ETL and Data Warehouse

Agenda

1. What is ETL?
2. Benefits of ETL Tool.
3. What is Data Warehouse?
4. Benefits of Data Warehousing.
5. Structure of Data warehouse.
6. Architecture of Data Warehouse
7. Data Lakes

While the data warehouse acts as the storage place for all our data and BI tools serve as the mechanism that consumes the data to give us insights, ETL is the intermediary that pushes all of the data from your tech stack and customer tools into the data warehouse for analysis.

The ETL phase is where our business will spend a good chunk of its time and energy when developing a warehouse solution.

What is ETL?

ETL (or Extract, Transform, Load) is a process of data integration that encompasses three steps — extraction, transformation, and loading. In a nutshell, ETL systems take large volumes of raw data from multiple sources, converts it for analysis, and loads that data into your warehouse.

ETL Process

01

Extraction

02

Transform

03

Load data

Extraction

In the first step, extracted data sets come from a source (e.g., Salesforce , Google AdWords, etc.) into a staging area.

The staging area acts as a buffer between the data warehouse and the source data.

The staging area is used for data cleansing and organization. A big challenge during the data extraction process is how your ETL tool handles structured and unstructured data.

Transformation

The data cleaning and organization stage is the transformation stage

All of that data from multiple source systems will be normalized and converted to a single system format. ETL yields transformed data through these methods:

Cleaning, filtering, joining, sorting, splitting, Deduplication, summarization.

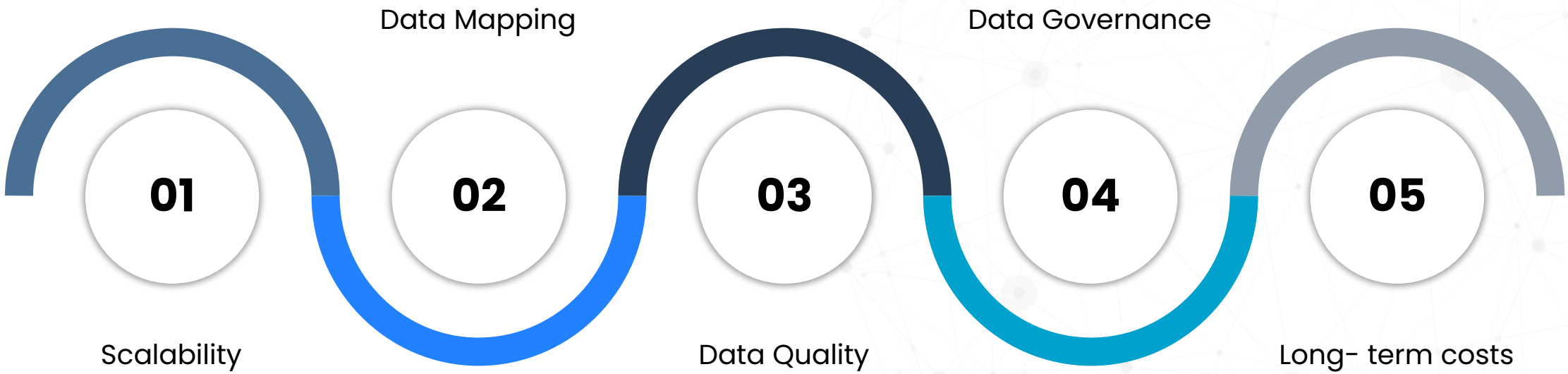
Loading

Finally, data that has been extracted to a staging area and transformed is loaded into your data warehouse

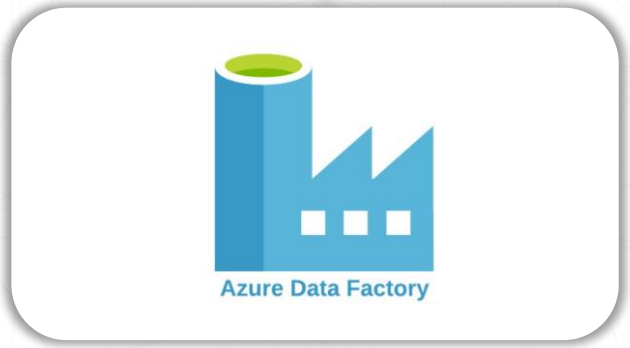
Depending upon your business needs, data can be loaded in batches or all at once.

The exact nature of the loading will depend upon the data source, ETL tools, and various other factors.

Benefits of ETL



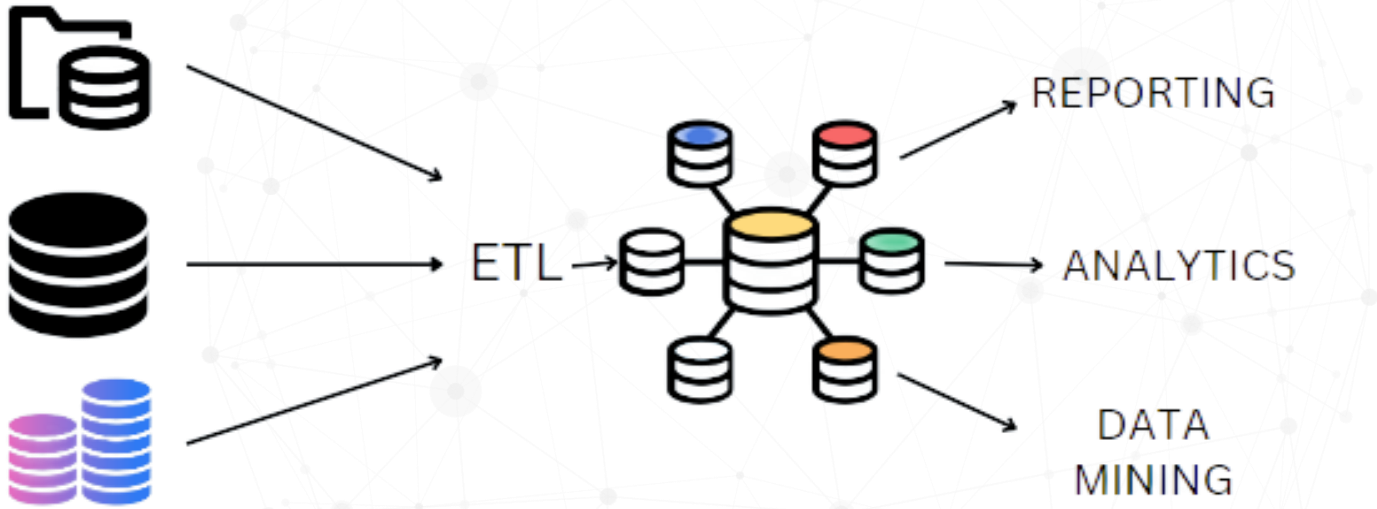
ETL Tools



Data Warehouse

What is Data Warehouse?

- Data warehouse is a large collection of business data that an organization is using to make decisions. The process of constructing and using data warehouse is known as **data warehousing.**
- A data warehouse is specially designed for **data analytics**, which involves reading large amounts of data to understand relationships and trends across the data.



The large amount of data comes from different places such as

- Internal applications (e.G. Marketing, sales, and finance)
- Customer-facing apps
- External partner systems
- And others.

The data warehouse stores the processed data so it's ready for decision makers to access.

This storage system also gives a multi-dimensional view of atomic and summary data.

The important functions which are needed to perform are:

- Data Extraction
- Data Cleaning
- Data Transformation
- Data Loading and Refreshing

Benefits of Data Warehousing

- The benefits of a data warehousing includes **improved data analytics, greater revenue and the ability to compete more strategically in the marketplace**. By efficiently feeding standardized, contextual data to an organization's business intelligence software, a data warehouse drives a more effective data strategy.
- Organizations that use a data warehouse to assist their analytics and business intelligence see a number of substantial benefits:
  - Better Data
  - Faster Decisions

Data Warehouse Structure

- A data warehouse architecture is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise.
- Each data warehouse is different, but all are characterized by standard vital components.

Three Common Architecture are

- Data Warehouse Architecture: Basic
- Data Warehouse Architecture: With Staging Area
- Data Warehouse Architecture: With Staging Area and Data Marts

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.

Why do we need a Data Lake?

- Organizations that successfully generate business value from their data, will outperform their peers.
- A survey saw organizations who implemented a data lake outperforming similar companies by **9% in organic revenue growth.** These leaders were able to do new types of analytics like machine learning over new sources like log files, data from click-streams, social media, and internet connected devices stored in the data lake.
- This helped them to identify, and act upon opportunities for business growth faster by
  - Attracting and retaining customers
  - Boosting productivity
  - Proactively maintaining devices
  - Making informed decisions

Data Lake vs Data Warehouse

Characteristics	Data warehouse	Data lake
Data	Relational from transactional systems, operational database, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (i.e., raw data)
Users	Business analysts	Data scientists, data developers, and business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine learning, predictive analytics, data discovery and profiling

Elements of a Data Lake

Data Movement

Data Lakes allow you to import any amount of data that can come in real-time.

Securely Store

Data Lakes allow you to store relational data like operational databases and data from line of business applications, and non-relational data like mobile apps, IoT devices, and social media.

Analytics

Data Lakes allow various roles in your organization like data scientists, data developers, and business analysts to access data with their choice of analytic tools and frameworks.

Predictive Analytics & ML

Data Lakes will allow organizations to generate different types of insights including reporting on historical data, and doing machine learning where models are built to forecast likely outcomes, and suggest a range of prescribed actions to achieve the optimal result.

Benefits of Data Lake

- Improved customer interactions
- Improve R&D innovation choices
- Increase operational efficiencies