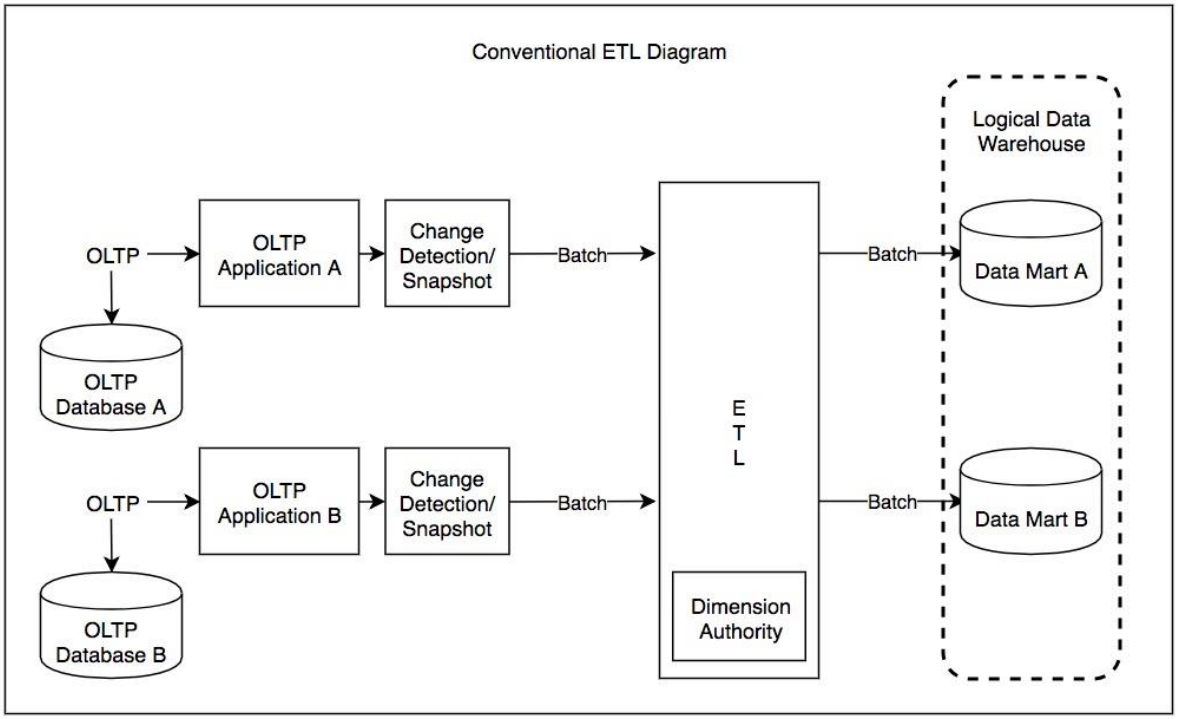
ETL

ETL stands for "Extract, Transform, and Load."



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

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.

Step 3: Loading

The final step in the ETL process is to load the newly transformed data into a new destination (data lake or data warehouse.) Data can be loaded all at once (full load) or at scheduled intervals (incremental load).

- **Full loading** — In an ETL full loading scenario, everything that comes from the transformation assembly line goes into new, unique records in the data warehouse or data repository. Though there may be times this is useful for research purposes, full loading produces datasets that grow exponentially and can quickly become difficult to maintain.

- **Incremental loading** — A less comprehensive but more manageable approach is incremental loading. Incremental loading compares incoming data with what's already on hand, and only produces additional records if new and unique information is found. This architecture allows smaller, less expensive data warehouses to maintain and manage business intelligence.