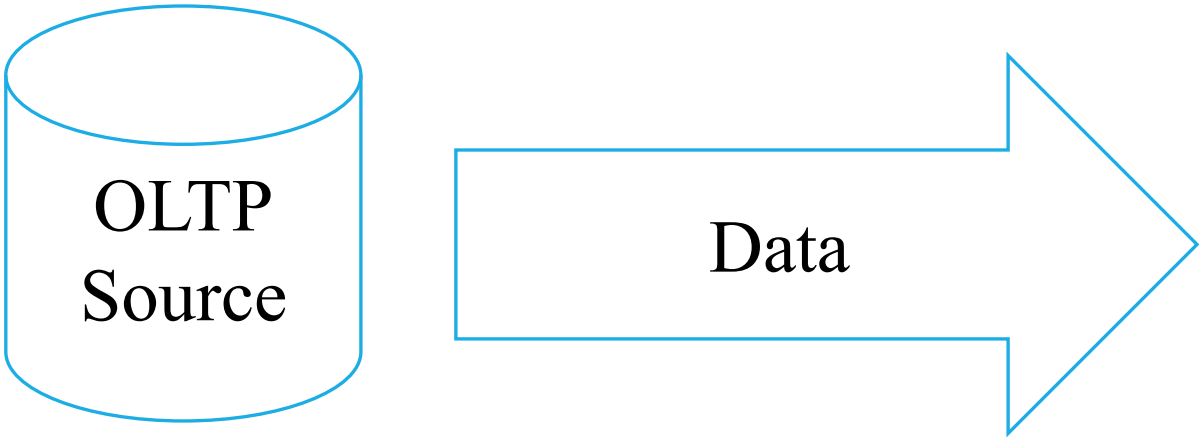
Data Extraction

School of Information Studies
Syracuse University

Kimball: Four Major ETL Operations

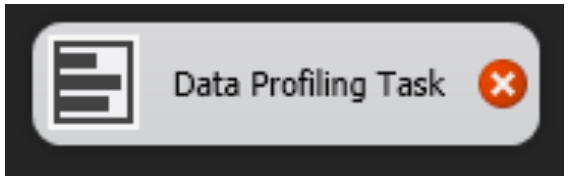
1. **Extract** the data from its source.
2. **Cleanse and Conform** to improve data accuracy and quality (transform).
3. **Deliver** the data into the presentation server (load).
4. **Manage** the ETL process itself.

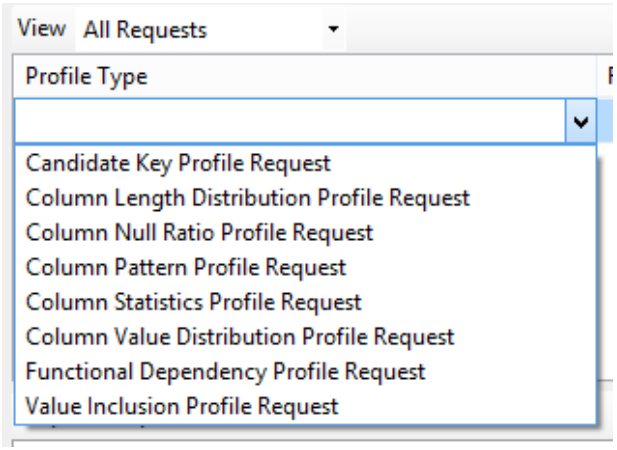
Data Extraction Subsystems



Data Profiling

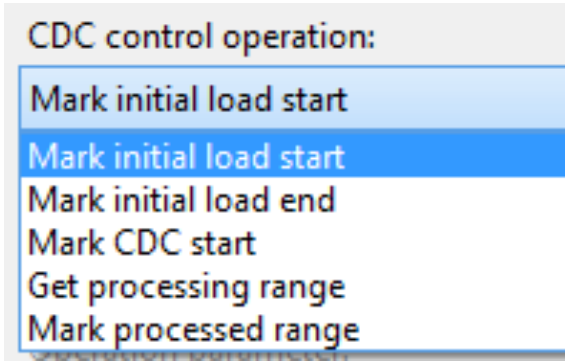
- Helps you to understand the source data.
  - Identify candidate business keys
  - Functional dependencies
  - Nulls
  - Etc.
- Helps us figure out the facts, dimensions, and source-to-target mapping.
- Valuable tool when you do not have the SQL chops to query the source data.





Change Data Capture System

- A means to detect which data are part of the **incremental load** (selective processing).
- Difficult to get right, needs a lot of testing.
- Common approaches:
  - **Audit columns** in source data (last update)
  - **Timed extracts** (e.g., yesterday's records)
  - **Diff compare** with CRC /Hash
  - **Database transactions logs**
  - **Triggers/message queues**



Extract System

- Getting data from the source system—a fundamental component!
- Two methods:
  - **File:** extracted output from a source system. Useful with third parties/legacy systems.
  - **Stream:** initiated data flows out of a system: middleware query, web service.
- Files are useful because they provide restart points without **requerying the source**.