**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.*

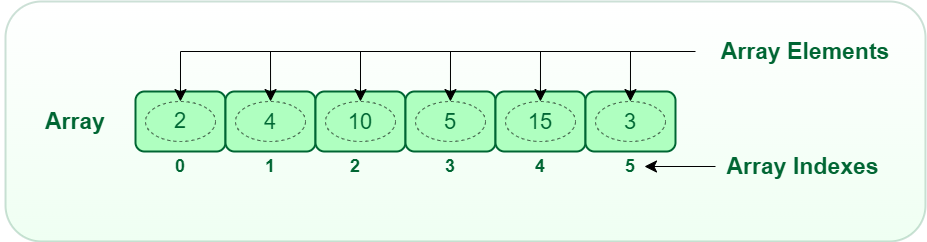
**For example,** we can store a list of items having the same data-type using the *array* data structure.



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

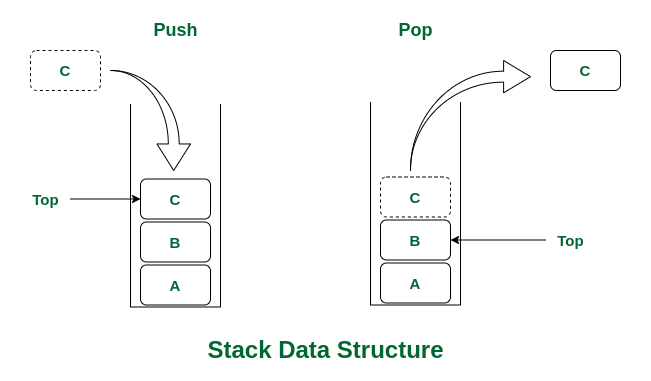
*A linked list is a linear data structure, in which the elements are not stored at contiguous memory locations. The elements in a linked list are linked using pointers as shown in the below image:*



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

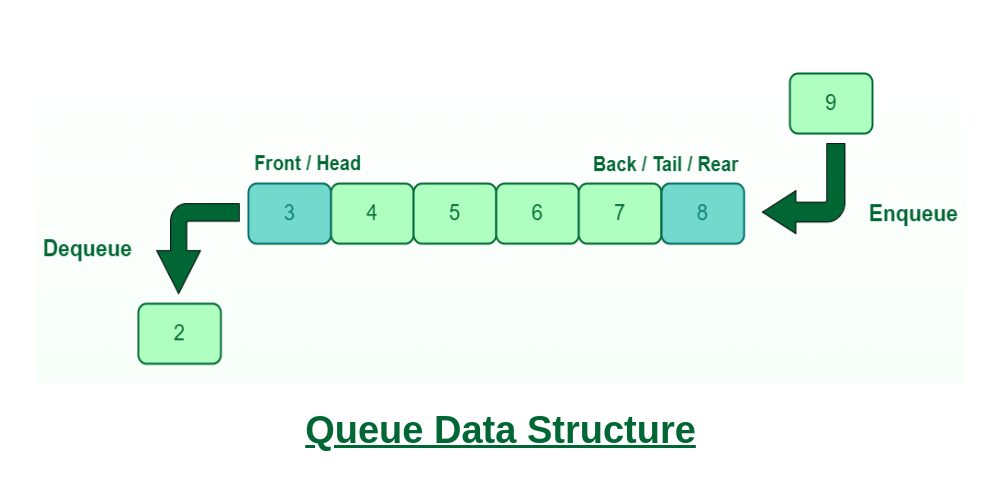
in Last Out) order.

Queue Data Structure

**What is Queue Data Structure?**

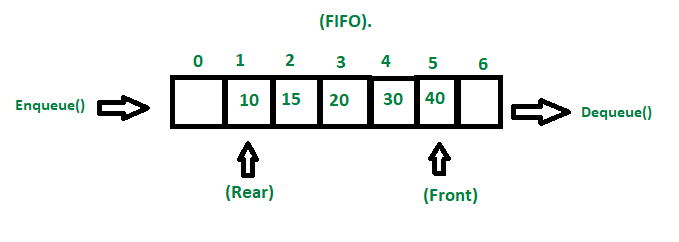
*A****Queue****is defined as a linear data structure that is open at both ends and the operations are performed in First In First Out (FIFO) order.*

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*

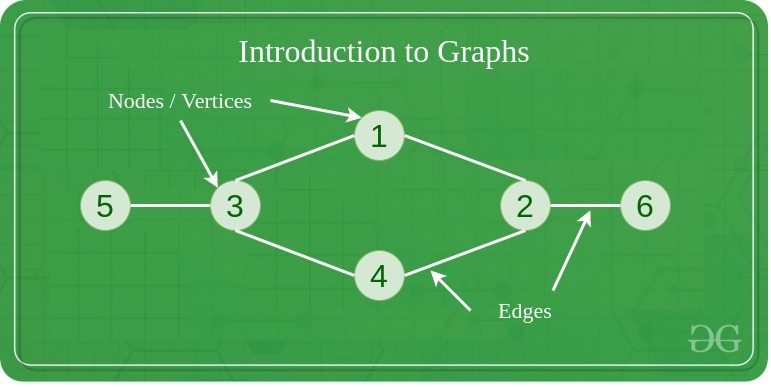
**Characteristics of Queue:**

* 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

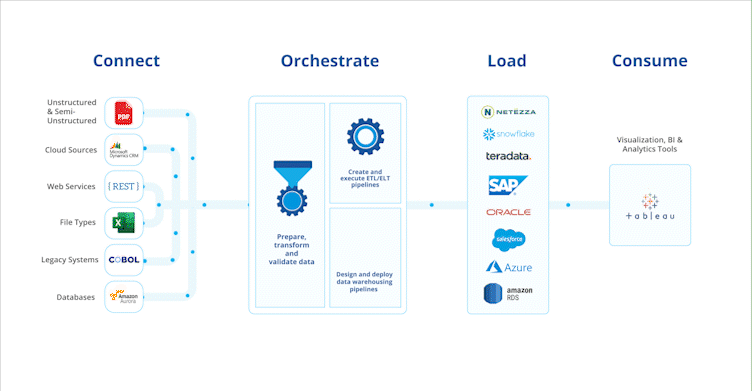
Data cubes are data lists with multidimensions (usually three or more dimensions) stored as a table. They are used to describe the time sequence of an image’s data and help assess gathered data from a range of standpoints.

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:

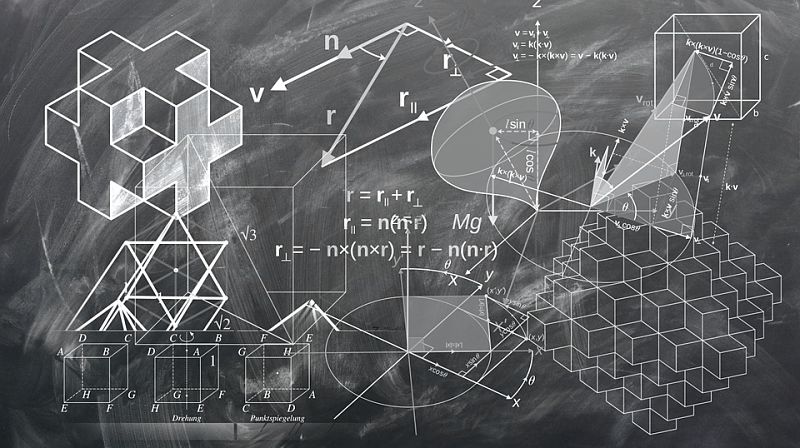
* **Cloud-based storage solutions:** These allow an organization to store data off-site and access it remotely
* **On-site servers:** These give organizations a greater sense of control over how data is stored and used
* **Other storage media:**These include hard drives, USB devices, CD-ROMs, and floppy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

## Data Analyst (DA)

Focusing first on profiles more oriented to data analysis, **Data Analyst is a profile that came before Data Scientist**. In some cases they are refrred to as "Junior Data Scientists ".

They have a fairly generalist role, covering **a wide range of functions that include mining, obtaining and/or retrieving data as well as its processing, advanced study and visualization**.

The study or advanced analysis of data is done based on algorithms, mathematical and statistical methods. Therefore, **this profile mainly requires knowledge of maths and statistics applied to data mining and machine learning**.



The latter means that it is also essential to know how to develop software (at least in current projects). Although its specialty is Machine Learning, the use of libraries of statistical methods such as [Panda](http://pandas.pydata.org/) requires in depth knowledge in the operation of each algorithm, as well as the basic functionality of the corresponding language, in this case Python. Another common language for a Data Analyst could be R.

In addition to the concepts of Machine Learning and the Python and R languages, Data Analysts stand out for their knowledge in the use of notebooks such as Jupyter, as well as knowledge of the Big Data environment in which they work, such as Spark or Hadoop.

It is also well valued that you have knowledge of SQL Databases and traditional Business Intelligence.

## Data Scientist

It is the "evolution of Data Analyst". In many cases they are considered the same profile with a different approach. For us, **it is a more specific role and less aligned with the business vision**.

Like the DA, **it requires knowledge of mathematics, statistics and Machine Learning**, programming languages ​​such as **R or Python, the use of notebooks and Big Data ecosystems**, but what we believe differentiates the [Data Scientist](https://en.paradigmadigital.com/techbiz/data-science-why-do-we-think-were-so-hip/) is that they are responsible for extracting value from data.

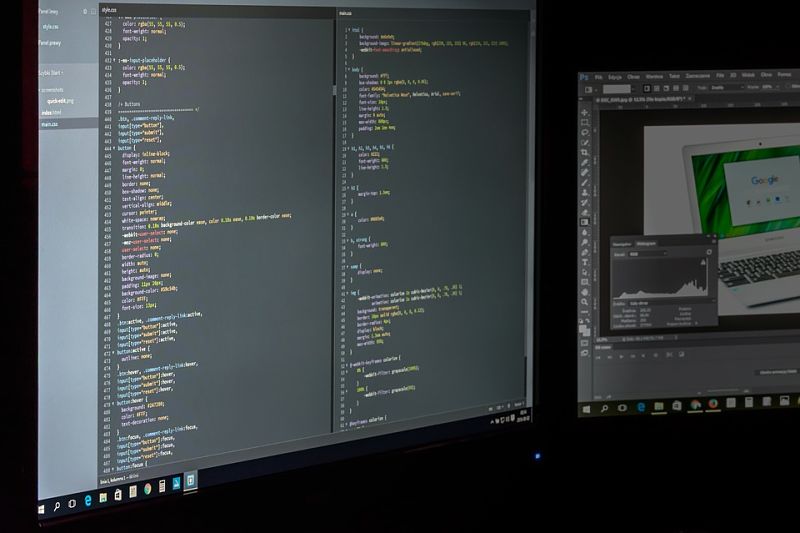
They also obtain, process and visualize data, although with a more focused role in prediction, based on the behaviors learned.

Considering a Data Scientist as a more modern version of Data Analyst, it is more appropriate for them to use more recent libraries such as TensorFlow for Deep Learning techniques based on neural networks.

Also many of its developments are linked to Artificial Intelligence techniques and neuro-linguistic programming (NLP). But, once again, they are quite similar profiles and the inclusion of technologies is not strict for one role or another.

In the case of Data Scientists that use tools such as [SAS Enterprise Miner](https://www.sas.com/en_us/software/enterprise-miner.html) to perform statistical analysis, there is a perception on the part of many that the tool itself does not require programming knowledge, a perception with which we currently disagree.

Although it is true that SAS in many cases provides a much more graphic and visual modelling capacity, it is still required to know how the algorithms behind each operation work, and in many cases, it will also be necessary to know the SAS programming language.



## Data Engineer

Already focusing on the storage and processing of data, we find ourselves with the role of Data Engineer. This is our role in the Aura project at Telefónica and here is one of the reasons why we are going to give it a lot of importance.

Perhaps the most relevant is that **it provides the Big Data project with a value very different from the one provided by a Data Scientist or Data Analyst**.

We know that the latter are the ones that work with the data, but where do they get it from? How does the environment in which they do their analysis work? **It is the task of the Data Engineer to prepare the entire ecosystem so that others can obtain their data clean and prepared for analysis**.

The Data Engineers are those who design, develop, build, test and maintain the data processing systems in the Big Data project.

You must know how the data is modeled as well as having a wide knowledge of the SQL databases, since in the Big Data world they are not excluded and in many cases they are still the origin of the data. They simply complement each other.

They perform and program data intakes (for example, from a relational model to a Spark processing engine). They also do cleaning, validation, data quality and aggregation processes so that the information reaches the Data Scientist as expected, and they configure the cluster in Spark (number of nodes and cores per node, GB of RAM) so that the statistical models are executed optimally.

What technologies do they use? **A Data Engineer should know Linux and Git** much like an engineer working on software projects. **Hadoop and Spark at the environment level; Map Reduce at the level of computational models; and HDFS, MongoDB and** [Cassandra](https://en.paradigmadigital.com/dev/cassandra-lady-nosql-databases/)**at the level of NoSQL technologies**.

In terms of programming languages **​​it is essential to know SQL**, since the relational model is still an important part in the generation and query of data.

I**t is also usually required to know one or two of the following languages: Python** for data processing (sometimes PySpark) and **Scala as the native language of Spark and Java** in many cases.

**Should a Data Engineer know the models used by the Data Scientist in depth?** In principle, you should know what it means to use one or another model for the environment, and what architecture is ideal for them to work in.

# **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.

