## What is a data lake?

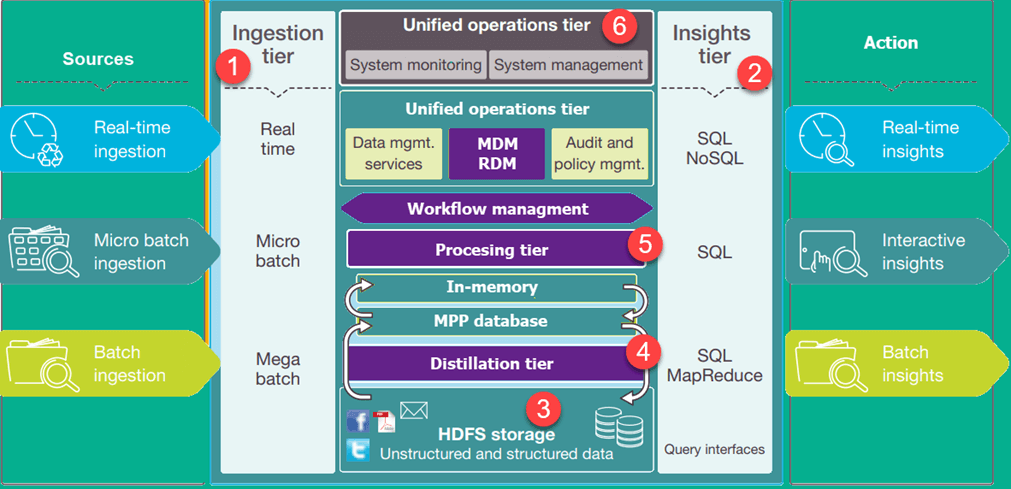
A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

Data Lake is like a large container which is very similar to real lake and rivers. Just like in a lake you have multiple tributaries coming in, a data lake has structured data, unstructured data, machine to machine, logs flowing through in real-time.

**Purpose of Data Lakes:**

The primary purpose of a data lake is to make organizational data from different sources accessible to various end-users like business analysts, data engineers, data scientists, product managers, executives, etc., to enable these personas to leverage insights in a cost-effective manner for improved business performance

## Data Lake Architecture



The figure shows the architecture of a Business Data Lake. The lower levels represent data that is mostly at rest while the upper levels show real-time transactional data. This data flow through the system with no or little latency. Following are important tiers in Data Lake Architecture:

1. **Ingestion Tier**: The tiers on the left side depict the data sources. The data could be loaded into the data lake in batches or in real-time
2. **Insights Tier:**The tiers on the right represent the research side where insights from the system are used. [SQL](https://www.guru99.com/sql.html), NoSQL queries, or even excel could be used for data analysis.
3. **HDFS**is a cost-effective solution for both structured and unstructured data. It is a landing zone for all data that is at rest in the system.
4. **Distillation tier** takes data from the storage tire and converts it to structured data for easier analysis.
5. **Processing tier** run analytical algorithms and users queries with varying real time, interactive, batch to generate structured data for easier analysis.
6. **Unified operations tier** governs system management and monitoring. It includes auditing and proficiency management, data management, workflow management.

## Layers of Data Lakes

* **Bronze Layer**: A one-on-one copy of the data from the source into the data lake. ‘Bronze data’ is raw untransformed unmodified data and all your sources land into this layer.
* **Silver Layer**: Once a business case has been identified and requires analysis, the ‘raw Bronze data’ is transformed into sets of data that add additional values. This can imply replacements of codes to meaningful values, adding sanity constraints, filtering-out unneeded information, … . Hence, resulting in concise useful datasets that may be used by other pieces of information as well.
* **Gold Layer**: The gold layer then provides a well-constructed dataset ready for analysis by data scientists and business analysts. The data is presented in such a way that appeals to them the most, which may include aggregations, joins and merges, encoding, etc.

## What is ETL?

ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system.

ETL provides the foundation for data analytics and machine learning workstreams. Through a series of business rules, ETL cleanses and organizes data in a way which addresses specific business intelligence needs, like monthly reporting, but it can also tackle more advanced analytics, which can improve back-end processes or end user experiences. ETL is often used by an organization to:

* Extract data from legacy systems
* Cleanse the data to improve data quality and establish consistency
* Load data into a target database

## How ETL works

The easiest way to understand how ETL works is to understand what happens in each step of the process.

### **Extract**

During data extraction, raw data is copied or exported from source locations to a staging area. Data management teams can extract data from a variety of data sources, which can be structured or unstructured. Those sources include but are not limited to:

* SQL or NoSQL servers
* CRM and ERP systems
* Flat files
* Email
* Web pages

### **Transform**

In the staging area, the raw data undergoes data processing. Here, the data is transformed and consolidated for its intended analytical use case. This phase can involve the following tasks:

* Filtering, cleansing, de-duplicating, validating, and authenticating the data.
* Performing calculations, translations, or summarizations based on the raw data. This can  include changing row and column headers for consistency, converting currencies or other units of measurement, editing text strings, and more.
* Conducting audits to ensure data quality and compliance
* Removing, encrypting, or protecting data governed by industry or governmental regulators
* Formatting the data into tables or joined tables to match the schema of the target data warehouse.

### **Load**

In this last step, the transformed data is moved from the staging area into a target data warehouse. Typically, this involves an initial loading of all data, followed by periodic loading of incremental data changes and, less often, full refreshes to erase and replace data in the warehouse. For most organizations that use ETL, the process is automated, well-defined, continuous and batch-driven. Typically, ETL takes place during off-hours when traffic on the source systems and the data warehouse is at its lowest.