Understanding Data Repositories and Big Data Platforms

In this lesson, you have learned the following information:

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

The different types of Data Repositories include:

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

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

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

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

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