**Data Engineering**

**Data**

* Data refers to any collection of facts, figures, symbols, or descriptions that can be recorded, stored, and manipulated. It can take various forms, including text, numbers, images, audio, video, and more.
* Data can be classified into various types based on:

1. Origin:

* Internal data
* External data

1. Frequency:

* Real-time Data: refers to information that is generated, processed, and made available for analysis or action immediately as it is produced. It represents the most current state of a system. As Social Media Feeds (Tweets, posts, and comments on social media platforms) are streamed in real-time, enabling immediate interactions and insights.
* Historical Data: refers to past records, and observations that have been collected, stored, and preserved over time. It represents a historical perspective on events, trends, and behaviors. As Sales Data Historical records of sales transactions, customer purchases, and inventory levels provide insights into past performance and trends.

1. Structure:

* Structured data: data that has a high level of organization and follows a specific format. It is typically stored in tables with rows and columns, where each column represents a distinct attribute, and each row represents a record.

**Ex:** relational databases, spreadsheets, CSV files.

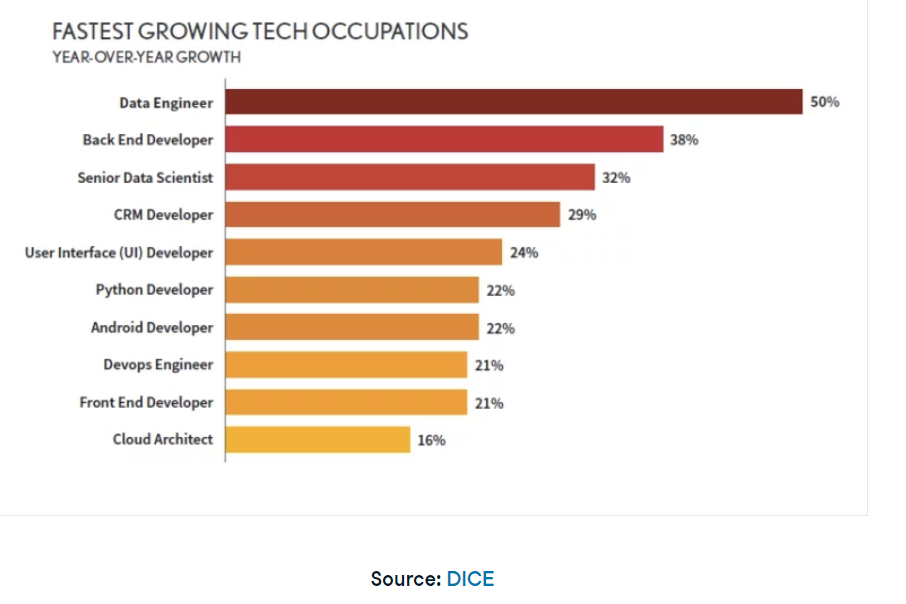
* Semi-structured data: lacks a strict, predefined structure like structured data but has some level of organization. It may contain keys that provide a partial structure, allowing for flexibility in representation. With some processes, it can be stored in the relational database.

**Ex:** XML and JSON files

* Unstructured data: Data does not adhere to any predefined schema or organization. this data does not fit in a relational database.

**Ex:** Text documents (Word files, PDFs), social media feeds, multimedia content (images, videos, audio recordings), sensor data streams.

**Data Engineering**

****

**A Data Engineer** is responsible for handling large amounts of data with various types. They work closely with data scientists and analysts to design data models and schemas that facilitate efficient data storage and retrieval.

Data engineers are responsible for the process of data collection and storage. They ensure that the large volume of data collected from different sources becomes accessible raw material for other data science specialists, such as data analysts and data scientists

the data engineer’s day-to-day runs fundamentally between two processes:

**ETL** (Extract, Transform, Load) Processes include developing data extraction, transformation, and loading tasks, and moving data between different environments.

**Data Cleaning** Processes so that it arrives in a normalized and structured fashion into the hands of analysts and data scientists.

However, the process of data collection and storage can be extremely complex. There may be different data sources involved, and these data sources may have different types of data. As the volume, variety, and velocity of the data at hand increase, so does the complexity of the data engineer’s work

To ensure that the tasks performed are timely, robust, and scalable, data engineers develop the so-called **data pipelines**

**Build, test, and maintain data pipeline.**

**Data Pipeline**

A data pipeline is a system for retrieving data from various sources and then ported to a data store, like a data lake data warehouse, or database.

### **Key reasons to use data pipelines:**

1. **Single-source limitations**: Pre-made datasets are often limited to single-source data, restricting the scope of analysis.
2. **Data integration**: Pipelines enable the combination of data from multiple sources, enriching the analysis.
3. **Automated analysis**: Pipelines facilitate regular, automated data collection and analysis, saving time and reducing manual errors.
4. **Data consistency**: They help in standardizing and transforming data from different sources to maintain consistency.
5. **Efficiency**: Automated pipelines streamline the data workflow, making the analysis process more efficient.

**Example**:  
Imagine you want to know if the average GPA of students in a school district correlates with the average income in the county or with the population size. For this analysis, you may need to pull GPA data from each school district in your analysis, income data from the IRS

”مصلحة الضرائب”, and population size from the Census Bureau “مكتبالاحصاء”.

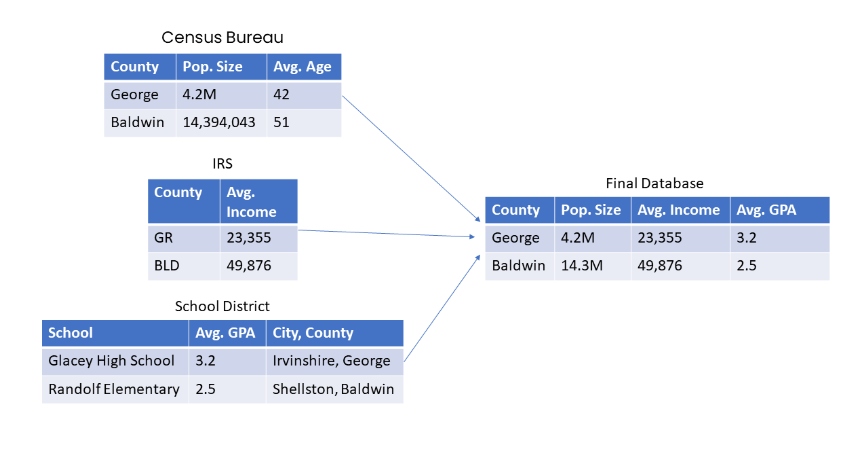
You could probably pull this information manually once. But imagine you need to repeat this analysis regularly. Building a reliable pipeline can help you automate this information collection, saving you time and headaches.

### **Example: Handling data inconsistencies**

In addition to just collecting the data in one place, pipelines can help you deal with inconsistencies in data from different sources. For example, the Census Bureau may identify counties by full name, while the IRS may identify counties by an abbreviated name.

### **Example: Data transformation**

Within the pipeline, you can set up a system to standardize the way counties are identified in the final repository. Or, perhaps the population is marked with different units in different entries, and you need to standardize which unit system is used in your final destination. These types of changes are called transformations.

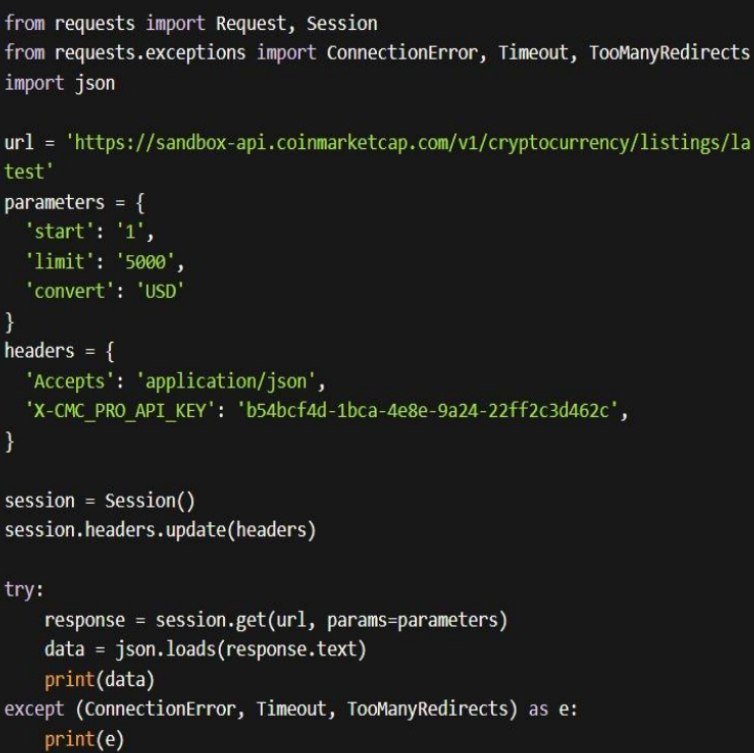


**ETL** (Extract, Transform, Load)

**1- Extract:**

* **Data Sources**: Begin by identifying the sources from which data needs to be extracted. These could be databases, flat files (like CSV or Excel), APIs, or web services.
* **Data Extraction**: This is the process of retrieving data from various sources and collecting it in a structured format for further analysis, processing, or storage.
* **Extracting Methods:**

1. Database Queries: SQL queries are commonly used to extract data from relational databases..
2. ETL Tools: Extract, Transform, Load (ETL) tools automate the extraction process by connecting to various data sources, performing transformations, and loading data into a target destination.
3. Streaming Data: Continuous extraction of real-time data streams from sources like sensors, IoT devices, or social media platforms.
4. Web Scraping: Extracting data from websites by parsing HTML pages and extracting relevant information using web scraping techniques.
5. File Parsing: Reading and parsing files in different formats (such as CSV, JSON, XML) to extract data.
6. **API Extraction**: Application Programming Interfaces (APIs) allow access to data from web services or software applications, enabling programmatic extraction of data.

****

**2-Transform** ( practically )

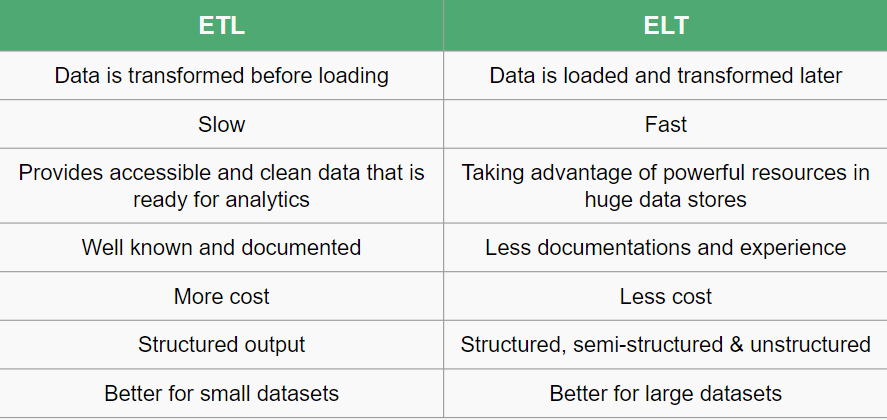
**Data Cleaning**: Cleanse the extracted data to remove inconsistencies, errors, or duplicates. This might involve fixing missing values, correcting formatting issues, standardizing data formats, or removing irrelevant data.

**Data Transformation**: Transform the data into a format that is suitable for analysis or storage. This could include aggregating data, splitting or merging fields, performing calculations, applying business rules, or enriching the data with additional information

**3- Loading:**

Once extracted and transformed, the data is typically loaded into a target destination, such as a data warehouse, database, data lake, or analytical platform, where it can be further processed, analyzed, or used for reporting and decision-making.

**ETL vs ELT**



## Transformation can be Multistage (ETL) or In-warehouse (ELT)

## 

## 

## 

## 

### **ETL Tool: Apache Spark**

Apache Spark is widely used for ETL processes due to its speed, scalability, and ease of use.

Apache Spark can **extract** data from various sources, including databases, data lakes, streaming platforms, and more. It provides connectors for reading data from different file formats (CSV, JSON), and databases (e.g., MySQL, PostgreSQL).

Developers can use Spark's DataFrame and Dataset APIs, as well as SQL queries, to perform various **transformations** such as filtering, joining, and sorting.

It also can **load** the processed data into various target destinations, including databases, data warehouses, data lakes, and external storage systems.

## 

## **Data Storage: Data Lakes, Data Warehouses, and Databases**

Data can be **stored** in several different ways, depending on **the type of data** and your **desired use**. Three terms you’re likely to come across concerning data storage are data lakes, databases, and data warehouses. Let’s briefly touch on each of these.

### **Data lake**

A data lake is a large, flexible storage repository that can hold both structured and unstructured data at scale. It allows organizations to store raw, unprocessed data from various sources, such as weblogs, sensors, social media, and more, in its **native format**.

Data lakes are known for their versatility, as they enable data scientists and analysts to explore, transform, and analyze data in a variety of ways. They are particularly valuable for big data and advanced analytics use cases where the schema and structure of the data are **not predefined**, giving organizations the flexibility to extract insights from diverse data sources.

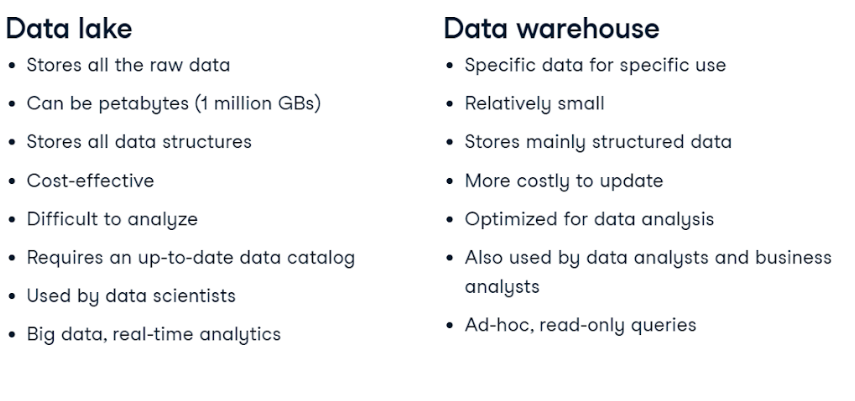
### **Database**

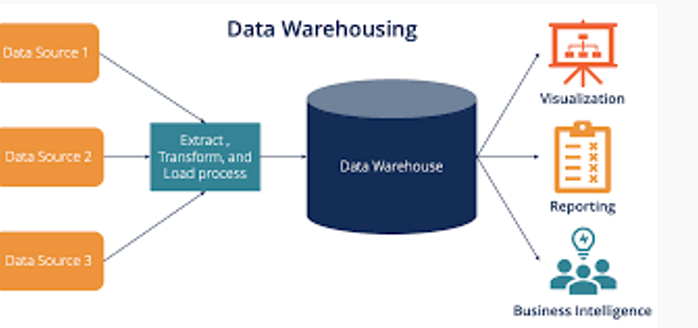
Databases are used for efficiently storing, managing, and retrieving **structured data**. While they share some similarities with data warehouses, databases are typically used for **transactional data, day-to-day operations**, and real-time applications. They come in various types, including relational databases, NoSQL databases, each suited to **specific** data storage and processing requirements. Databases ensure data integrity, support concurrent access, and provide various mechanisms for querying and retrieving data.

### **Data warehouse**

Data warehouses, on the other hand, are **highly** structured and optimized repositories designed for storing, organizing, and querying structured data primarily for analytical purposes. They typically follow a predefined schema, enforce data quality and consistency, and are optimized for **complex** querying and reporting tasks.

Data warehouses are ideal for businesses looking to make informed decisions based on **historical data**, as they provide a single source of truth for standardized reporting and analytics. Data is usually extracted from various sources, transformed into a consistent format, and loaded into the data warehouse for analysis.





## **Data Pipeline Use Cases and Stakeholders**

Data pipelines are generally developed and maintained by data engineers.

Once the data is collected, cleaned, and organized at the end of the pipeline, the data can be pulled by various stakeholders, including **data analysts, data scientists, or business analysts.**

The data can then be used for real-time analytics, dumped into an updating dashboard, or added to a machine learning model. Sometimes, the end of the pipeline may be an application or dashboard used by a business leader or individual.

Pipelines are used by virtually every sector that utilizes large amounts of data. In some **large organizations**, like those in the tech sector, their pipelines will be complex, highly regulated, and managed by a team of specialized engineers.

In **smaller organizations**, for example, a single researcher trying to collect data from multiple sensors worldwide, the pipeline may be straightforward, with only a few steps.

Pipelines are also crucial for machine learning models. These models consume huge amounts of data; a proper pipeline is imperative for smooth functioning.

### **Data warehouse Tool: Google Cloud BigQuery**

### **What is Google Cloud BigQuery?**

### BigQuery is a serverless, multi-cloud data warehouse that simplifies the process of working with all types of data so you can focus on getting valuable business insights quickly. At the core of Google’s data cloud, BigQuery allows you to simplify data integration, cost-effectively and securely scale analytics, share rich data experiences with built-in business intelligence, and train and deploy ML models with a simple SQL interface, helping to make your organization’s operations more data-driven. The most important thing is “**it’s free**” and there are some other paid plans for more features and it costs 1$ / MONTHS**.**

**References:**

<https://www.datacamp.com/blog/how-to-become-a-data-engineer>

[An Introduction to Data Pipelines for Aspiring Data Professionals | DataCamp](https://www.datacamp.com/tutorial/introduction-to-data-pipelines-for-data-professionals)

[What is the difference between ETL and ELT - Public - مستندات Google](https://docs.google.com/document/d/1LWiYglGAcDKGdFJotFV-GMaRCZrEYcTbyvevso0ESJo/edit)

[Difference between Data Warehouse and Data Mart - GeeksforGeeks](https://www.geeksforgeeks.org/difference-between-data-warehouse-and-data-mart/)