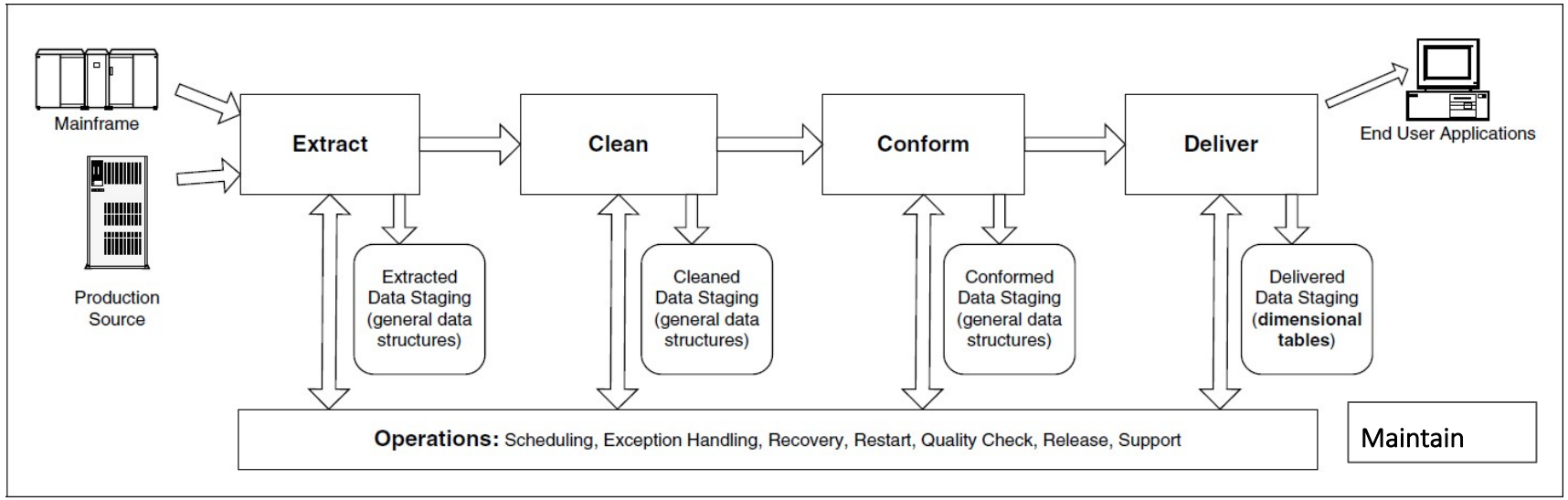
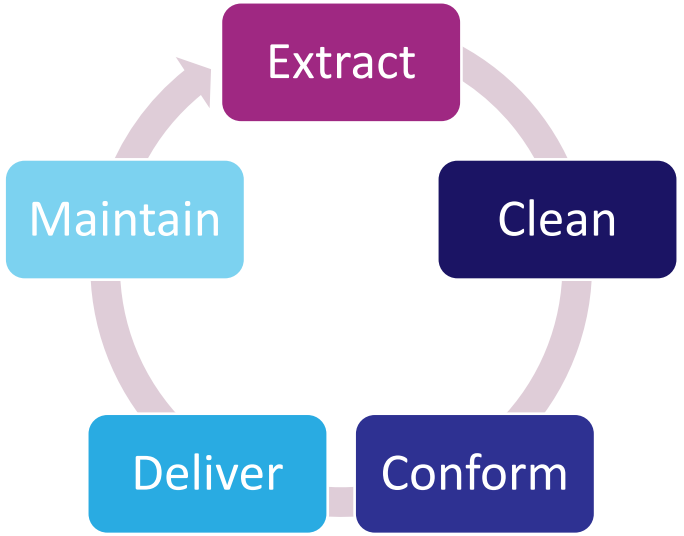
ETL Overview

Katz
Katz School
of Science and Health

Extract Transform Load (ETL)

- ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL full-form is Extract, Transform and Load.
- In a typical data warehouse solution, it is used to extract, clean, conform, and deliver the data and includes activities such as scheduling, exception handling, updates, recovery, restart, quality assurance, releases and notifications and support.
- Approximately 70% of the effort is dedicated to the ETL Process





Why do we do it?

- Integrate data from multiple data sources
- To manage query and transactions of data.
- Definitions and adoptions of data, metadata and their storages.
- Accessing the data seamlessly.
- Transparency, support for heterogeneity, extensibility and scalability.

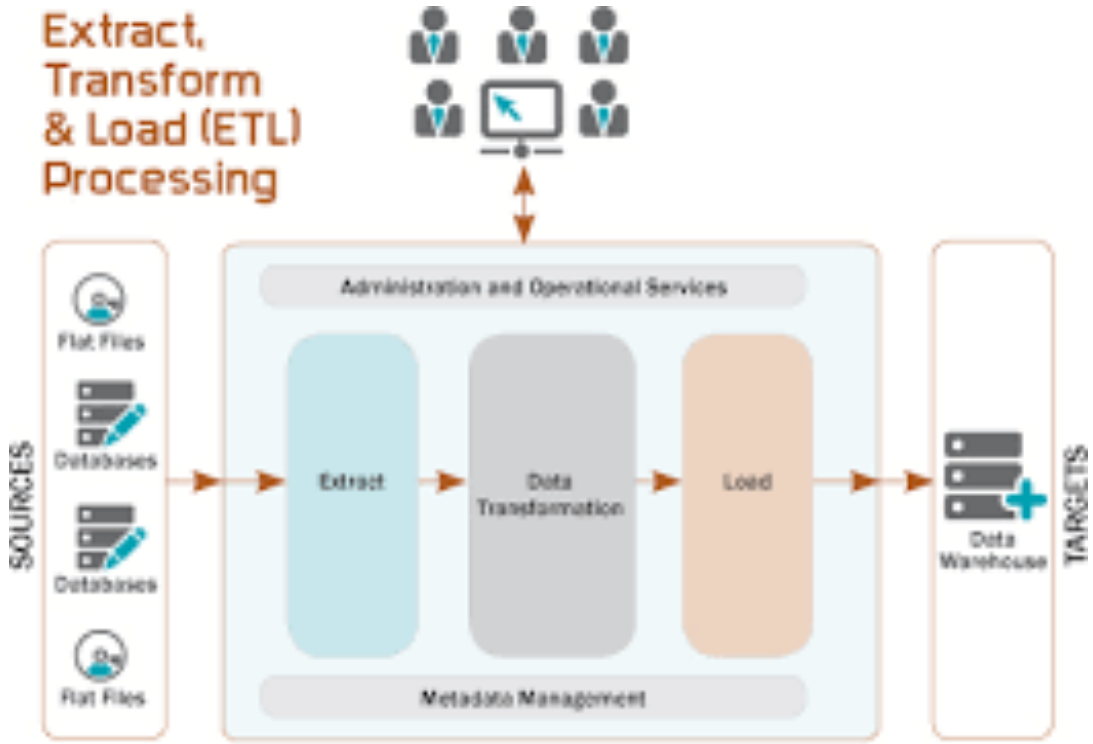
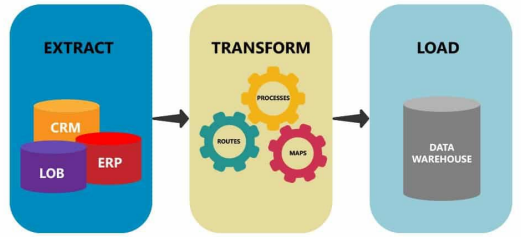


FIGURE 1: BASIC ETL FUNCTIONALITY

Uses of ETLs

ETL can be used for many types of data manipulation and does not necessarily have to be strictly enforced for use in a data warehouse:

- **Data Migration**: Process of transferring data between storage types or formats. An *automated migration* frees up human resources from tedious tasks. Design, extraction, cleansing, load and verification are done for moderate to high complexity jobs.
- **Data Consolidation**: Usually associated with moving data from remote locations to a central location or combining data.
- **Data Integration**: Process of combining data residing at different sources and providing a unified view. Emerges in both commercial and scientific fields and is focus of extensive theoretical work. Also referred to as *Enterprise Information Integration*.



ETL - Extract, Transform, Load

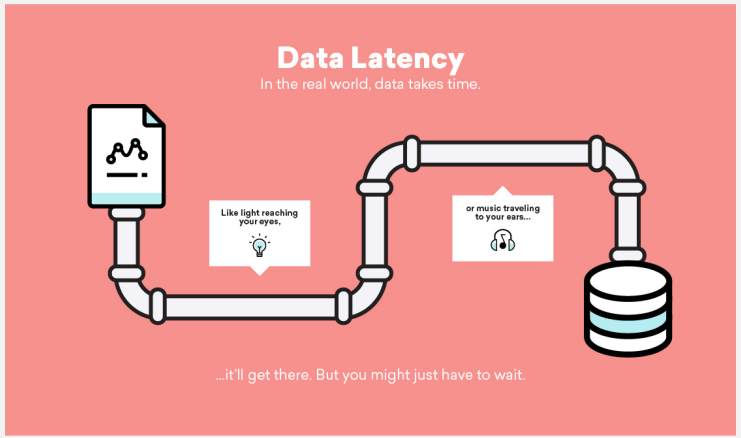
- **Master Data Management**: Processes and tools to define and manage non-transactional data. Provides for collecting, aggregating, matching, consolidating, quality-assuring, persisting and distributing data to an organization to ensure consistency and control.
- **Data Warehouse**: Repository of electronically stored data. ETL facilitates populating, reporting and analysis. Includes business intelligence as well as metadata retrieval and management tools.
- **Data Synchronization**: Process of making sure two or more locations contain the same up-to-date files. Add, change, or delete a file from one location, synchronization will mirror the action at the new location.

ETL Requirements

- **Business needs.** information requirements of the consumers of the solution that may not be explicit during the elicitation of those requirements. You will need…
- **Compliance.** Storing and retaining archives, transactional flor and changed data, documentation of algorithms and adjustments, proof of security of the data copies are important compliance considerations around ETL development
- **Data quality via Data Profiling.** Analytical method for looking at data for the purpose of developing a thorough understanding of the content, structure, and quality. Can reveal issues with the source database

Data profiling activities

- Collecting descriptive statistics like min, max, count and sum.
- Collecting data types, length and recurring patterns.
- Tagging data with keywords, descriptions or categories.
- Performing data quality assessment, risk of performing joins on the data.
- Discovering metadata and assessing its accuracy.
- Identifying distributions, key candidates, foreign-key candidates, functional dependencies, embedded value dependencies, and performing inter-table analysis.



Data latency is a challenge to manage because many source data systems may be available to provide data at different frequencies. The ETL must handle these varying frequencies

ETL Requirements

- Security. Often need to publish data to broad audience at odds with security need to restrict access. LDAp, access controls, IAM, are all security measures to control access to data.
- Data integration and 360 degree view. Usually results in conforming dimensions and facts. Full view of the customer.
- Data latency. Determines how quickly the data must be delivered to business users.
- Archived and Lineage. All staged data should be archived unless a decision is made not to recover it in the future.
- Bi Delivery. Interface with the BI Application.
- Skillsets. Consider the team at hand and the resources available.
- Legacy licenses. Think about software dependencies and legacy systems