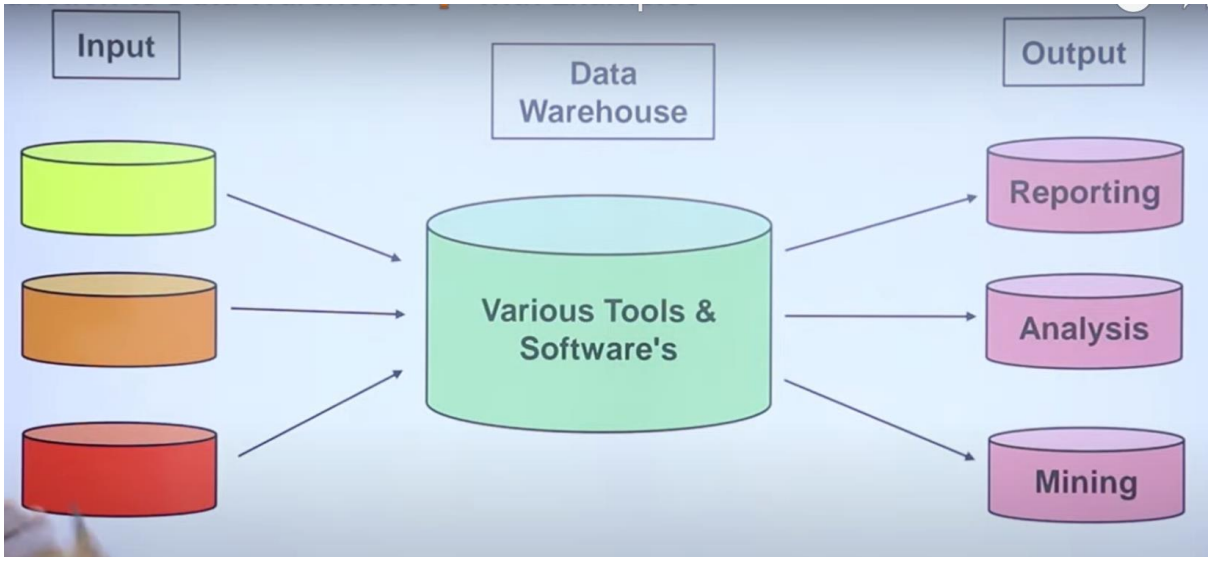


What is ETL?

Extract, transform, and load (ETL) is the process of combining data from multiple sources into a large, central repository called a data warehouse. ETL uses a set of business rules to clean and organize raw data and prepare it for storage, data analytics, and machine learning (ML). You can address specific business intelligence needs through data analytics (such as predicting the outcome of business decisions, generating reports and dashboards, reducing operational inefficiency, and more).

Why is ETL important?

Organizations today have both structured and unstructured data from various sources including:

- Customer data from online payment and customer relationship management (CRM) systems
- Inventory and operations data from vendor systems
- Sensor data from Internet of Things (IoT) devices
- Marketing data from social media and customer feedback
- Employee data from internal human resources systems

By applying the process of extract, transform, and load (ETL), individual raw datasets can be prepared in a format and structure that is more consumable for analytics purposes, resulting in more meaningful insights. For example, online retailers can analyze data from points of sale to forecast demand and manage inventory. Marketing teams can integrate CRM data with customer feedback on social media to study consumer behavior.

How does ETL benefit business intelligence?

Extract, transform, and load (ETL) improves business intelligence and analytics by making the process more reliable, accurate, detailed, and efficient.

Historical context

ETL gives deep historical context to the organization's data. An enterprise can combine legacy data with data from new platforms and applications. You can view older datasets alongside more recent information, which gives you a long-term view of data.

Consolidated data view

ETL provides a consolidated view of data for in-depth analysis and reporting. Managing multiple datasets demands time and coordination and can result in inefficiencies and delays. ETL combines databases and various forms of data into a single, unified view. The data integration process improves the data quality and saves the time required to move, categorize, or standardize data. This makes it easier to analyze, visualize, and make sense of large datasets.

Accurate data analysis

ETL gives more accurate data analysis to meet compliance and regulatory standards. You can integrate ETL tools with data quality tools to profile, audit, and clean data, ensuring that the data is trustworthy.

Task automation

ETL automates repeatable data processing tasks for efficient analysis. ETL tools automate the data migration process, and you can set them up to integrate data changes periodically or even at runtime. As a result, data engineers can spend more time innovating and less time managing tedious tasks like moving and formatting data.

How has ETL evolved?

Extract, transform, and load (ETL) originated with the emergence of relational databases that stored data in the form of tables for analysis. Early ETL tools attempted to convert data from transactional data formats to relational data formats for analysis.

Traditional ETL

Raw data was typically stored in transactional databases that supported many read and write requests but did not lend well to analytics. You can think of it as a row in a spreadsheet. For example, in an ecommerce system, the transactional database stored the purchased item, customer details, and order details in one transaction. Over the year, it contained a long list of transactions with repeat entries for the same customer who purchased multiple items during the year. Given the data duplication, it became cumbersome to analyze the most popular items or purchase trends in that year.

To overcome this issue, ETL tools automatically converted this transactional data into relational data with interconnected tables. Analysts could use queries to identify relationships between the tables, in addition to patterns and trends.

Modern ETL

As ETL technology evolved, both data types and data sources increased exponentially. Cloud technology emerged to create vast databases (also called data sinks). Such data sinks can receive data from multiple sources and have underlying hardware resources that can scale over time. ETL tools have also become more sophisticated and can work with modern data sinks. They can convert data from legacy data formats to modern data formats. Examples of modern databases follow.

Data warehouses

A data warehouse is a central repository that can store multiple databases. Within each database, you can organize your data into tables and columns that describe the data types in the table. The data warehouse software works across multiple types of storage hardware—such as solid state drives (SSDs), hard drives, and other cloud storage—to optimize your data processing.

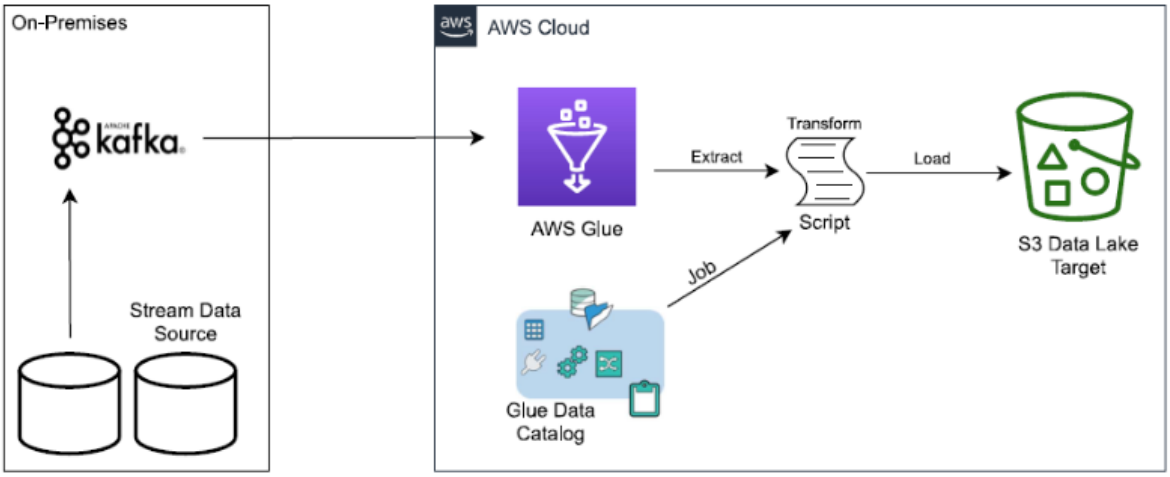
Data lakes

With a data lake, you can store your structured and unstructured data in one centralized repository and at any scale. You can store data as is without having to first structure it based on questions you might have in the future. Data lakes also allow you to run different types of analytics on your data, like SQL queries, big data analytics, full-text search, real-time analytics, and machine learning (ML) to guide better decisions.

How does ETL work?

Extract, transform, and load (ETL) works by moving data from the source system to the destination system at periodic intervals. The ETL process works in three steps:

1. Extract the relevant data from the source database
2. Transform the data so that it is better suited for analytics
3. Load the data into the target database



What is data extraction?

In data extraction, extract, transform, and load (ETL) tools extract or copy raw data from multiple sources and store it in a staging area. A staging area (or landing zone) is an intermediate storage area for temporarily storing extracted data. Data staging areas are often transient, meaning their contents are erased after data extraction is complete. However, the staging area might also retain a data archive for troubleshooting purposes.

How frequently the system sends data from the data source to the target data store depends on the underlying change data capture mechanism. Data extraction commonly happens in one of the three following ways.

Update notification

In update notification, the source system notifies you when a data record changes. You can then run the extraction process for that change. Most databases and web applications provide update mechanisms to support this data integration method.

Incremental extraction

Some data sources can't provide update notifications but can identify and extract data that has been modified over a given time period. In this case, the system checks for changes at periodic intervals, such as once a week, once a month, or at the end of a campaign. You only need to extract data that has changed.

Full extraction

Some systems can't identify data changes or give notifications, so reloading all data is the only option. This extraction method requires you to keep a copy of the last extract to check which records are new. Because this approach involves high data transfer volumes, we recommend you use it only for small tables.

What is data transformation?

In data transformation, extract, transform, and load (ETL) tools transform and consolidate the raw data in the staging area to prepare it for the target data warehouse. The data transformation phase can involve the following types of data changes.

Basic data transformation

Basic transformations improve data quality by removing errors, emptying data fields, or simplifying data. Examples of these transformations follow.

Data cleansing

Data cleansing removes errors and maps source data to the target data format. For example, you can map empty data fields to the number 0, map the data value “Parent” to “P,” or map “Child” to “C.”

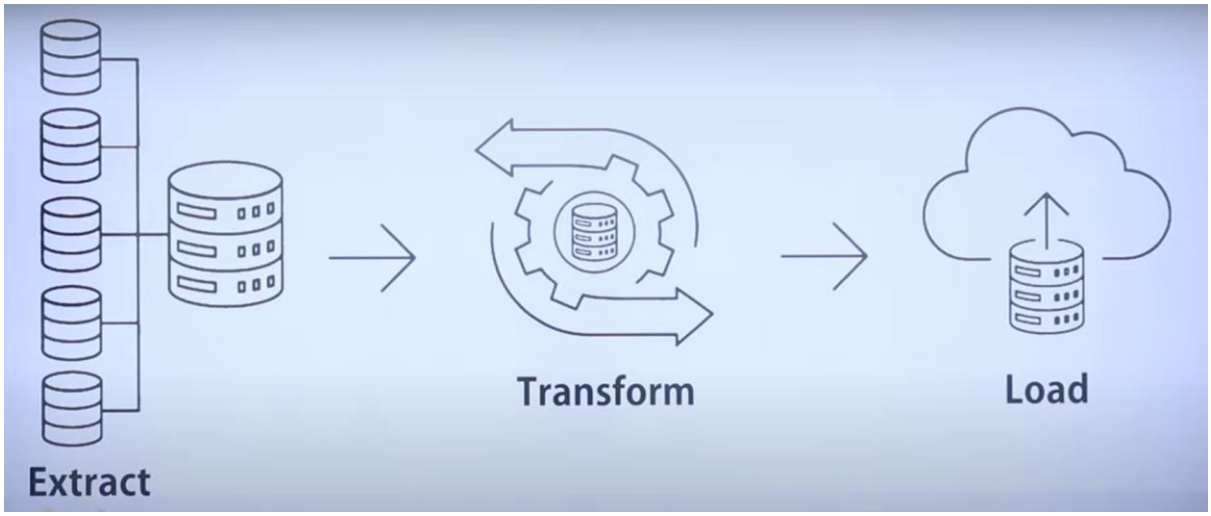
Data deduplication

Deduplication in data cleansing identifies and removes duplicate records.

Data format revision

Format revision converts data, such as character sets, measurement units, and date/time values, into a consistent format. For example, a food company might have different recipe databases with ingredients measured in kilograms and pounds. ETL will convert everything to pounds.

SQL Server Integration Services (SSIS) is a Microsoft SQL Server database built to be a fast and flexible data warehousing tool to perform high-performance data integrations.



What is data loading?

In data loading, extract transform, and load (ETL) tools move the transformed data from the staging area into the target data warehouse. For most organizations that use ETL, the process is automated, well defined, continual, and batch driven. Two methods for loading data follow.

Full load

In full load, the entire data from the source is transformed and moved to the data warehouse. The full load usually takes place the first time you load data from a source system into the data warehouse.

Incremental load

In incremental load, the ETL tool loads the delta (or difference) between target and source systems at regular intervals. It stores the last extract date so that only records added after this date are loaded. There are two ways to implement incremental load.

A data mart is a data storage system that contains information specific to an organization's business unit. It contains a small and selected part of the data that the company stores in a larger storage system. Companies use a data mart to analyze department-specific information more efficiently. It provides summarized data that key stakeholders can use to quickly make informed decisions.

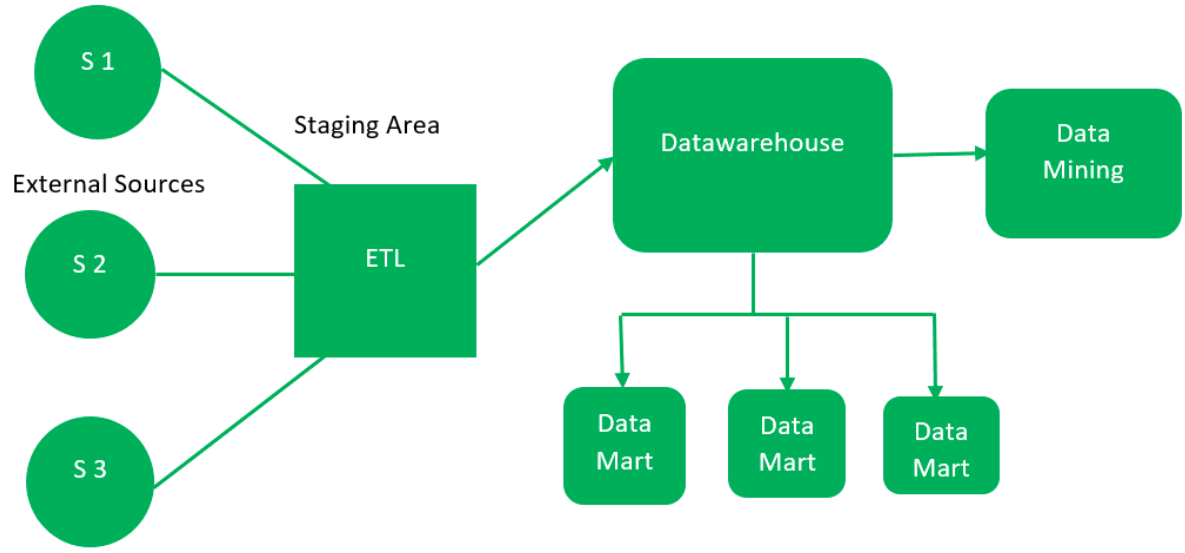
For example, a company might store data from various sources, such as supplier information, orders, sensor data, employee information, and financial records in their data warehouse or data lake. However, the company stores information relevant to, for instance, the marketing department, such as social media reviews and customer records, in a data mart.

Data Warehouse Architecture

Last Updated : 22 Apr, 2023

A **data-warehouse** is a heterogeneous collection of different data sources organised under a unified schema. There are 2 approaches for constructing data-warehouse: Top-down approach and Bottom-up approach are explained as below.

1. Top-down approach:



The essential components are discussed below:

1. **External Sources –**
   External source is a source from where data is collected irrespective of the type of data. Data can be structured, semi structured and unstructured as well.

2. **Stage Area –**
   Since the data, extracted from the external sources does not follow a particular format, so there is a need to validate this data to load into datawarehouse. For this purpose, it is recommended to use **ETL** tool.
   - **E(Extracted):** Data is extracted from External data source.
   - **T(Transform):** Data is transformed into the standard format.
   - **L(Load):** Data is loaded into datawarehouse after transforming it into the standard format.

3. **Data-warehouse –**
   After cleansing of data, it is stored in the datawarehouse as central repository. It actually stores the meta data and the actual data gets stored in the data marts. **Note** that datawarehouse stores the data in its purest form in this top-down approach.

4. **Data Marts –**
   Data mart is also a part of storage component. It stores the information of a particular function of an organisation which is handled by single authority. There can be as many number of data marts in an organisation depending upon the functions. We can also say that data mart contains subset of the data stored in datawarehouse.

5. **Data Mining –**
   The practice of analysing the big data present in datawarehouse is data mining. It is used to find the hidden patterns that are present in the database or in datawarehouse with the help of algorithm of data mining.
   This approach is defined by **Inmon** as – datawarehouse as a central repository for the complete organisation and data marts are created from it after the complete datawarehouse has been created.

Advantages of Top-Down Approach –

1. Since the data marts are created from the datawarehouse, provides consistent dimensional view of data marts.

2. Also, this model is considered as the strongest model for business changes. That's why, big organisations prefer to follow this approach.

3. Creating data mart from datawarehouse is easy.
4. Improved data consistency: The top-down approach promotes data consistency by ensuring that all data marts are sourced from a common data warehouse. This ensures that all data is standardized, reducing the risk of errors and inconsistencies in reporting.
5. Easier maintenance: Since all data marts are sourced from a central data warehouse, it is easier to maintain and update the data in a top-down approach. Changes can be made to the data warehouse, and those changes will automatically propagate to all the data marts that rely on it.
6. Better scalability: The top-down approach is highly scalable, allowing organizations to add new data marts as needed without disrupting the existing infrastructure. This is particularly important

for organizations that are experiencing rapid growth or have evolving business needs.

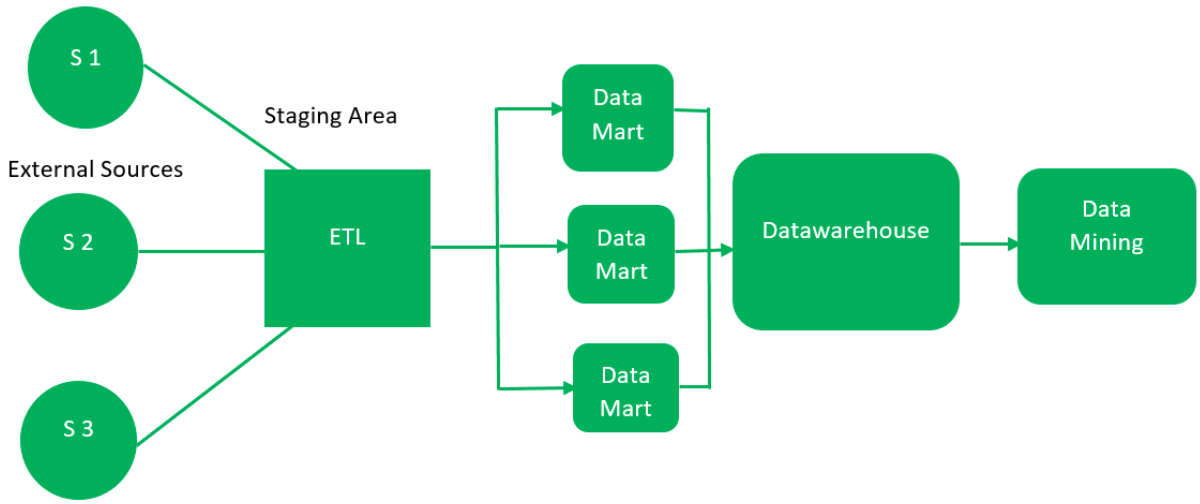
7. Improved governance: The top-down approach facilitates better governance by enabling centralized control of data access, security, and quality. This ensures that all data is managed consistently and that it meets the organization's standards for quality and compliance.
8. Reduced duplication: The top-down approach reduces data duplication by ensuring that data is stored only once in the data warehouse. This saves storage space and reduces the risk of data inconsistencies.
9. Better reporting: The top-down approach enables better reporting by providing a consistent view of data across all data marts. This makes it easier to create accurate and timely reports, which can improve decision-making and drive better business outcomes.
10. Better data integration: The top-down approach enables better data integration by ensuring that all data marts are sourced from a common data warehouse. This makes it easier to integrate data from different sources and provides a more complete view of the organization's data.

Disadvantages of Top-Down Approach –

1. The cost, time taken in designing and its maintenance is very high.
2. Complexity: The top-down approach can be complex to implement and maintain, particularly for large organizations with complex data needs. The design and implementation of the data warehouse and data marts can be time-consuming and costly.
3. Lack of flexibility: The top-down approach may not be suitable for organizations that require a high degree of flexibility in their data reporting and analysis. Since the design of the data warehouse and data marts is pre-determined, it may not be possible to adapt to new or changing business requirements.
4. Limited user involvement: The top-down approach can be dominated by IT departments, which may lead to limited user involvement in the design and implementation process. This can result in data marts that do not meet the specific needs of business users.
5. Data latency: The top-down approach may result in data latency, particularly when data is sourced from multiple systems. This can impact the accuracy and timeliness of reporting and analysis.
6. Data ownership: The top-down approach can create challenges around data ownership and control. Since data is centralized in the data warehouse, it may not be clear who is responsible for maintaining and updating the data.

7. Cost: The top-down approach can be expensive to implement and maintain, particularly for smaller organizations that may not have the resources to invest in a large-scale data warehouse and associated data marts.
8. Integration challenges: The top-down approach may face challenges in integrating data from different sources, particularly when data is stored in different formats or structures. This can lead to data inconsistencies and inaccuracies.

2. Bottom-up approach:



1. First, the data is extracted from external sources (same as happens in top-down approach).

2. Then, the data go through the staging area (as explained above) and loaded into data marts instead of datawarehouse. The data marts are created first and provide reporting capability. It addresses a single business area.

3. These data marts are then integrated into datawarehouse.

This approach is given by **Kinball** as – data marts are created first and provides a thin view for analyses and datawarehouse is created after complete data marts have been created.

Advantages of Bottom-Up Approach –

1. As the data marts are created first, so the reports are quickly generated.

2. We can accommodate more number of data marts here and in this way datawarehouse can be extended.

3. Also, the cost and time taken in designing this model is low comparatively.
4. Incremental development: The bottom-up approach supports incremental development, allowing for the creation of data marts one at a time. This allows for quick wins and incremental improvements in data reporting and analysis.
5. User involvement: The bottom-up approach encourages user involvement in the design and implementation process. Business users can provide feedback on the data marts and reports, helping to ensure that the data marts meet their specific needs.
6. Flexibility: The bottom-up approach is more flexible than the top-down approach, as it allows for the creation of data marts based on specific business needs. This approach can be particularly useful for organizations that require a high degree of flexibility in their reporting and analysis.
7. Faster time to value: The bottom-up approach can deliver faster time to value, as the data marts can be created more quickly than a centralized data warehouse. This can be particularly useful for smaller organizations with limited resources.
8. Reduced risk: The bottom-up approach reduces the risk of failure, as data marts can be tested and refined before being incorporated into a larger data warehouse. This approach can also help to identify and address potential data quality issues early in the process.
9. Scalability: The bottom-up approach can be scaled up over time, as new data marts can be added as needed. This approach can be particularly useful for organizations that are growing rapidly or undergoing significant change.
10. Data ownership: The bottom-up approach can help to clarify data ownership and control, as each data mart is typically owned and managed by a specific business unit. This can help to ensure that data is accurate and up-to-date, and that it is being used in a consistent and appropriate way across the organization.

Disadvantage of Bottom-Up Approach –

1. This model is not strong as top-down approach as dimensional view of data marts is not consistent as it is in above approach.
2. Data silos: The bottom-up approach can lead to the creation of data silos, where different business units create their own data marts without considering the needs of other parts of the organization. This can lead to inconsistencies and redundancies in the data, as well as difficulties in integrating data across the organization.
3. Integration challenges: Because the bottom-up approach relies on the integration of multiple data marts, it can be more difficult to integrate data from different sources and ensure consistency across

the organization. This can lead to issues with data quality and accuracy.

4. Duplication of effort: In a bottom-up approach, different business units may duplicate effort by creating their own data marts with similar or overlapping data. This can lead to inefficiencies and higher costs in data management.
5. Lack of enterprise-wide view: The bottom-up approach can result in a lack of enterprise-wide view, as data marts are typically designed to meet the needs of specific business units rather than the organization as a whole. This can make it difficult to gain a comprehensive understanding of the organization's data and business processes.
6. Complexity: The bottom-up approach can be more complex than the top-down approach, as it involves the integration of multiple data marts with varying levels of complexity and granularity. This can make it more difficult to manage and maintain the data warehouse over time.
7. Risk of inconsistency: Because the bottom-up approach allows for the creation of data marts with different structures and granularities, there is a risk of inconsistency in the data. This can make it difficult to compare data across different parts of the organization or to ensure that reports are accurate and reliable.