbucks

The Starbucks app is popular for its rewarding system. The app frequently rewards coffee lovers with free drinks. Tempting deals and other rewards. However, at the back end, the app utilizes business intelligence to track the users’ preferences for their preferred drinks and throw them with tempting deals. And while on the surface it may appear tempting for users to get a free deal for their favorite drink, the ultimate aim of the company is to retain the regular customers and draw them towards the store through customized deals.

### UBER

When Uber launched its food delivery service “UberEats” in 2015, the company immediately becomes an important component in the restaurant and food delivery business. Just like the ride-sharing application, Uber Eats utilizes smart data tracking and business intelligence applications to track and store many details about the restaurants.

## What is a Data Warehouse?

A [data warehouse](https://www.oracle.com/database/what-is-a-data-warehouse/) is a centralized repository where an organization can store substantial amounts of data from multiple source systems and locations. These sources and locations include:

* Customer relationship management (CRM) systems
* Enterprise resource planning (ERP) systems
* SaaS tools
* Relational databases
* Transactional databases

Essentially, data warehouses house all the essential data businesses need to run analyses and collect the valuable business insights reflected in that data. The data warehouse is the ultimate destination supporting [business intelligence](https://www.integrate.io/glossary/what-is-business-intelligence/) (BI) activities such as identifying trends and making smarter organizational decisions.

Some of the top benefits of a data warehouse include:

* **Consistency:**Data warehouses compile data from multiple sources and clean it, creating consistency among all data.
* **Security:**A data warehouse provides security because it's a stable, nonvolatile entity that doesn't change over time.
* **Saves time:**A data warehouse helps organizations and individual employees save time by acquiring data within seconds.
* **Data governance:**A data warehouse can make it easier to adhere to data governance guidelines such as [GDPR](https://gdpr-info.eu/) and CCPA rather than relying on legacy systems to safeguard sensitive data.
* **Removes data silos:**A data warehouse removes existing silos in your organization, such as legacy systems in different departments unable to communicate.

Consider this example of a data warehouse in action: you’re trying to figure out the overall value of your leads in Salesforce. You could [push your Salesforce data into your data warehouse](https://www.integrate.io/integrations/salesforce/), set up a schema, and run a query that tells you which marketing activities led to your highest-value prospects. The results of the query outline how to target those high-value prospects more easily with focused marketing materials, move them through your sales funnels, and — ultimately — increase revenue.

Examples of data warehouses include Snowflake, Amazon Redshift, Microsoft Azure, and IBM Db2.

 UNIT 4

## 8 Steps in Data Warehouse Design

Here are the eight core steps that go into data warehouse design:

### 1. Defining Business Requirements (or Requirements Gathering)

Data warehouse design is a business-wide journey. Data warehouses touch all areas of your business, so every department needs to be on board with the design. Since your warehouse is only as powerful as the data it contains, aligning departmental needs and goals with the overall project is critical to your success.

So, if you currently can't combine all your sales data with all your marketing data, your overall query results are missing some critical components. Knowing which leads are valuable can help you get more value from your marketing data.

Every department needs to understand the purpose of the data warehouse, how it benefits them, and what kinds of results they can expect from your warehousing solution.

This Requirements Gathering stage should focus on the following objectives:

* Aligning departmental goals with the overall project
* Determining the scope of the project in relation to business processes
* Discovering your current and future needs by diving deep into your data (finding out what data is useful for analysis) and your current tech stack (where your data is currently siloed and not being used)
* Creating a disaster recovery plan in the case of system failure
* Thinking about each layer of security (e.g., threat detection, threat mitigation, identity controls, monitoring, risk reduction, etc.)
* Anticipating compliance needs and mitigating regulatory risks

You can think of this as your overall data warehouse blueprint. But this phase is more about determining your business needs, aligning those to your data warehouse, and, most importantly, getting everyone on board with the data warehousing solution.

### 2. Setting Up Your Physical Environments

Data warehouses typically have three primary physical environments — development, testing, and production. This mimics standard software development best practices, and your three environments exist on completely separate physical servers.

Why do you need three separate environments?

* You need a way to test changes before they move into the production environment.
* Some security best practices require that testers and developers never have access to production data.
* Running tests against data typically uses extreme data sets or random sets of data from the production environment — and you need a unique server to execute these tests en masse.
* Having a development environment is a necessity, and dev environments exist in a unique state of flux compared to production or test environments.
* Production environments have much higher workloads (your whole business is using it), so trying to run tests or develop in that environment can be stressful for both team members and servers.
* Data integrity is much easier to track, and issues are easier to contain when you have three environments running. It makes headhunting issues less stressful on your workloads, and data flow in production and testing environments can be stalled without impacting end users.
* Running tests can often introduce breakpoints and hang your entire server. That's not something you want happening in your production environment.
* Imagine sharing resources between production, testing, and development. You don’t want that! Testing, development, and production environments all have different resource needs, and trying to combine all functions into one server can be catastrophic for performance.

Remember, [BI development](https://www.integrate.io/blog/top-business-intelligence-tools/) is an ongoing process that really never grinds to a halt. This is especially true in [Agile](https://www.integrate.io/glossary/what-is-agile-methodology/)/DevOps approaches to the software development lifecycle, which all require separate environments due to the sheer magnitude of constant changes and adaptations.

You can choose to run more than these three environments, and some business users choose to add additional environments for specific business needs. Integrate.io has seen staging environments that are separate from testing solely for quality assurance work, as well as demo and integration environments specifically for testing integrations.

You should have these three core environments, but you can layer in additional settings to fit your unique business goals.

### 3. Data Warehouse Design: Introducing Data Modeling

Data modeling is the process of visualizing data distribution in your warehouse. Think of it as a blueprint. Before you start building a house, it's important to know what goes where and why it goes there. That's what data modeling is to data warehouses.

Data modeling helps you:

* Visualize the relationships between data
* Set standardized naming conventions
* Create relationships between data sets
* Establish compliance and security processes
* Align your processes with your overarching IT goals

The above benefits of data modeling help improve decision-making throughout your organization.

However, data modeling is probably the most complex phase of data warehouse design, and there are multiple data modeling techniques businesses can choose from for warehouse design. Before jumping into a few of the most popular data modeling techniques, let's take a look at the differences between data warehouses and data marts:

A data warehouse is a system to store data in (or push data into) to run analytics and queries. A [data mart](https://www.integrate.io/blog/data-mart-vs-data-warehouse/), on the other hand, is an area within a data warehouse that stores data for a specific business function.

So, say you've built your entire data warehouse. That's great! But does it account for how different departments will use the data? Your sales team will use that data warehouse in a vastly different way than your legal team. Plus, certain workflows and data sets are only valuable to certain teams. Data marts are where all those team-specific data sets are stored, and related queries are processed.

Data modeling typically takes place at the data mart level and branches out into your data warehouse. It's the logic behind how you store certain data in relation to other data.

The three most popular data models for warehouses are:

1. Snowflake schema
2. Star schema
3. Galaxy schema

You should choose and develop a data model to guide your overall data architecture within your warehouse. The model you choose will impact the structure of your data warehouse and data marts — which impacts the ways that you utilize [ETL tools](https://www.integrate.io/blog/top-7-etl-tools/) like Integrate.io and run queries on that data.

**Related Reading:** [Snowflake Schema vs Star Schema](https://www.integrate.io/blog/snowflake-schemas-vs-star-schemas-what-are-they-and-how-are-they-different/)

### 4. Choosing Your Extract, Transform, Load (ETL) Solution

[ETL or Extract, Transform, Load](https://www.integrate.io/blog/what-is-etl/) is the process used to pull data out of your current tech stack or existing storage solutions and put it into your warehouse. It goes something like this:

* You **extract** data from a source system and place it into a staging area.
* You **transform** that data into the best format for data analytics. You also remove any duplicated data or inconsistencies that can make analysis difficult.
* You then **load** the data to a data warehouse before pushing it through BI tools like Tableau and Looker.

Normally, [ETL](https://www.informatica.com/resources/articles/what-is-etl.html) is a complicated process that requires manual pipeline-building and lots of code. Building these pipelines can take weeks or even months and might require a data engineering team. That’s where ETL solutions come in. They automate many tasks associated with this data management and integration process, freeing up resources for your team.

You should pay careful attention to the ETL solution you use so you can improve business decisions. Since ETL is responsible for the bulk of the in-between work, choosing a subpar tool or developing a poor [ETL process](https://www.integrate.io/blog/etl-data-warehousing-explained-etl-tool-basics/) can break your entire warehouse. You want optimal speeds, high availability, good visualization, and the ability to build easy, replicable, and consistent data pipelines between all your existing architecture and your new warehouse.

This is where ETL tools like Integrate.io are valuable. Integrate.io creates hyper-visualized data pipelines between all your valuable tech architecture while cleaning and nominalizing that data for compliance and ease of use.

Remember, a good ETL process can mean the difference between a slow, painful-to-use data warehouse and a simple, functional warehouse that's valuable throughout every layer of your organization.

ETL will likely be the go-to for pulling data from systems into your warehouse. Its counterpart [Extract, Load, Transfer](https://dashboard.integrate.io/auth/signup?tool=CDC&_gl=1) (ELT) negatively impacts the performance of most custom-built warehouses since data is loaded directly into the warehouse before data organization and cleansing occur. However, there might be other data integration use cases that suit the ELT process. Integrate.io not only executes ETL but can handle ELT, Reverse ETL, and Change Data Capture (CDC), as well as provide data observability and data warehouse insights.

**Related Reading**

### 5. Online Analytic Processing (OLAP) Cube

OLAP (Online Analytical Processing) cubes are commonly used in the data warehousing process to enable faster, more efficient analysis of large amounts of data. OLAP cubes are based on multidimensional databases that store summarized data and allow users to quickly analyze information from different dimensions.

Here's how an OLAP cube fits into the data warehouse design:

* OLAP cubes are designed to store pre-aggregated data that has been processed from various sources in a data warehouse. The data is organized into a multi-dimensional structure that enables users to view and analyze it from different perspectives.
* OLAP cubes are created using a process called cube processing, which involves aggregating and storing data in a way that enables fast retrieval and analysis. Cube processing can be performed on a regular basis to ensure that the data is up-to-date and accurate.
* OLAP cubes enable users to perform complex analytical queries on large volumes of data in real-time, making it easier to identify trends, patterns, and anomalies. Users can also slice and dice data in different ways to gain deeper insights into their business operations.
* OLAP cubes support drill-down and roll-up operations, which allow users to navigate through different levels of data granularity. Users can drill down to the lowest level of detail to view individual transactions or roll up to higher levels of aggregation to view summary data.
* OLAP cubes can be accessed using a variety of tools, including spreadsheets, reporting tools, and business intelligence platforms. Users can create reports and dashboards that display the data in a way that is meaningful to them.

You'll likely need to address OLAP cubes if you're designing your entire database from scratch, or if you're maintaining your own OLAP cube — which typically requires specialized personnel.

### 6. Data Warehouse Design: Creating the Front End

So far, this guide has only covered back-end processes. There needs to be front-end visualization, so users can immediately understand and apply the results of data queries.

That's the job of your front end. There are plenty of tools on the market that help with visualization. BI tools like Tableau (or PowerBI for those using BigQuery) are great for visualization. You can also develop a custom solution — though that's a significant undertaking.

Most small-to-medium-sized businesses lean on established BI kits like those mentioned above. But, some businesses may need to develop their own BI tools to meet ad-hoc analytic needs. For example, a Sales Ops manager at a large company may need a specific BI tool for territory strategies. This tool would probably be custom-developed given the scope of the company’s sales objectives.

You should pay keen attention to reporting during this stage. How often does reporting need to be done? Do you need each person to create their own reports? Questions like these should guide you to a BI toolkit that fits your unique requirements.

**Pro-tip**: Keep it simple. Your employees don't care about most of the fancy features or deep complexities. They just want something that works for them and makes their lives easier.

### 7. Optimizing Queries

Optimizing queries is a critical part of data warehouse design. One of the primary goals of building a data warehouse is to provide fast and efficient access to data for decision-making. During the design process, data architects need to consider the types of queries that users will be running and design the data warehouse schema and indexing accordingly.

Optimizing your queries is a complex process that's hyper-unique to your specific needs. But there are some general rules of thumb.

We **heavily recommend** the following during database design:

* **Ensure your production, testing, and development environments have mirrored resources.** This mirroring prevents the server from hanging when you push projects from one environment to the next.
* **Try to minimize data retrieval.**Don't run SELECT on the whole database if you only need a column of results. Instead, run your SELECT query by targeting specific columns. This is especially important if you're paying for your query power separately.
* **Understand the limitations of your OLAP vendor.**BigQuery uses a hybrid SQL language, and RedShift is built on top of a Postgre fork. Knowing the little nuances baked into your vendor can help you maximize workflows and speed up queries.

### 8. Establishing a Rollout Plan

Once you're ready to launch your warehouse, it's time to start thinking about education, training, and use cases. Most of the time, it will be a week or two before your end-users start seeing any functionality from that warehouse (at least at scale). But they should be adequately trained in its use before the rollout is completed.

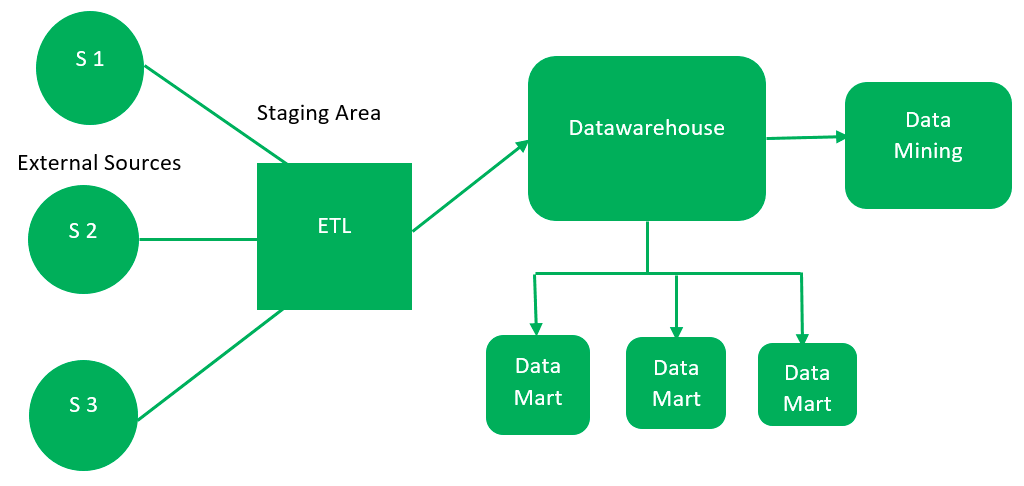
A rollout plan typically includes the following steps:

1. **Identifying the target audience:**This involves determining which groups or individuals within the organization will benefit from using the data warehouse.
2. **Determining the data requirements**: This involves identifying the types of data that the target audience needs access to and ensuring that this data is available within the data warehouse.
3. **Developing user-friendly interfaces:** This involves creating user interfaces that are intuitive and easy to use, and that provide users with the ability to interact with the data in meaningful ways.
4. **Testing and refining:** This involves conducting user testing to ensure that the data warehouse meets the needs of its users, and making adjustments as necessary.
5. **Training users:** This involves providing training and support to users to help them understand how to use the data warehouse effectively.
6. **Deploying the data warehouse:**This involves introducing the data warehouse to its intended users, and ensuring that the rollout process goes smoothly.

# Data Warehouse Architecture

A **data-warehouse** is a heterogeneous collection of different data sources organised under a unified schema. There are 2 approaches for constructing data-warehouse: Top-down approach and Bottom-up approach are explained as below.

**1. Top-down approach:**



1. **External Sources –**   
   External source is a source from where data is collected irrespective of the type of data. Data can be structured, semi structured and unstructured as well.
2. **Stage Area –**   
   Since the data, extracted from the external sources does not follow a particular format, so there is a need to validate this data to load into datawarehouse. For this purpose, it is recommended to use **ETL** tool.
   * **E(Extracted):** Data is extracted from External data source.
   * **T(Transform):** Data is transformed into the standard format.
   * **L(Load):** Data is loaded into datawarehouse after transforming it into the standard format.
3. **Data-warehouse –**   
   After cleansing of data, it is stored in the datawarehouse as central repository. It actually stores the meta data and the actual data gets stored in the data marts. **Note** that datawarehouse stores the data in its purest form in this top-down approach.
4. **Data Marts –**   
   Data mart is also a part of storage component. It stores the information of a particular function of an organisation which is handled by single authority. There can be as many number of data marts in an organisation depending upon the functions. We can also say that data mart contains subset of the data stored in datawarehouse.
5. **Data Mining –**   
   The practice of analysing the big data present in datawarehouse is data mining. It is used to find the hidden patterns that are present in the database or in datawarehouse with the help of algorithm of data mining.

This approach is defined by **Inmon** as – datawarehouse as a central repository for the complete organisation and data marts are created from it after the complete datawarehouse has been created. 

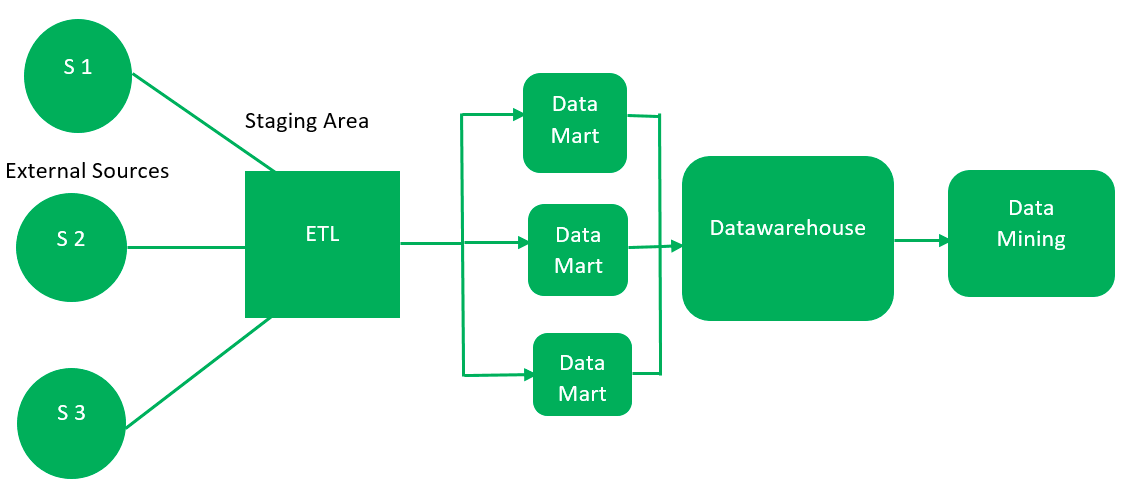
**Advantages of Top-Down Approach –**

1. Since the data marts are created from the datawarehouse, provides consistent dimensional view of data marts.
2. Also, this model is considered as the strongest model for business changes. That’s why, big organisations prefer to follow this approach.
3. Creating data mart from datawarehouse is easy.

**Disadvantages of Top-Down Approach –**

1. The cost, time taken in designing and its maintenance is very high.

**2. Bottom-up approach:**



1. First, the data is extracted from external sources (same as happens in top-down approach).
2. Then, the data go through the staging area (as explained above) and loaded into data marts instead of datawarehouse. The data marts are created first and provide reporting capability. It addresses a single business area.
3. These data marts are then integrated into datawarehouse.

This approach is given by **Kinball** as – data marts are created first and provides a thin view for analyses and datawarehouse is created after complete data marts have been created.

**Advantages of Bottom-Up Approach –**

1. As the data marts are created first, so the reports are quickly generated.
2. We can accommodate more number of data marts here and in this way datawarehouse can be extended.
3. Also, the cost and time taken in designing this model is low comparatively.

**Disadvantage of Bottom-Up Approach –**

1. This model is not strong as top-down approach as dimensional view of data marts is not consistent as it is in above approach.

## Data source types

Though the diversity of content, format, and location for data is only increasing with contributions from technologies such as [IoT](https://www.talend.com/resources/internet-of-things/) and the adoption of [big data](https://www.talend.com/resources/future-big-data/) methodologies, it remains possible to classify most data sources into two broad categories: machine data sources and file date sources.

Though both share the same basic purpose — pointing to the data’s location and describing similar connection characteristics — machine and file data sources are stored, accessed, and used in different ways.

### **Machine data sources**

Machine data sources have names defined by users, must reside on the machine that is ingesting data, and cannot be easily shared. Like other data sources, machine data sources provide all the information necessary to connect to data, such as relevant software drivers and a driver manager, but users need only ever refer to the DSN as shorthand to invoke the connection or query the data.

The connection information is stored in environment variables, database configuration options, or a location internal to the machine or application being used. An Oracle data source, for example, will contain a server location for accessing the remote DBMS, information about which drivers to use, the driver engine, and any other relevant parts of a typical connection string, such as system and user IDs and authentication.

### **File data sources**

File data sources contain all of the connection information inside a single, shareable, computer file (typically with a .dsn extension). Users do not decide which name is assigned to file data sources, as these sources are not registered to individual applications, systems, or users, and in fact do not have a DSN like that of machine data sources. Each file stores a connection string for a single data source.

File data sources, unlike machine sources, are editable and copyable like any other computer file. This allows users and systems to  share a common connection (by moving the data source between individual machines or servers), and for the streamlining of data connection processes (for example by keeping a source file on a shared resource so it may be used simultaneously by multiple applications and users).

It is important to note that ‘unshareable’ .dsn files also exist. These are the same type of file as described above, but they exist on a single machine and cannot be moved or copied. These files point directly to machine data sources. This means that unshareable file data sources are wrappers for machine data sources, serving as a proxy for applications which expect only files but also need to connect to machine data.

## What Does Data Loading Mean?

Data loading is the process of copying and loading data or data sets from a source file, folder or application to a database or similar application. It is usually implemented by copying digital data from a source and pasting or loading the data to a data storage or processing utility.

**Types of Data Loading**

Soon after your departure from the extraction phase, you will be faced with the decision of which loading process that you would like to deploy. The **data** **loading** process is the physical movement of the data from the computer systems storing the source database(s) to that which will store the data warehouse database. The entire process of transferring data to a data warehouse repository is referred to in the following ways:

* **Full Load:**This is where *all*of your data is selected, moved in bulk, and then replaced by new data. Although it is not as complex to navigate through, loading time is much slower. With the overwhelming amount of data being moved at once, it is much easier for data to get lost within the big move.
* **Incremental Load:** This is where you are moving new data in intervals. Due to its intricate nature, delivery time is much faster than its counterpart. However, this speed comes at a cost. Incremental loads are more likely to encounter problems due to the nature of having to manage them as individual batches rather than one big group. **Incremental Load** Periodically applies ongoing changes as per the requirement. After the data is loaded into the data warehouse database, verify the referential integrity between the dimensions and the fact tables to ensure that all records belong to the appropriate records in the other tables. The DBA must verify that each record in the fact table is related to one record in each dimension table that will be used in combination with that fact table.
* **Initial Load**: For the very first time loading all the data warehouse tables.
* **Full Refresh**: Deleting the contents of a table and reloading it with fresh data.

UNIT 5

**Data selection** is defined as the process of determining the appropriate **data type** and **source**, as well as suitable **instruments** to collect data. Data selection precedes the actual practice of data collection.

## DATA COLLECTION METHODS USED IN BUSINESS ANALYTICS

### **1. Surveys**

Surveys are physical or digital questionnaires that gather both qualitative and quantitative data from subjects. One situation in which you might conduct a survey is gathering attendee feedback after an event. This can provide a sense of what attendees enjoyed, what they wish was different, and areas in which you can improve or save money during your next event for a similar audience.

While physical copies of surveys can be sent out to participants, online surveys present the opportunity for distribution at scale. They can also be inexpensive; running a survey can cost nothing if you use a free tool. If you wish to target a specific group of people, partnering with a market research firm to get the survey in front of that demographic may be worth the money.

Something to watch out for when crafting and running surveys is the effect of bias, including:

* **Collection bias**: It can be easy to accidentally write survey questions with a biased lean. Watch out for this when creating questions to ensure your subjects answer honestly and aren’t swayed by your wording.
* **Subject bias**: Because your subjects know their responses will be read by you, their answers may be biased toward what seems socially acceptable. For this reason, consider pairing survey data with behavioral data from other collection methods to get the full picture.

Related: [3 Examples of Bad Survey Questions & How to Fix Them](https://online.hbs.edu/blog/post/3-survey-question-mistakes-and-how-to-fix-them)

### **2. Transactional Tracking**

Each time your customers make a purchase, tracking that data can allow you to make decisions about targeted marketing efforts and understand your customer base better.

Often, e-commerce and point-of-sale platforms allow you to store data as soon as it’s generated, making this a seamless data collection method that can pay off in the form of customer insights.

### **3. Interviews and Focus Groups**

Interviews and focus groups consist of talking to subjects face-to-face about a specific topic or issue. Interviews tend to be one-on-one, and focus groups are typically made up of several people. You can use both to gather qualitative and quantitative data.

Through interviews and focus groups, you can gather feedback from people in your target audience about new product features. Seeing them interact with your product in real-time and recording their reactions and responses to questions can provide valuable data about which product features to pursue.

As is the case with surveys, these collection methods allow you to ask subjects anything you want about their opinions, motivations, and feelings regarding your product or brand. It also introduces the potential for bias. Aim to craft questions that don’t lead them in one particular direction.

One downside of interviewing and conducting focus groups is they can be time-consuming and expensive. If you plan to conduct them yourself, it can be a lengthy process. To avoid this, you can hire a market research facilitator to organize and conduct interviews on your behalf.

### **4. Observation**

Observing people interacting with your website or product can be useful for data collection because of the candor it offers. If your user experience is confusing or difficult, you can witness it in real-time.

Yet, setting up observation sessions can be difficult. You can use a third-party tool to record users’ journeys through your site or observe a user’s interaction with a beta version of your site or product.

While less accessible than other data collection methods, observations enable you to see firsthand how users interact with your product or site. You can leverage the qualitative and quantitative data gleaned from this to make improvements and double down on points of success.

### **5. Online Tracking**

To gather behavioral data, you can implement pixels and cookies. These are both tools that track users’ online behavior across websites and provide insight into what content they’re interested in and typically engage with.

You can also track users’ behavior on your company’s website, including which parts are of the highest interest, whether users are confused when using it, and how long they spend on product pages. This can enable you to improve the website’s design and help users navigate to their destination.

Inserting a pixel is often free and relatively easy to set up. Implementing cookies may come with a fee but could be worth it for the quality of data you’ll receive. Once pixels and cookies are set, they gather data on their own and don’t need much maintenance, if any.

It’s important to note: Tracking online behavior can have legal and ethical privacy implications. Before tracking users’ online behavior, ensure you’re in compliance with local and industry [data privacy standards](https://online.hbs.edu/blog/post/data-privacy).

### **6. Forms**

Online forms are beneficial for gathering qualitative data about users, specifically demographic data or contact information. They’re relatively inexpensive and simple to set up, and you can use them to gate content or registrations, such as webinars and email newsletters.

You can then use this data to contact people who may be interested in your product, build out demographic profiles of existing customers, and in remarketing efforts, such as email workflows and content recommendations.

### **7. Social Media Monitoring**

Monitoring your company’s social media channels for follower engagement is an accessible way to track data about your audience’s interests and motivations. Many social media platforms have analytics built in, but there are also third-party social platforms that give more detailed, organized insights pulled from multiple channels.

You can use data collected from social media to determine which issues are most important to your followers. For instance, you may notice that the number of engagements dramatically increases when your company posts about its sustainability efforts.

## What is data cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

### **Step 1: Remove duplicate or irrelevant observations**

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process. Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyze. For example, if you want to analyze data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient and minimize distraction from your primary target—as well as creating a more manageable and more performant dataset.

### **Step 2: Fix structural errors**

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabeled categories or classes. For example, you may find “N/A” and “Not Applicable” both appear, but they should be analyzed as the same category.

### **Step 3: Filter unwanted outliers**

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with. However, sometimes it is the appearance of an outlier that will prove a theory you are working on. Remember: just because an outlier exists, doesn’t mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

### **Step 4: Handle missing data**

You can’t ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

1. As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
2. As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
3. As a third option, you might alter the way the data is used to effectively navigate null values.

### **Step 5: Validate and QA**

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation:

* Does the data make sense?
* Does the data follow the appropriate rules for its field?
* Does it prove or disprove your working theory, or bring any insight to light?
* Can you find trends in the data to help you form your next theory?
* If not, is that because of a data quality issue?

## What Are the OUTPUTS of Data Mining?

Since we live and work in a data-centric world, it’s essential to get as many advantages as possible. Data mining provides us with the means of resolving problems and issues in this challenging information age. Data mining benefits include:

* It helps companies gather reliable information
* It’s an efficient, cost-effective solution compared to other data applications
* It helps businesses make profitable production and operational adjustments
* Data mining uses both new and legacy systems
* It helps businesses make informed decisions
* It helps detect credit risks and fraud
* It helps data scientists easily analyze enormous amounts of data quickly
* Data scientists can use the information to detect fraud, build risk models, and improve product safety
* It helps data scientists quickly initiate automated predictions of behaviors and trends and discover hidden patterns.

**10 Data Mining Techniques**

## 1. Clustering

Clustering is a technique used to represent data visually — such as in graphs that show buying trends or sales demographics for a particular product.

### **What Is Clustering in Data Mining?**

Clustering refers to the process of grouping a series of different data points based on their characteristics. By doing so, data miners can seamlessly divide the data into subsets, allowing for more informed decisions in terms of broad demographics (such as consumers or users) and their respective behaviors.

### **Methods for Data Clustering**

* **Partitioning method:**This involves dividing a data set into a group of specific clusters  for evaluation based on the criteria of each individual cluster. In this method, data points belong to just one group or cluster.
* **Hierarchical method:** With the hierarchical method, data points are a single cluster, which are grouped based on similarities. These newly created clusters can then be analyzed separately from each other.
* **Density-based method:** A machine learning method where data points plotted together are further analyzed, but data points by themselves are labeled “noise” and discarded.
* **Grid-based method:** This involves dividing data into cells on a grid, which then can be clustered by individual cells rather than by the entire database. As a result, grid-based clustering hase a fast processing time.
* **Model-based method:**In this method, models are created for each data cluster to locate the best data to fit that particular model.

### **Examples of Clustering in Business**

Clustering helps businesses manage their data more effectively. For example, retailers can use clustering models to determine which customers buy particular products, on which days, and with what frequency. This can help retailers target products and services to customers in a specific demographic or region.

Clustering can help grocery stores group products by a variety of characteristics (brand, size, cost, flavor, etc.) and better understand their sales tendencies. It can also help car insurance companies that want to identify a set of customers who typically have high annual claims in order to price policies more effectively. In addition, banks and financial institutions might use clustering to better understand how customers use in-person versus virtual services to better plan branch hours and staffing.

## 2. Association

Association rules are used to find correlations, or associations, between points in a data set.

### **What Is Association in Data Mining?**

Data miners use association to discover unique or interesting relationships between variables in databases. Association is often employed to help companies determine marketing research and strategy.

### **Methods for Data Mining Association**

Two primary approaches using association in data mining are the single-dimensional and multi-dimensional methods.

* **Single-dimensional association:** This involves looking for one repeating instance of a data point or attribute. For instance, a retailer might search its database for the instances a particular product was purchased.
* **Multi-dimensional association:** This involves looking for more than one data point in a data set. That same retailer might want to know more information than what a customer purchased — such as their age, method of purchase (cash or credit card), or age.

### **Examples of Association in Business**

The analysis of impromptu shopping behavior is an example of association — that is, retailers notice in data studies that parents shopping for childcare supplies are more likely to purchase specialty food or beverage items for themselves during the same trip. These purchases can be analyzed through statistical association.

Association analysis carries many other uses in business. For retailers, it’s particularly helpful in making purchasing suggestions. For example, if a customer buys a smartphone, tablet, or video game device, association analysis can recommend related items like cables, applicable software, and protective cases.

Additionally, association is used by the government to employ census data and plan for public services; it is also used by doctors to diagnose various illnesses and conditions more effectively.

## 3. Data Cleaning

Data cleaning is the process of preparing data to be mined.

### **What Is Data Cleaning in Data Mining?**

Data cleaning involves organizing data, eliminating duplicate or corrupted data, and filling in any null values. When this process is complete, the most useful information can be harvested for analysis.

### **Methods for Data Cleaning**

* **Verifying the data:** This involves checking that each data point in the data set is in the proper format (e.g, telephone numbers, social security numbers).
* **Converting data types**: This ensures data is uniform across the data set. For instance, numeric variables only contain numbers, while string variables can contain letters, numbers, and characters.
* **Removing irrelevant data:** This clears useless or inapplicable data so full emphasis can be placed on necessary data points.
* **Eliminating duplicate data points:** This helps speed up the mining process by boosting efficiency and reducing errors.
* **Removing errors:** This eliminates typing mistakes, spelling errors, and input errors that could negatively affect analysis outcomes.
* **Completing missing values:** This provides an estimated value for all data and reduces missing values, which can lead to skewed or incorrect results.

### **Examples of Data Cleaning in Business**

According to Experian, [95 percent of businesses say they have been impacted by poor data quality](https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm#tab-6). Working with incorrect data wastes time and resources, increases analysis costs (because models need to be repeated), and often leads to faulty analytics.

Ultimately, no matter how great their models or algorithms are, businesses suffer when their data is incorrect, incomplete, or corrupted.

## 4. Data Visualization

Data visualization is the translation of data into graphic form to illustrate its meaning to business stakeholders.

### **What Is Data Visualization in Data Mining?**

Data can be presented in visual ways through charts, graphs, maps, diagrams, and more. This is a primary way in which data scientists display their findings.

### **Methods for Data Visualization**

Many methods exist for representing data visually. Here are a few:

* **Comparison charts:** Charts and tables express relationships in the data, such as monthly product sales over a one-year period.
* **Maps:**Data maps are used to visualize data pertaining to specific geographic locations. Through maps, data can be used to show population density and changes; compare populations of neighboring states, counties, and countries; detect how populations are spread over geographic regions; and compare characteristics in one region to those in other regions.
* **Heat maps:**This is a popular visualization technique that represents data through different colors and shading to indicate patterns and ranges in the data. It can be used to track everything from a region’s temperature changes to its food and pop culture trends.
* **Density plots:** These visualizations track data over a period of time, creating what can look like a mountain range. Density plots make it easy to represent occurrences of single events over time (e.g., month, year, decade).
* **Histograms:** These are similar to density plots but are represented by bars on a graph instead of a linear form.
* **Network diagrams:** These diagrams show how data points relate to each other by using a series of lines (or links) to connect objects together.
* **Scatter plots:** These graphs represent data point relationships on a two-variable axis. Scatter plots can be used to compare unique variables such as a country’s life expectancy or the amount of money spent on healthcare annually.
* **Word clouds:** These graphics are used to highlight specific word or phrase instances appearing in a body of text; the larger the word’s size in the cloud, the more frequent its use.

### **Examples of Data Visualization in Business**

Representing data visually is an important skill because it makes data readily understandable to executives, clients, and customers. [According to Markets and Markets](https://www.marketsandmarkets.com/Market-Reports/data-visualization-tools-market-94728248.html), the market size for global data visualization tools is expected to nearly double (to $10.2 billion) by 2026.

Companies can make faster, more informed decisions when presented with data that is easy to understand and interpret. Today, this is typically accomplished through effective, visually accessible mediums such as graphs, 3D models, and even augmented reality. As a result, it’s a good idea for aspiring data professionals to consider learning such skills through a [data science and visualization bootcamp](https://bootcamp.pe.gatech.edu/data/).

## 5. Classification

Classification is a fundamental technique in data mining and can be applied to nearly every industry. It is a process in which data points from large data sets are assigned to categories based on how they’re being used.

### **What Is Classification in Data Mining?**

In data mining, classification is considered to be a form of clustering — that is, it is useful for extracting comparable points of data for comparative analysis. Classification is also used to designate broad groups within a demographic, target audience, or user base through which businesses can gain stronger insights.

### **Methods for Data Mining Classification**

* **Logistic regression:** This algorithm attempts to show the probability of a specific outcome within two possible results. For example, an email service can use logistic regression to predict whether or not an email is spam.
* **Decision trees:** Once data is classified, follow-up questions can be asked, and the results diagrammed into a chart called a decision tree. For example, if a computer company wants to predict the likelihood of laptop purchases, it may ask, *Is the potential buyer a student?* The data is classified into “Yes” and “No” decision trees, with other questions to be asked afterward in a similar fashion.
* **K-nearest neighbors (KNN):** This is an algorithm that tries to identify an unknown object by comparing it to others. For instance, grocery chains might use the K-nearest neighbors algorithm to decide whether to include a sushi or hot meals station in their new store layout based on consumer habits in the local marketplace.
* **Naive Bayes:** Based on the Bayes Theorem of Probability, this algorithm uses historical data to predict whether similar events will occur based on a different set of data.
* **Support Vector Machine (SVM):** This machine learning algorithm is often used to define the line that best divides a data set into two classes. An SVM can help classify images and is used in facial and handwriting recognition software.

### **Examples of Classification in Business**

Financial institutions classify consumers based on many variables to market new loans or project credit card risks. Meanwhile, weather apps classify data to project snowfall totals and other similar figures. Grocery stores also use classification to group products by the consumers who buy them, helping forecast buying patterns.

## 6. Machine Learning

Machine learning is the process by which computers use algorithms to learn on their own. An increasingly relevant part of modern technology, machine learning makes computers “smarter” by teaching them how to perform tasks based on the data they have gathered.

### **What Is Machine Learning in Data Mining?**

In data mining, machine learning’s applications are vast. Machine learning and data mining fall under the umbrella of data science but aren’t interchangeable terms. For instance, computers perform data mining as part of their machine learning functions.

### **Methods for Machine Learning**

* **Supervised learning:** In this method, algorithms train machines to learn using pre-labeled data with correct values, which the machines then classify on their own. It’s called supervised because the process trains (or “supervises”) computers to classify data and predict outcomes. Supervised machine learning is used in data mining classification.
* **Unsupervised learning:** When computers handle unlabeled data, they engage in unsupervised learning. In this case, the computer classifies the data itself and then looks for patterns on its own. Unsupervised models are used to perform clustering and association.
* **Semi-supervised learning:** Semi-supervised learning uses a combination of labeled and unlabeled data, making it a hybrid of the above models.
* **Reinforcement learning:** This is a more layered process in which computers learn to make decisions based on examining data in a specific environment. For example, a computer might learn to play chess by examining data from thousands of games played online.

### **Examples of Machine Learning in Business**

With machine learning, companies can use computers to quickly identify all sorts of data patterns (in sales, product usage, buying habits, etc.) and develop business plans using those insights. This is a growing need in many industries.

[According to a MicroStrategy survey](https://www3.microstrategy.com/getmedia/50ea9c13-feb7-4b9a-b976-8b04ca39abb2/Global-State-of-Enterprise-Analytics-Report-MicroStrategy_2018), 18 percent of analytics professionals said machine learning and AI will have the most significant impact on their strategies over the next five years. Learning more [advanced topics like machine learning](https://bootcamp.pe.gatech.edu/data/curriculum/) is thus becoming imperative for data scientists.

**Request information about the Data Science and Analytics Boot Camp to learn more about machine learning and other topics covered in the program curriculum.**

## 7. Neural Networks

Computers process large amounts of data much faster than human brains but don’t yet have the capacity to apply common sense and imagination in working with the data. Neural networks are one way to help computers reason more like humans.

### **What Are Neural Networks in Data Mining?**

Artificial neural networks attempt to digitally mimic the way the human brain operates. Neural networks combine many computer processors (similar to the way the brain uses neurons) to process data, make decisions, and learn as a human would — or at least as closely as possible.

### **Neural Network Methods**

Neural networks consist of three main layers: input, “hidden,” and output. Data enters through the input layer, is processed in the hidden layer, and is resolved in the output layer where any relevant action based on the data is then taken. The hidden layer can consist of many processing layers, depending on the amount of data being used and learning taking place.

Supervised and unsupervised learning also apply to neural networks; neural networks use these types of algorithms to “train” themselves to function in ways similar to the human brain.

### **Examples of Neural Networks in Business**

Neural networks have a wide range of applications. They can help businesses predict consumer buying patterns and focus marketing campaigns on specific demographics. They can also help retailers make accurate sales forecasts and understand how to use dynamic pricing. Furthermore, they help to improve diagnostic and treatment methods in healthcare, improving care and performance.

## 8. Outlier Detection

Outlier detection is a key component of maintaining safe databases. Companies use it to test for fraudulent transactions, such as abnormal credit card usage that might suggest theft.

### **What Is Outlier Detection in Data Mining?**

While other data mining methods seek to identify patterns and trends, outlier detection looks for the unique: the data point or points that differ from the rest or diverge from the overall sample. Outlier detection finds errors, such as data that was input incorrectly or extracted from the wrong sample. Natural data deviations can be instructive as well.

### **Methods for Outlier Detection**

* **Numeric outlier:** Outliers are detected based on the Interquartile Range, or the middle 50 percent of values. Data points outside that range are considered outliers.
* **Z-score:** The Z-Score denotes how many standard deviations a data point is from the sample’s mean. This is also known as extreme value analysis.
* **DBSCAN:** This stands for “density-based spatial clustering of applications with noise” and is a method that defines data as core points, border points, and noise points, which are the outliers.
* **Isolation forest:** This method isolates anomalies in large sets of data (the forest) with an algorithm that searches for those anomalies instead of profiling normal data points.

### **Examples of Outlier Detection in Business**

Almost every business can benefit from understanding anomalies in their production or distribution lines and how to fix them. Retailers can use outlier detection to learn why their stores witness an odd increase in purchases, such as snow shovels being bought in the summer, and how to respond to such findings.

Generally, outlier detection is employed to enhance logistics, instill a culture of preemptive damage control, and create a smoother environment for customers, users, and other key groups.

## 9. Prediction

Predictive modeling seeks to turn data into a projection of future action or behavior. These models examine data sets to find patterns and trends, then calculate the probabilities of a future outcome.

### **What Is Prediction in Data Mining?**

Predictive modeling is among the most common uses of data mining and works best with large data sets that represent a broad sample size.

### **Methods for Prediction**

Predictive modeling uses some of the same techniques and terminology as other data mining processes. Here are four examples:

* **Forecast modeling:** This is a common technique in which the computer answers a question (for instance, *How much milk should a store have in stock on Monday?*) by analyzing historical data.
* **Classification modeling:** Classification places data into groups where it can be used to answer direct questions.
* **Cluster modeling:** By clustering data into groups with shared characteristics, a predictive model can be used to study those data sets and make decisions.
* **Time series modeling:** This model analyzes data based on when the data was input. A study of sales trends over a year is an example of time series modeling.

## Data Mining Best Practices

Regardless of which specific technique you use, here are key data mining best practices to help you maximize the value of your process. They can be applied to any of the 15 aforementioned techniques.

* **Preserve the data.** This should be obvious. Data must be maintained militantly, and it must not be archived, deleted, or overwritten once processed. You went through a lot of trouble to get that data prepared for generating insight, now vigilance must be applied to maintenance.
* **Have a clear idea of what you want out of the data.** This predicates your sampling and modeling efforts, never mind your searches. The first question is what do you want out of this strategy, such as knowing customer behaviors.
* **Have a clear modeling technique.** Be prepared to go through many modeling prototypes as you narrow down your data ranges and the questions you are asking. If you aren’t getting the answers you want, ask them a different way.
* **Clearly identify the business problems.** Be specific, don’t just say sell more stuff. Identify fine grain issues, determine where they occur in the sale, pre- or post-, and what the problem actually is.
* **Look at post-sale as well.** Many mining efforts focus on getting the sale but what happens after the sale — returns, cancellations, refunds, exchanges, rebates, write-offs – are equally important because they are a portent to future sales. They help identifying customers who will be more or less likely to make future purchases.
* **Deploy on the front lines.** It’s too easy leave the data mining inside the corporate firewall, since that’s where the warehouse is located and all data comes in. But preparatory work on the data before it is sent in can be done in remote sites, as can application of sales, marketing, and customer relations models.

## Myth #1: Data mining is an extremely complicated process and difficult to understand.

Algorithms behind data mining may be complex, but with the right tools, data mining can be easy to use and can change the way you run your business. Data mining tools allow you to easily see and understand your data with simple to understand graphs, queries, and visualizations gives you insight as to how your business is performing. You can then identify problems and potential issues and make analytics-based decisions to improve upon your inefficiencies.

Data mining tools are not as complex or hard to use as people think they may be. They are designed to be easy to understand so that businesses are able to interpret the information that is produced. Data mining is extremely advantageous and should not be intimidating to those who are considering utilizing it.

## Myth #2: Data mining is another trend that will soon die out, allowing us to return to standard business practice.

Quantitative practices have been employed by businesses for quite some time. Data mining is just a more developed practice that has come about since the beginning of the 20th century. Data is everywhere and the size of some databases are tremendous, making it extremely difficult for discovery to be done manually. With the easy-to-use functionality, cost and time reduction benefits, and the ability to conduct an analysis of your company’s performance within a quick to deploy, and easy to understand solution makes it hard to believe that something so advantageous and beneficial will ever fade away. If anything, data mining will be an everlasting and growing tool that will help us for years to come.

## Myth #3: Data mining techniques are so advanced that they can replace domain knowledge.

Expertise and experience of the business and its markets cannot be replaced by data mining techniques. Knowledge about the new analytical methods that arise is important but, without knowledge of the business and its markets, there is of no use to these methods. Therefore, it is critical to have an understanding of both.

If you are conducting an analysis of a company’s data, it is important to have someone who is an expert in the field to make sense of the information produced and vice versa. If there is someone with knowledge about the business and its markets, it’s important to have an expert in data mining conduct analysis with tools and modeling to help improve their business knowledge. Data mining essentially cannot exist without domain knowledge.

## Myth #4: Only big databases are worth mining.

Although data mining is more commonly used for analyzing big data sets, it can be used for any size. Just about any amount of data can produce valuable information that can be used for businesses to detect issues and potential issues. Even these sample size datasets allow for businesses to find inefficiencies in which they can proactively or plan to improve upon. It may be more beneficial to pull certain data from a large data warehouse to conduct an analysis versus the entire database itself. You just need to know which data you want to analyze to produce valuable outcomes and conclusions.

## Myth #5: Data mining is useful only in certain industries.

Although data mining may be most commonly used within highly data-focused and innovation-driven industries, it is a tool that can be used in any industry. There will always be an instance in which data mining might not be worth the return on investment. But just like how the size of the database does not matter, neither does the industry. There is value that can come from any type of data you analyze.

**DATAMINING MISTAKES**

* **Skipping data quality checks:** Most data miners think developing predictive models is more fun than reviewing data for quality problems. But if you fail to detect and correct data quality problems, you could end up with worthless predictions.
* **Missing the point:** You've discovered something fascinating! That's nice, but if it isn't also relevant to the business problem you set out to solve, well, it isn't relevant at all. Get back on track.
* **Believing that a pattern in the data proves a cause-and-effect relationship:** You explore a dataset and notice that when Variable A increases, Variable B increases, too. This could occur because Variable A influences Variable B, or because Variable B influences Variable A. On the other hand, it could be that both are influenced by some other variable that you have not considered. Or it could be a one-time coincidence. Who can say?
* **Stretching conclusions too far:** Don't presume that the relationships you observe in data will recur in different circumstances. If your data was collected in a cool environment, don't assume that things will work the same way in a hot factory setting.
* **Betting on results that don't make sense:** Data mining methods are informal and not usually backed up by scientific method and theory, so your results had better at least make business sense. If there's no common sense explanation for the results you present, your executive management probably won't take it seriously, and they shouldn't.
* **Falling in love with a particular modeling method:** There is no single type of data mining model that fits every situation.
* **Putting a model into production without adequate testing:** Don't bet your business on a predictive model until you have tested it with holdout data and on a small scale in the field.
* **Ignoring results you don't like:** If you ignore your data now, it will come back one day and say, "I told you so."
* **Using data mining to address every data analysis need:** Data mining has tremendous value, yet some applications still call for rigorous data collection methods, formal statistical analysis, and scientific method.
* **Presuming that traditional data analysis techniques no longer matter:** Refer to the previous bullet.

**UNIT 6**

# **What is Data Visualization?**

Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.

Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans.

Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.

In the world of Big Data, the data visualization tools and technologies are required to analyze vast amounts of information.

Data visualizations are common in your everyday life, but they always appear in the form of graphs and charts. The combination of multiple visualizations and bits of information are still referred to as Infographics.

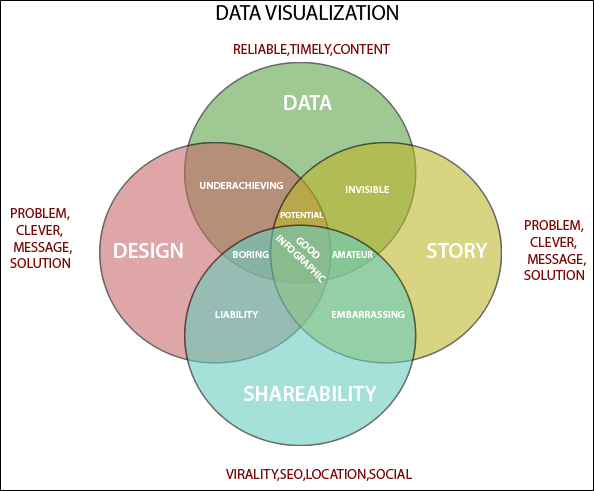
Data visualizations are used to discover unknown facts and trends. You can see visualizations in the form of line charts to display change over time. Bar and column charts are useful for observing relationships and making comparisons. A pie chart is a great way to show parts-of-a-whole. And maps are the best way to share geographical data visually.

Today's data visualization tools go beyond the charts and graphs used in the Microsoft Excel spreadsheet, which displays the data in more sophisticated ways such as dials and gauges, geographic maps, heat maps, pie chart, and fever chart.

What makes Data Visualization Effective?

Effective data visualization are created by communication, data science, and design collide. Data visualizations did right key insights into complicated data sets into meaningful and natural.

American statistician and Yale professor **Edward Tufte** believe useful data visualizations consist of ?complex ideas communicated with clarity, precision, and efficiency.



To craft an effective data visualization, you need to start with clean data that is well-sourced and complete. After the data is ready to visualize, you need to pick the right chart.

After you have decided the chart type, you need to design and customize your visualization to your liking. Simplicity is essential - you don't want to add any elements that distract from the data.

## Importance of Data Visualization

Data visualization is important because of the processing of information in human brains. Using graphs and charts to visualize a large amount of the complex data sets is more comfortable in comparison to studying the spreadsheet and reports.

Data visualization is an easy and quick way to convey concepts universally. You can experiment with a different outline by making a slight adjustment.

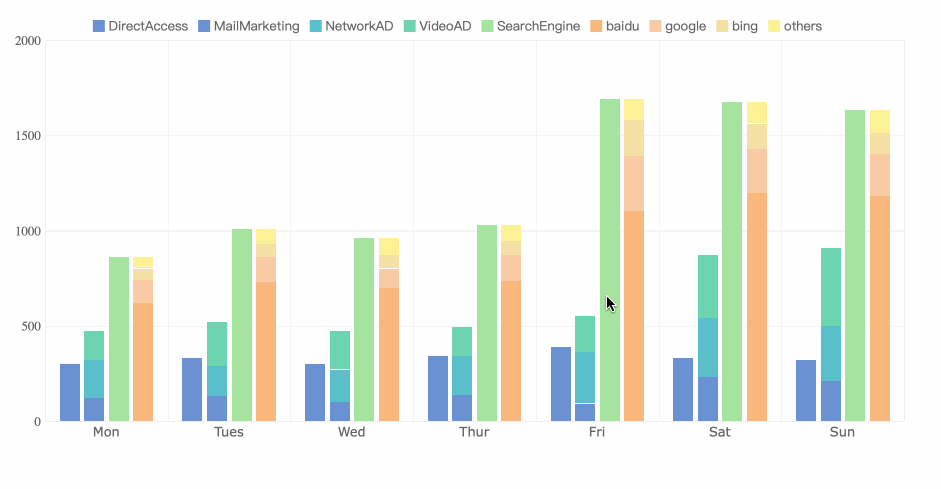
**Data visualization have some more specialties such as:**

* Data visualization can identify areas that need improvement or modifications.
* Data visualization can clarify which factor influence customer behavior.
* Data visualization helps you to understand which products to place where.
* Data visualization can predict sales volumes.

Data visualization tools have been necessary for democratizing data, analytics, and making data-driven perception available to workers throughout an organization. They are easy to operate in comparison to earlier versions of BI software or traditional statistical analysis software. This guide to a rise in lines of business implementing data visualization tools on their own, without support from IT.

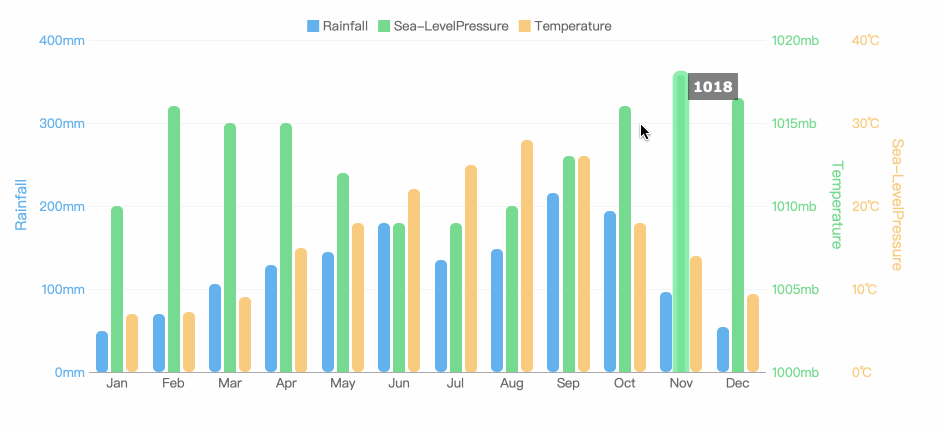
**1. Column Chart**

Column charts use vertical columns to show numerical comparisons between categories, and the number of columns should not be too large (the labels of the axis may appear incomplete if there are too many columns).



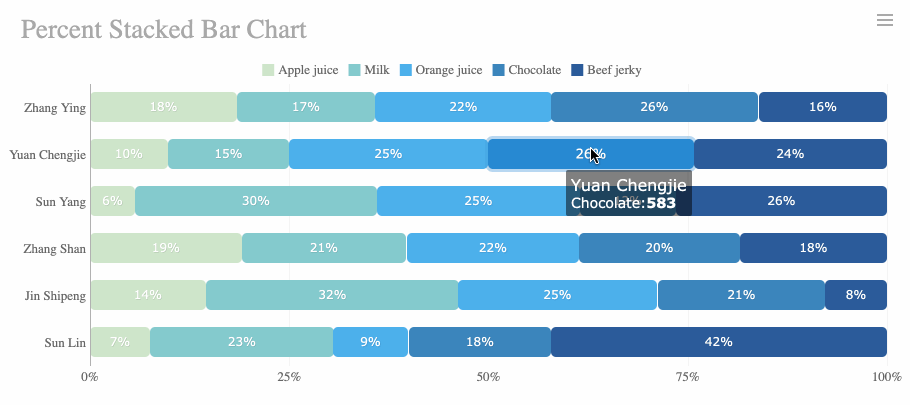
From FineReport

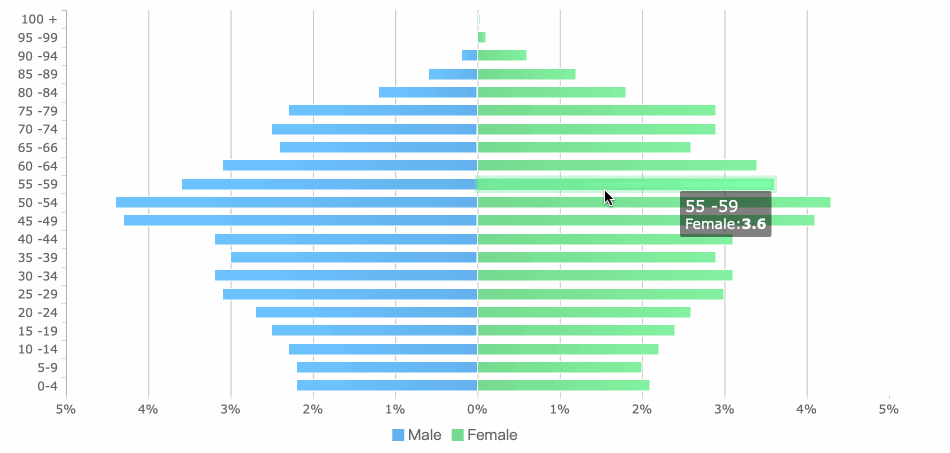
The column chart takes advantage of the height of the column to reflect the difference in the data, and the human eye is sensitive to height differences. The limitation is that it is only suitable for small and medium-sized data sets.



**2. Bar Chart**

Bar charts are similar to column charts, but the number of bars can be relatively large. Compared with the column chart, the positions of its two axes are changed.

[](https://www.finereport.com/en/wp-content/uploads/2019/12/2019072204A.gif)

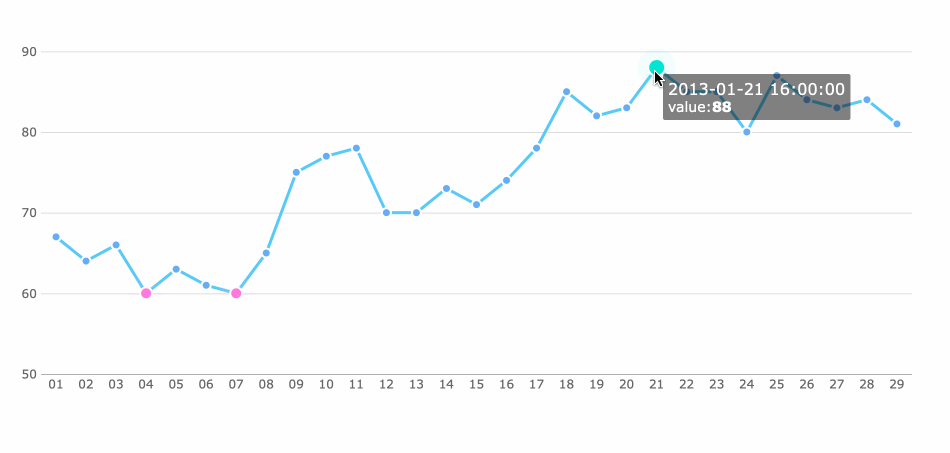


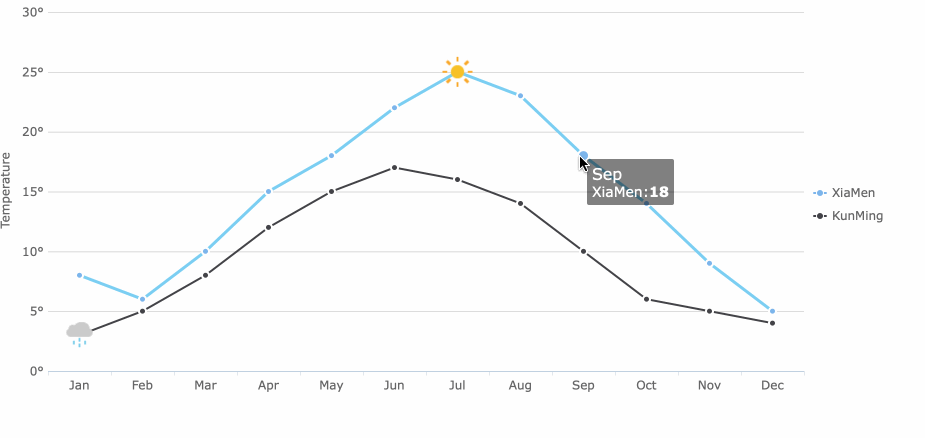
**Application Scenario**: comparison of data (the category name can be longer because there is more space on the Y axis)

**3. Line Chart**

A line chart is used to show the change of data over a continuous time interval or time span. It is characterized by a tendency to reflect things as they change over time or ordered categories.

It should be noted that the number of data records of the line graph should be greater than 2, which can be used for trend comparison of large data volume. And it is better not to exceed 5 polylines on the same graph.



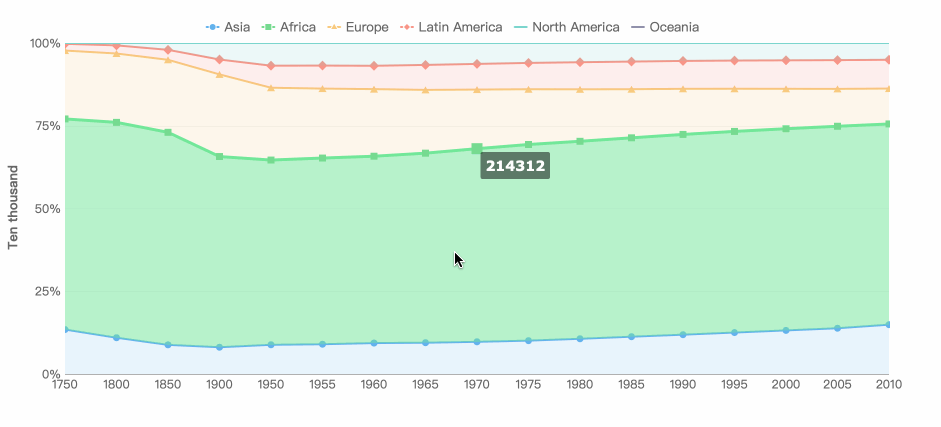


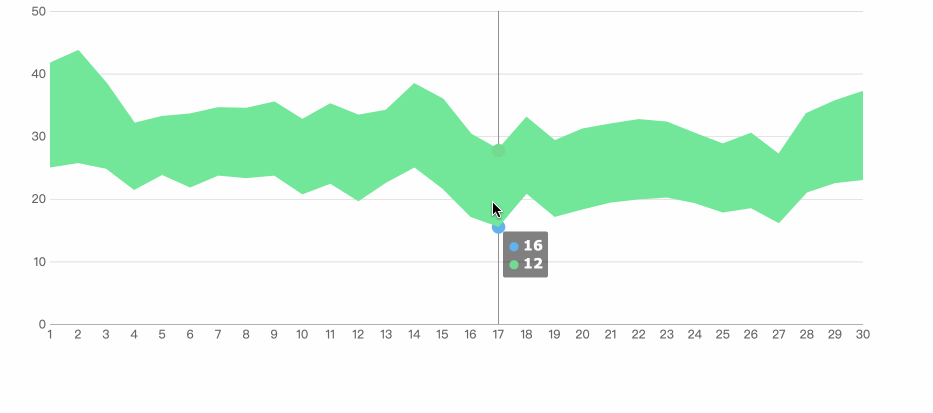
**Application Scenario**: trend of data volume over time, comparison of series trends

**4. Area Chart**

The area chart is formed on the basis of the line chart. It fills the area between the polyline and the axis in the line chart with color. The filling of the color can better highlight the trend information.

The fill color of the area chart should have a certain transparency. The transparency can help the user to observe the overlapping relationship between different series. The area without transparency will cause the different series to cover each other.



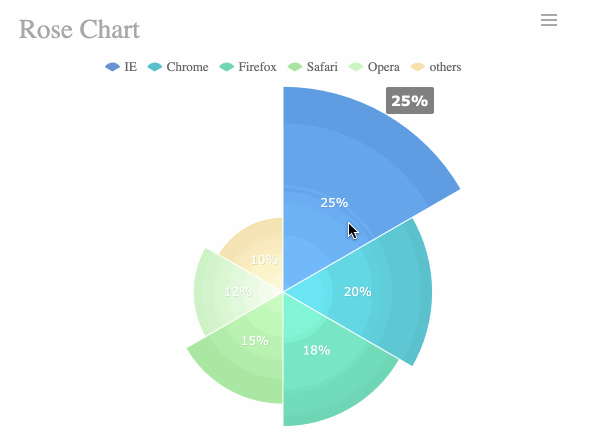


**Application Scenario**: series ratio, time trend ratio

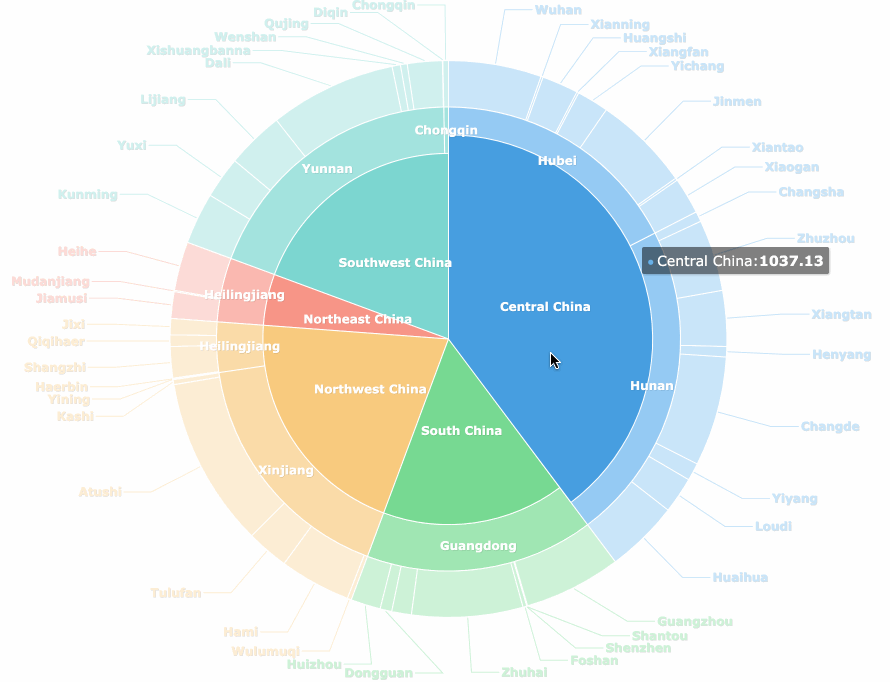
**5. Pie Chart**

Pie charts are widely used in various fields to represent the proportion of different classifications, and to compare various classifications by the arc.

The pie chart is not suitable for multiple series of data, because as the series increase, each slice becomes smaller, and finally the size distinction is not obvious.

[](https://www.finereport.com/en/wp-content/uploads/2019/12/2019072210A.gif)

A pie chart can also be made into a multi-layer pie chart, showing the proportion of different categorical data, while also reflecting the hierarchical relationship.

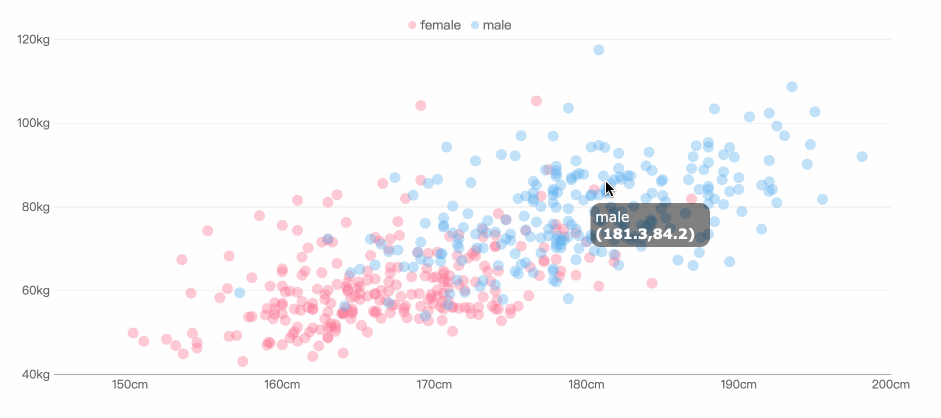
[](https://www.finereport.com/en/wp-content/uploads/2019/12/2019072211A.gif)

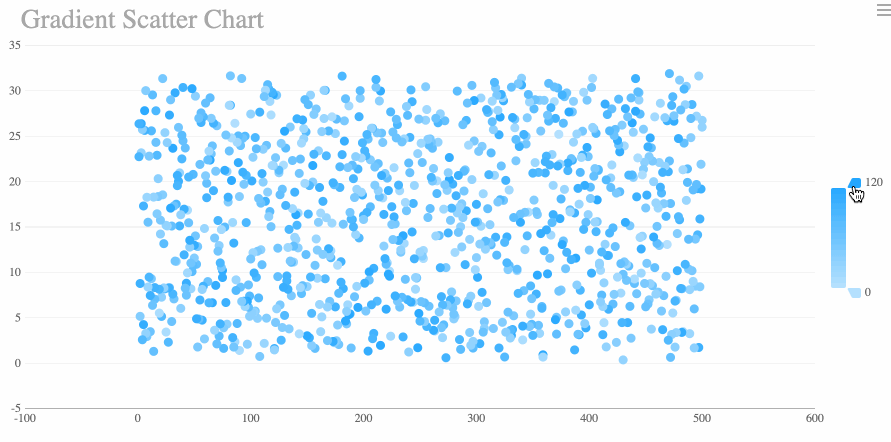
**Application Scenario**: series ratio, series size comparison (rose diagram)

**6. Scatter Plot**

The scatter plot shows two variables in the form of points on a rectangular coordinate system. The position of the point is determined by the value of the variable. By observing the distribution of the data points, we can infer the correlation between the variables.

Making a scatter plot requires a lot of data, otherwise the correlation is not obvious.



[](https://www.finereport.com/en/wp-content/uploads/2019/12/2019072213A.gif)

**Application Scenario**: correlation analysis, data distribution

**7. Bubble Chart**

A bubble chart is a multivariatechart that is a variant of a scatter plot. Except for the values of the variables represented by the X and Y axes, the area of each bubble represents the third value.

We should note that the size of the bubble is limited, and too many bubbles will make the chart difficult to read.

# Data Visualization Tips

## Tap into data storytelling

"Data visualization is powerful. Still, data storytelling is the thing driving decision-makers. People take action when they interact with a data-evidenced proof of the story." — Monika Piekarska

Believe it or not, [data tells a story](https://www.beautiful.ai/blog/data-tells-a-story-so-tell-it). We know that most people are more likely to engage with your presentation if there’s a strong narrative to back it up, so tapping into data storytelling when presenting your information and statistics can help it land better with your audience. Data storytelling is how you choose to communicate your insights to make them more meaningful and relevant to your audience, and can be used to support your overarching message. What do your reports tell you about your customer, proof of your company’s success, or your missed opportunities? By making your data more relatable, you’re making it easier for your audience to comprehend.

## Know when to use the right charts

Charts and infographics are the essence of data visualization. They take your most meaningful data sets and display them in more digestible ways. However, if you don’t [use the right charts](https://www.beautiful.ai/blog/how-to-find-a-graph-or-chart-that-makes-your-data-look-good), it can have the opposite effect and make your data even more confusing. Understanding your data means understanding which charts will help drive your point home. Don’t panic, our [Smart Slides](https://www.beautiful.ai/slide-templates) can help inform your decision.

If you don’t know where to start, you can dip your toes in by perusing our inspiration gallery of [pre-built presentation](http://www.beautiful.ai/presentations) and slide templates curated by expert designers. This allows you to explore some of the different use cases for each of the different types of charts and graphs, and get inspired to create your own. If you see something you like, you can customize it with your own content and then toggle between different layouts to see which format fits best. Not only does this give non-designers the tools they need to get started, it also helps presenters structure their data in new and interesting ways.

## Keep it clean

“You can achieve simplicity in the design of effective charts, graphs and tables by remembering three fundamental principles: restrain, reduce, emphasize.” — Garr Reynolds

Just because your data is complex doesn’t mean your slides should be. Read that again. Keeping your slides clean and simplistic is key for data visualization if you want your audience to retain any of the information you’re presenting to them. Each slide should feature one key takeaway, and it should be obvious to your audience— don’t send them on an easter egg hunt to figure it out.

## Use colors to highlight key points

Colors are your friend. They can help you convey things within your slide more effectively. We suggest using different colors to provide contrast between data sets. Use your boldest colors to represent the more important pieces of information, and more subtle hues to indicate the rest. While your colors should be on-brand and consistent with the rest of your presentation, it’s okay to play around with different hues here. Colors are an easy way to tell the audience exactly what you want them to pay attention to, and how it should make them feel.

## Include supporting visuals

Images, icons, and shapes can help provide additional context for your data. In fact, if some members of your audience find themselves more creative than logical, including the right supporting visuals can help paint a picture for them. While data should be the star of the show in data visualization, certain visual assets can put things into perspective and make them more relatable.

You might also use Beautiful.ai’s [Elements](https://www.beautiful.ai/blog/introducing-elements-a-more-flexible-version-of-beautiful-ai) for special annotations. You can add additional icons, arrows, or text to call out important pieces of information to craft your message in a more meaningful way.

## Be intentional with text

Data visualization is all about painting a picture with strong visual elements, so be intentional with your use of text. You want your audience to focus on the graphs, charts, or percentages on the screen, so don’t distract them with a lengthy paragraph to read. You might need to use bullet points to add context to the metrics, or to call out important aspects of a report, but limit it where you can.

## Bring your data to life with animations

Last, but certainly not least, [bring your data to life with dynamic animations](https://www.beautiful.ai/blog/bring-remote-presentations-to-life-with-animations). Even despite your best efforts to jazz up your data, you might still lose your audience to boredom (it happens). Don’t worry, we have the secret sauce to rein their focus back in. Dynamic animations are subtle movements for when each slide advances, and they’re a surefire way to catch the eye of your audience and pull their attention back to your presentation. You can select the animation style, and speed, so that the data in your graphs and charts build with your story.

