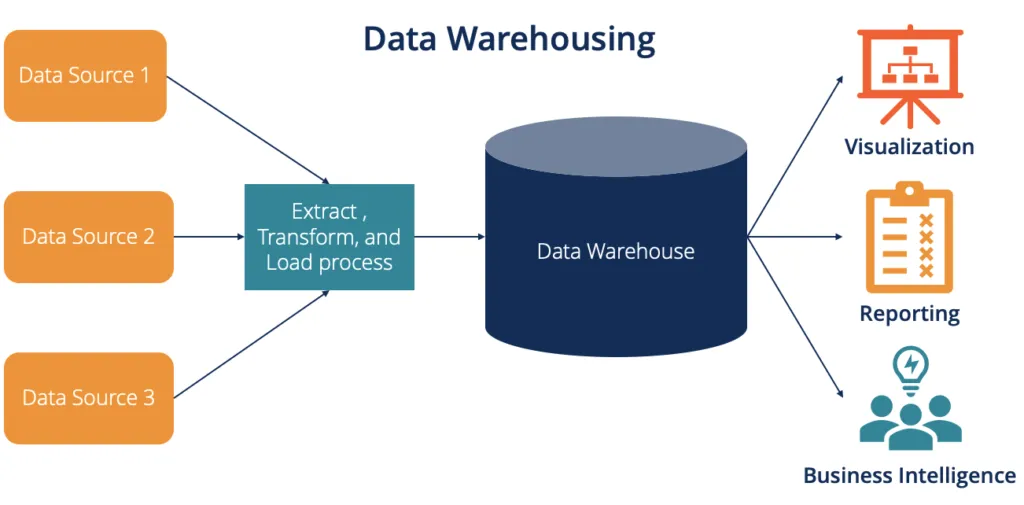
**Data Warehouse**

**What is Data Warehouse?**

***A data warehouse is a central repository of information that can be analysed to make more informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence.***

****

**A Data Warehouse (DW) is a relational database that is designed for query and analysis rather than transaction processing. It includes historical data derived from transaction data from single and multiple sources.**

**A Data Warehouse provides integrated, enterprise-wide, historical data and focuses on providing support for decision-makers for data modeling and analysis.**

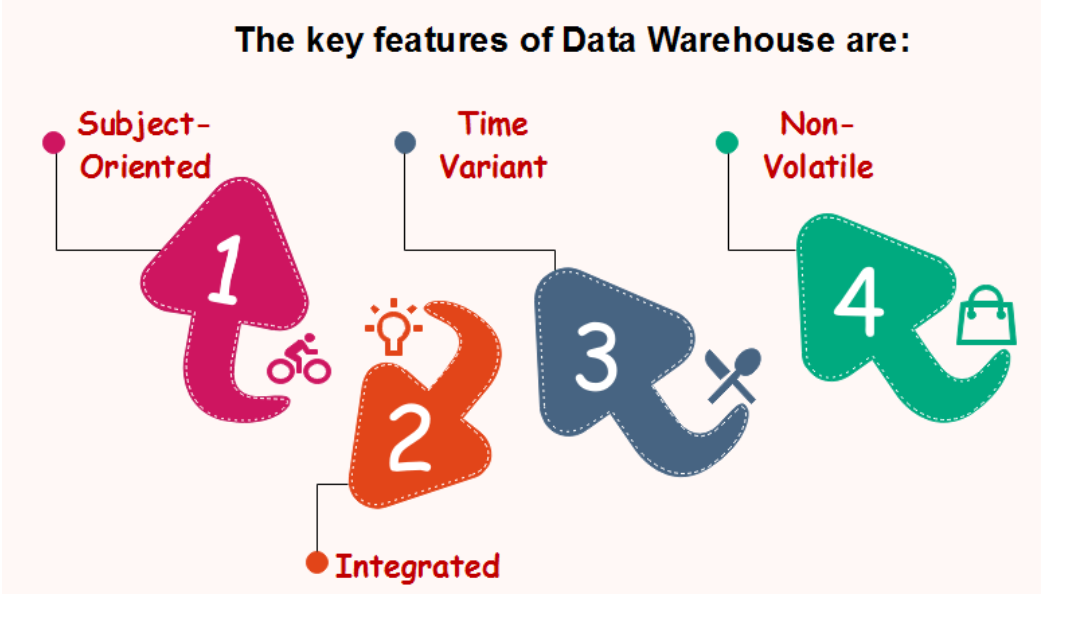
A Data Warehouse is a group of data specific to the entire organization, not only to a particular group of users.

It is not used for daily operations and transaction processing but used for making decisions.

A Data Warehouse can be viewed as a data system with the following attributes:

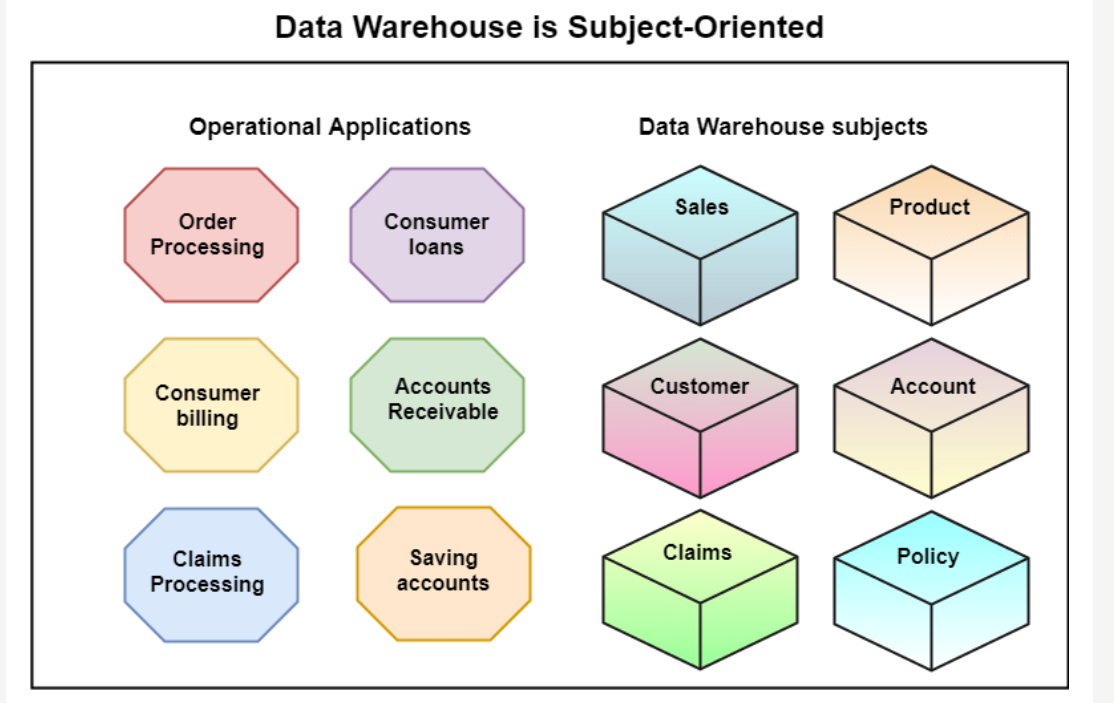
* It is a database designed for investigative tasks, using data from various applications.
* It supports a relatively small number of clients with relatively long interactions.
* It includes current and historical data to provide a historical perspective of information.
* Its usage is read-intensive.
* It contains a few large tables.

"Data Warehouse is a subject-oriented, integrated, and time-variant store of information in support of management's decisions."

****

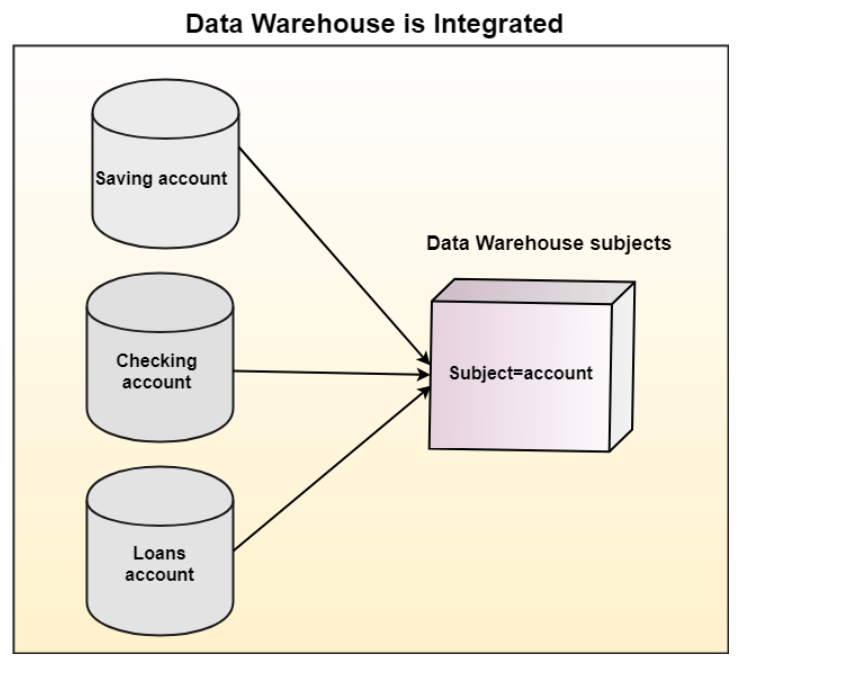
Subject-Oriented

A data warehouse target on the modeling and analysis of data for decision-makers. Therefore, data warehouses typically provide a concise and straightforward view around a particular subject, such as customer, product, or sales, instead of the global organization's ongoing operations. This is done by excluding data that are not useful concerning the subject and including all data needed by the users to understand the subject.

****

Integrated

A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, and online transaction records. It requires performing data cleaning and integration during data warehousing to ensure consistency in naming conventions, attributes types, etc., among different data sources.

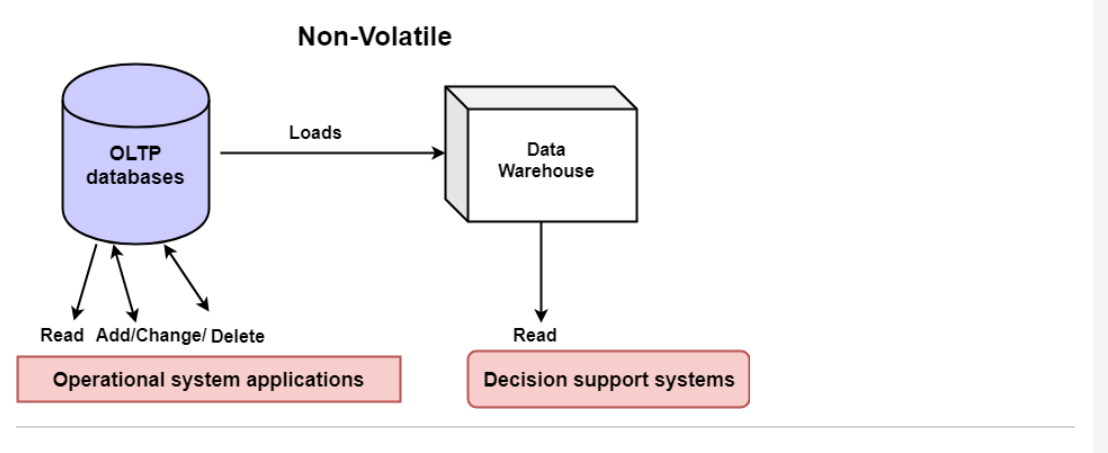
****

Time-Variant

Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse. These variations with a transactions system, where often only the most current file is kept.

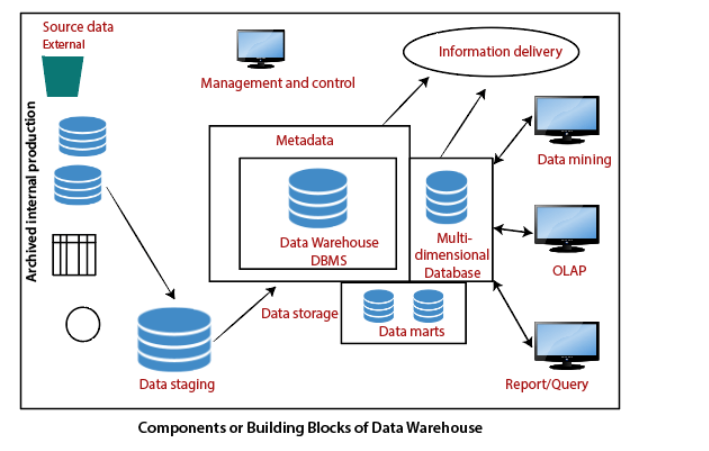
Non-Volatile

The data warehouse is a physically separate data storage, which is transformed from the source operational RDBMS. The operational updates of data do not occur in the data warehouse, i.e., update, insert, and delete operations are not performed. It usually requires only two procedures in data accessing: Initial loading of data and access to data. Therefore, the DW does not require transaction processing, recovery, and concurrency capabilities, which allows for substantial speedup of data retrieval. Non-Volatile defines that once entered into the warehouse, and data should not change.

****

****

# **Components or Building Blocks of Data Warehouse :**



**Source Data Component:**

**Production Data:** This type of data comes from the different operating systems of the enterprise. Based on the data requirements in the data warehouse, we choose segments of the data from the various operational modes.

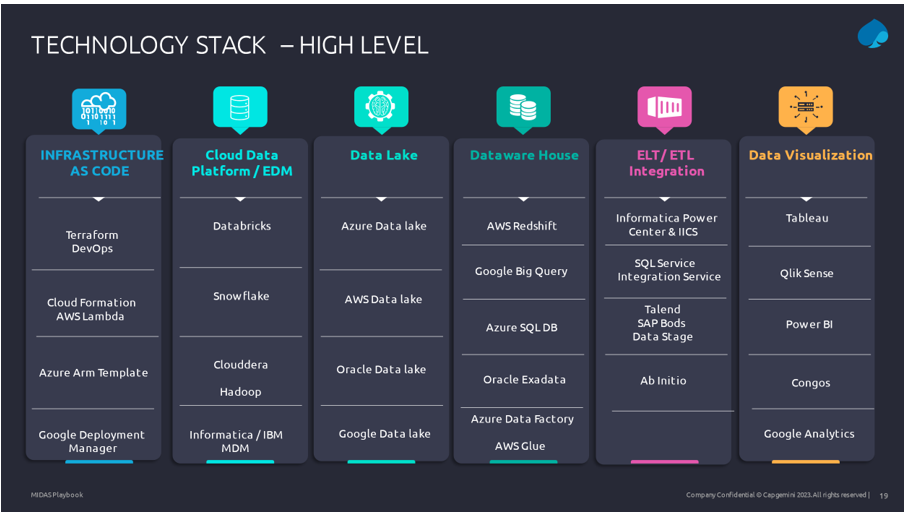
**Internal Data:** In each organization, the client keeps their "**private**" spreadsheets, reports, customer profiles, and sometimes even department databases. This is the internal data, part of which could be useful in a data warehouse.

**Archived Data:** Operational systems are mainly intended to run the current business. In every operational system, we periodically take the old data and store it in achieved files.

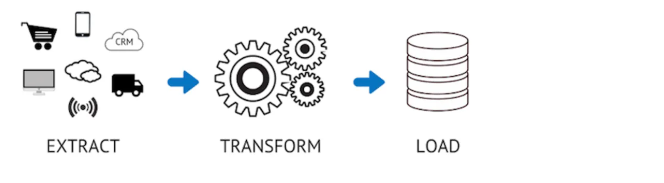
**External Data:** Most executives depend on information from external sources for a large percentage of the information they use. They use statistics associating to their industry produced by the external department.

## Data Staging Component:

## 



# ETL Process in Data Warehouse



**In the world of data warehousing, if you need to bring data from multiple different data sources into one, centralized database, you must first:**

* **EXTRACT data from its original source**
* **TRANSFORM data by deduplicating it, combining it, and ensuring quality, to then**
* **LOAD data into the target database**

**ETL tools enable data integration strategies by allowing companies to gather data from multiple data sources and consolidate it into a single, centralized location. ETL tools also make it possible for different types of data to work together.**

**A typical ETL process collects and refines different types of data, then delivers the data to a data lake or data warehouse such as**[**Redshift**](https://www.talend.com/resources/amazon-redshift/)**,**[**Azure**](https://www.talend.com/resources/microsoft-azure/)**, or [BigQuery](https://www.talend.com/resources/google-bigquery/).**

**How ETL works**

**The ETL process is comprised 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 the 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**
* **Although it can be done manually with a team of data engineers, hand-coded data extraction can be time-intensive and prone to errors. ETL tools automate the extraction process and create a more efficient and reliable workflow.**

**Step 2: Transformation**

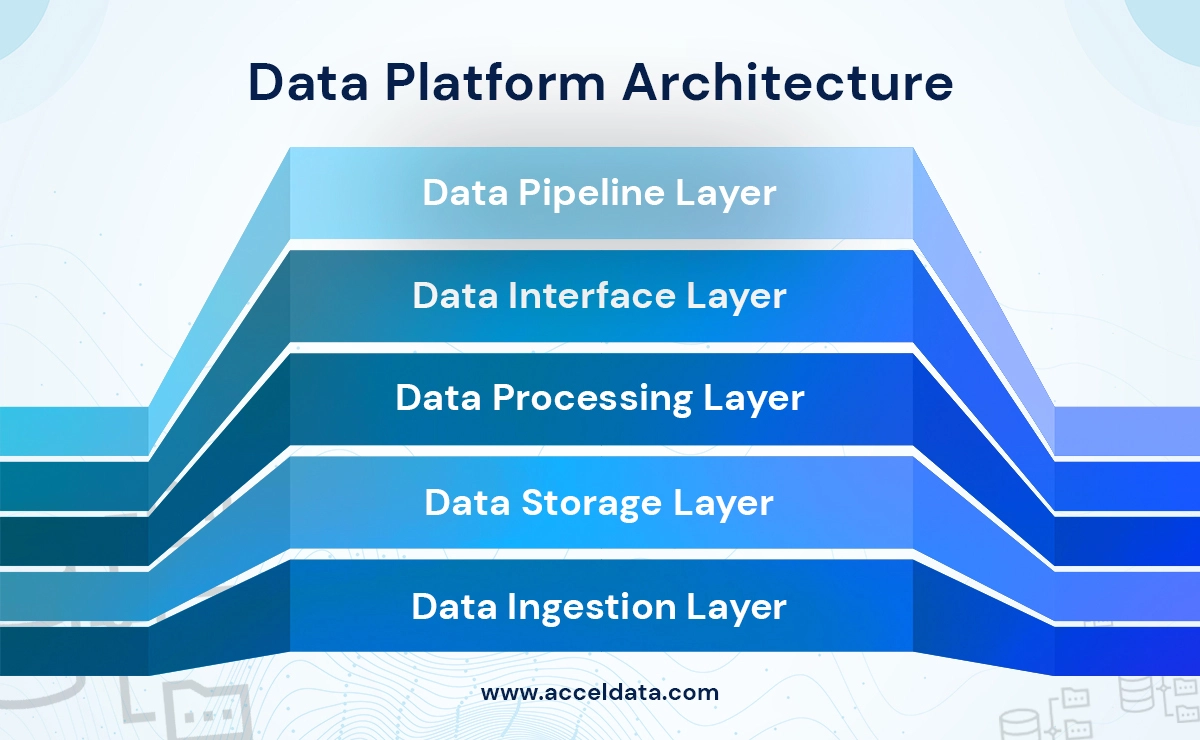
**During this phase of the ETL process, rules and regulations can be applied that ensure data quality and accessibility. You can also apply rules to help your company meet reporting requirements. The process of data transformation is comprised of several sub-processes:**

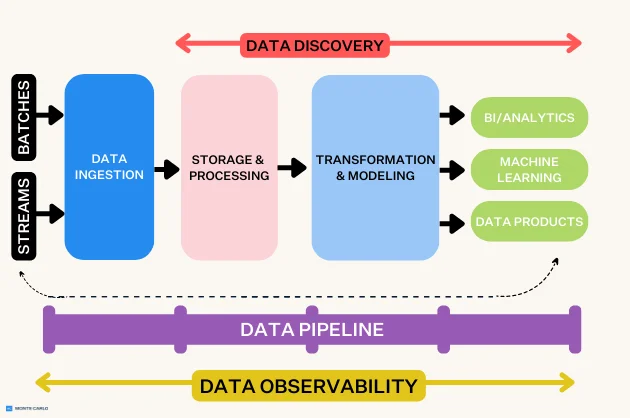
* **Cleansing — inconsistencies and missing values in the data are resolved.**
* **Standardization — formatting rules are applied to the dataset.**
* **Deduplication — redundant data is excluded or discarded.**
* **Verification — unusable data is removed and anomalies are flagged.**
* **Sorting — data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

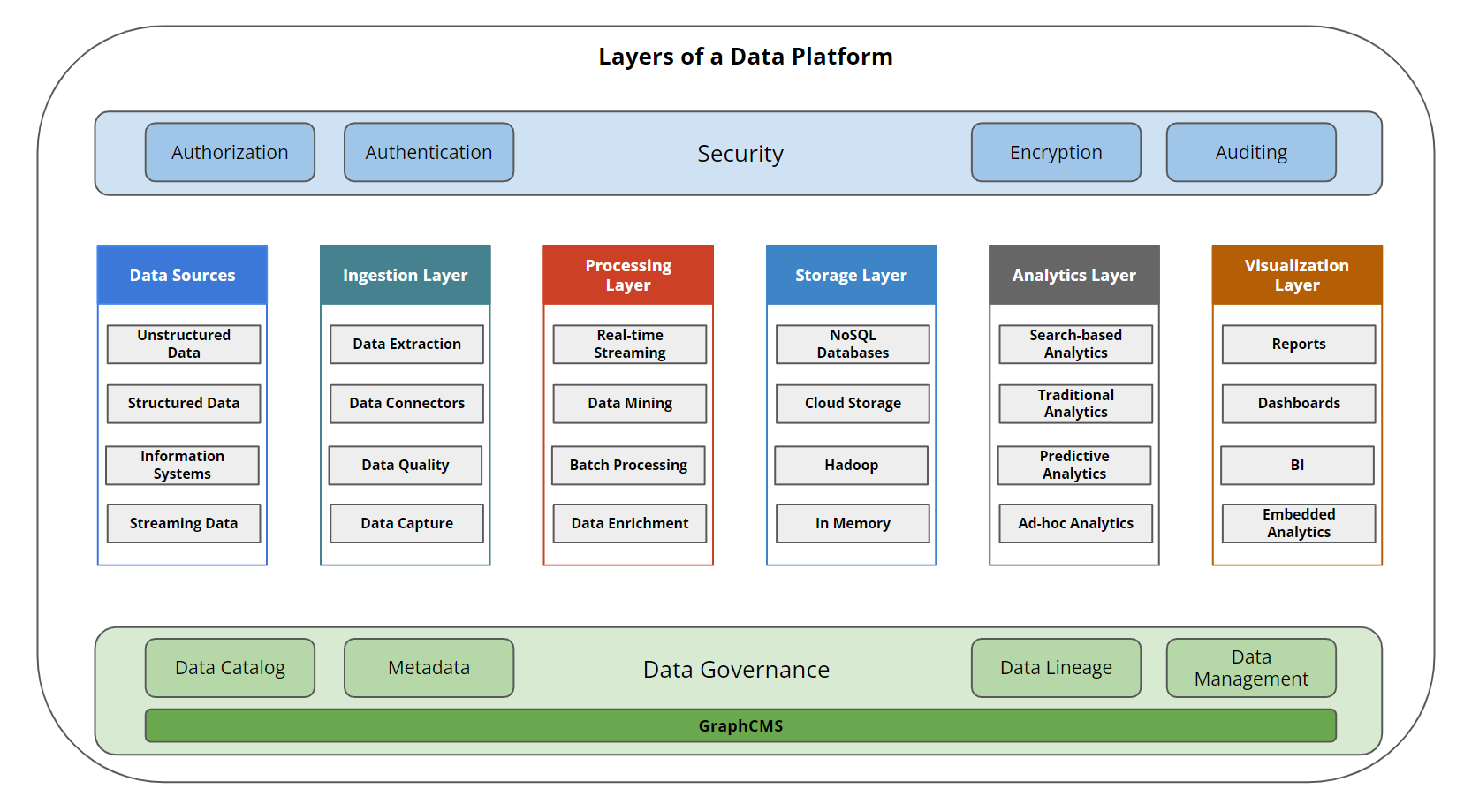
**Transformation is generally considered to be the most important part of the ETL process. Data transformation improves data integrity — removing duplicates and ensuring that raw data arrives at its new destination fully compatible and ready to use.**

**Data Platform Architecture :**

**A data platform architecture diagram shows all the different components and service areas that make up effective data management. Specifically, these diagrams will look different for each organization because each organization deals with different sources and has different data pipelines.**

****

****

****