

ETL, ELT, and Data Pipelines

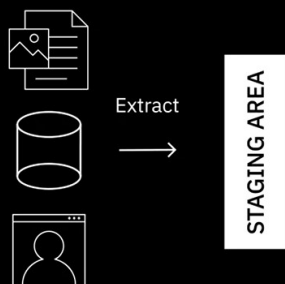
Extract, Transform, and Load Process

Extract, Transform, and Load Process is an automated process which includes:

- Gathering raw data
- Extracting information needed for reporting and analysis
- Cleaning, standardizing, and transforming data into usable format
- Loading data into a data repository



Extract, Transform, and Load Process



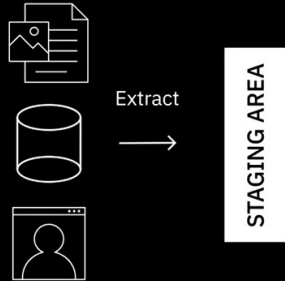
Extraction can be through:

- Batch processing—large chunks of data moved from source to destination at scheduled intervals

BLEND   **Stitch**



Extract, Transform, and Load Process

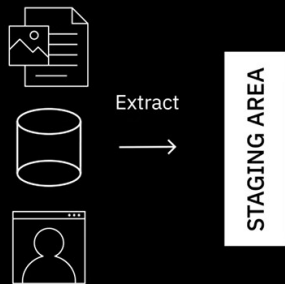


Extraction can be through:

- Stream processing—data pulled in real-time from source, transformed in transit, and loaded into data repository



Extract, Transform, and Load Process

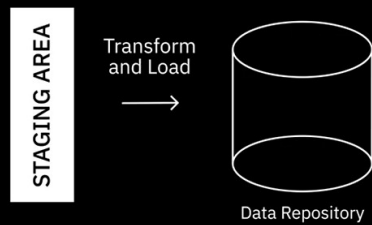


Transforming Data

- Standardizing date formats and units of measurement
- Removing duplicate data
- Filtering out data that is not required
- Enriching data
- Establishing key relationships across tables
- Applying business rules and data validations



Extract, Transform, and Load Process

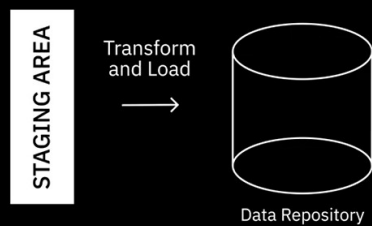


Loading is the transportation of processed data in to a data repository. It can be

- Initial loading—populating all of the data in the repository
- Incremental loading—applying updates and modifications periodically
- Full refresh—erasing a data table and reloading fresh data



Extract, Transform, and Load Process



Load Verification includes checks for:

- Missing or null values
- Server performance
- Load failures



Extract, Transform, and Load Process



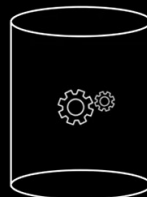
Extract, Load, and Transform Process



Extract & Load



Transformations



Data
Repository



Data Lake



Data
Warehouse



Extract, Load, and Transform Process

Advantages:

- Shortens the cycle between extraction and delivery
- Allows you to ingest volumes of raw data as immediately as the data becomes available
- Affords greater flexibility to analysts and data scientists for exploratory data analytics
- Transforms only that data which is required for a particular analysis so it can be leveraged for multiple use cases



Data Pipelines

A Data Pipeline

- Can be used for both batch and streaming data
- Supports both long-running batch queries and smaller interactive queries
- Typically loads data into a data lake but can also load data into a variety of target destinations—including other applications and visualization tools



Data Flow



