# Summary and Highlights

# 

In this lesson, you have learned the following information:

A data analyst ecosystem includes the infrastructure, software, tools, frameworks, and processes used to gather, clean, analyze, mine, and visualize data.

Based on how well-defined the structure of the data is, data can be categorized as:

* Structured Data, that is data which is well organized in formats that can be stored in databases.
* Semi-Structured Data, that is data which is partially organized and partially free form.
* Unstructured Data, that is data which can not be organized conventionally into rows and columns.

Data comes in a wide-ranging variety of file formats, such as delimited text files, spreadsheets, XML, PDF, and JSON, each with its own list of benefits and limitations of use.

Data is extracted from multiple data sources, ranging from relational and non-relational databases to APIs, web services, data streams, social platforms, and sensor devices.

Once the data is identified and gathered from different sources, it needs to be staged in a data repository so that it can be prepared for analysis. The type, format, and sources of data influence the type of data repository that can be used.

Data professionals need a host of languages that can help them extract, prepare, and analyze data. These can be classified as:

* Querying languages, such as SQL, used for accessing and manipulating data from databases.
* Programming languages such as Python, R, and Java, for developing applications and controlling application behavior.
* Shell and Scripting languages, such as Unix/Linux Shell, and PowerShell, for automating repetitive operational tasks.

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.