What is Big Data?

extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions

Top Data Terms you should know?

**Database:**

A database is an organized collection of structured information, or data, typically stored electronically in a computer system. A database is usually controlled by a database management system (DBMS)

**Data warehouse:**

A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data**.**

.

**Data Marts:** A data mart is a simple form of data warehouse focused on a single subject or line of business. With a data mart, teams can access data and gain insights faster, because they don’t have to spend time searching within a more complex data warehouse or manually aggregating data from different sources.

**Data Lakehouse:** A data lakehouse is a new, open data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data.

**Data Mesh:** A data mesh is a decentralized data architecture that organizes data by a specific business domain—for example, marketing, sales, customer service, and more—providing more ownership to the producers of a given dataset. The producers’ understanding of the domain data positions them to set data governance policies focused on documentation, quality, and access. This, in turn, enables self-service use across an organization.

**Can a database be used as a DWH?**

Yes, a database can be used as a data warehouse. In fact, many data warehouses are built using relational database management systems (RDBMS) such as Oracle, Microsoft SQL Server, and MySQL. These databases are designed to handle large volumes of structured data and are capable of performing complex queries and analyses.

To use a database as a data warehouse, you would typically create a schema that is optimized for reporting and analysis. This might involve creating denormalized tables or using indexing and partitioning to improve query performance. You would also likely use a data integration tool to extract data from various source systems and load it into the data warehouse.

It's worth noting that there are also specialized data warehousing platforms available that are designed specifically for this purpose. These platforms often offer additional features such as columnar storage, compression, and advanced analytics capabilities. However, if you already have a relational database in place, it may be more cost-effective to use it as a data warehouse rather than investing in a separate platform.

**Major Difference between Structure and un-structured Data**

| Structured data | Unstructured data |
| --- | --- |
| Structured data is quantitative data that consists of numbers and values. | Unstructured data is qualitative data that consists of audio, video, sensors, descriptions, and more. |
| Structured data is used in machine learning and drives machine learning algorithms. | Unstructured data is used in natural language processing and text mining. |
| Structured data is stored in tabular formats like excel sheets or SQL databases. | Stored as audio files, videos files, or NoSQL databases |
| Structured data has a pre-defined data model. | Unstructured data does not have a pre-defined data model. |
| Structured data is sourced from online forms, GPS sensors, network logs, web server logs, OLTP systems, and the like. | Unstructured data is sourced from email messages, word-processing documents, pdf files, and so on. |
| Structured data is stored in data warehouses | Unstructured data is stored in data lakes |
| Structured data requires less storage space and is highly scalable. | Unstructured data requires more storage space and is difficult to scale. |

**Data Engineers duties?**

The duties of a data engineer can vary depending on the company, industry, and specific job responsibilities, but some common duties may include:

Designing and implementing data pipelines and ETL processes to move data from source systems to data warehouses or data lakes

Developing and maintaining data storage solutions, such as databases and data warehouses, and ensuring data integrity and security

Optimizing data retrieval and storage systems for efficient and reliable access to large volumes of data

Collaborating with cross-functional teams, such as data analysts and data scientists, to identify and meet their data needs

Troubleshooting data-related issues and developing solutions to prevent data loss or corruption

Developing and maintaining documentation of data pipelines, storage solutions, and processes for future reference and training purposes

Staying up-to-date with the latest data technologies and trends to ensure that data infrastructure and processes remain efficient and effective

Ensuring compliance with data privacy and security regulations, such as GDPR and HIPAA.

Overall, data engineers play a critical role in designing, building, and maintaining the data infrastructure that enables organizations to collect, store, process, and analyze large volumes of data to gain insights and make informed decisions.

What is ETL?

The ETL process consists of 3 steps that enable data integration from source to destination: data extraction, data transformation, and data loading.

### **Step 1: Extraction**

Most businesses manage data from a variety of data sources and use a number of data analysis tools to produce business intelligence. To execute such a complex data strategy, the data must be able to travel freely between systems and apps.

Before data can be moved to a new destination, it must first be extracted from its source — such as a data warehouse or data lake. In this first step of ETL process structured and unstructured data is imported and consolidated into a single repository. Volumes of data can be extracted from a wide range of data sources, including:

* Existing databases and legacy systems
* Cloud, hybrid, and on-premises environments
* Sales and marketing applications
* Mobile devices and apps
* CRM systems
* Data storage platforms
* Data warehouses
* Analytics tools

### 

### 

### 

### **Step 2: Transformation**

During this phase of the ETL process, rules and regulations can be applied that ensure data quality and accessibility. You can also apply rules to help your company meet reporting requirements. The process of data transformation is comprised of several sub-processes:

* Cleansing — inconsistencies and missing values in the data are resolved.
* Standardization — formatting rules are applied to the dataset.
* Deduplication — redundant data is excluded or discarded.
* Verification — unusable data is removed and anomalies are flagged.
* Sorting — data is organized according to type.
* Other tasks — any additional/optional rules can be applied to improve data quality.

Transformation is generally considered to be the most important part of the ETL process. Data transformation improves data integrity — removing duplicates and ensuring that raw data arrives at its new destination fully compatible and ready to use.

### **Step 3: Loading**

The final step in the ETL process is to load the newly transformed data into a new destination (data lake or data warehouse.) Data can be loaded all at once (full load) or at scheduled intervals (incremental load).

**Historical Load:** Historical load is the one-time initial load of data that the Source already had before the creation of the Pipeline.

**Full loading** — In an ETL full loading scenario, everything that comes from the transformation assembly line goes into new, unique records in the data warehouse or data repository. Though there may be times this is useful for research purposes, full loading produces datasets that grow exponentially and can quickly become difficult to maintain.

**Incremental loading** — A less comprehensive but more manageable approach is incremental loading. Incremental loading compares incoming data with what’s already on hand, and only produces additional records if new and unique information is found. This architecture allows smaller, less expensive data warehouses to maintain and manage business intelligence.

ELT — the next generation of ETL

ELT is a modern take on the older process of extract, transform, and load in which transformations take place before the data is loaded. Over time, running transformations before the load phase is found to result in a more complex data replication process. While the purpose of ETL is the same as ELT, the method is evolved for better processing.

## **ELT vs ETL**

Traditional ETL software extracts and transforms data from different sources before loading it into a data warehouse or data lake. With the introduction of the cloud data warehouse, there was no longer the need for data cleanup on dedicated ETL hardware before loading into your data warehouse or data lake. The cloud enables a push-down ELT architecture with two steps changed from the ETL pipeline.

EXTRACT

* Extract the data from multiple data sources and connectors

LOAD

* Load it into the cloud data warehouse

TRANSFORM

* Transform it using the power and scalability of the target cloud platform

If you are still on premises and your data isn't coming from several different sources, ETL tools still fit your data analytics needs. But as more businesses move to a cloud data architecture (or hybrid), ELT processes are more adaptable and scalable to evolving needs of cloud-based businesses.

## **ETL process vs ELT processes**

ETL tools require processing engines for running transformations prior to loading data into a destination. On the other hand, with ELT, businesses use the processing engines in the destinations to efficiently transform data within the target system itself. This removal of an intermediate step streamlines the data loading process.