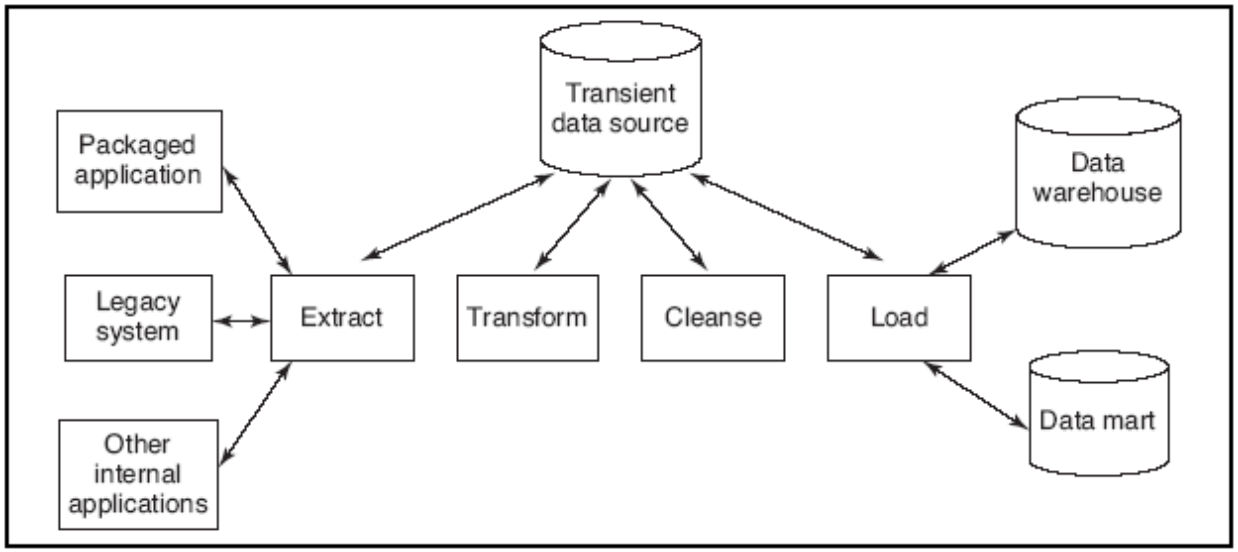
Lection 5

Definition

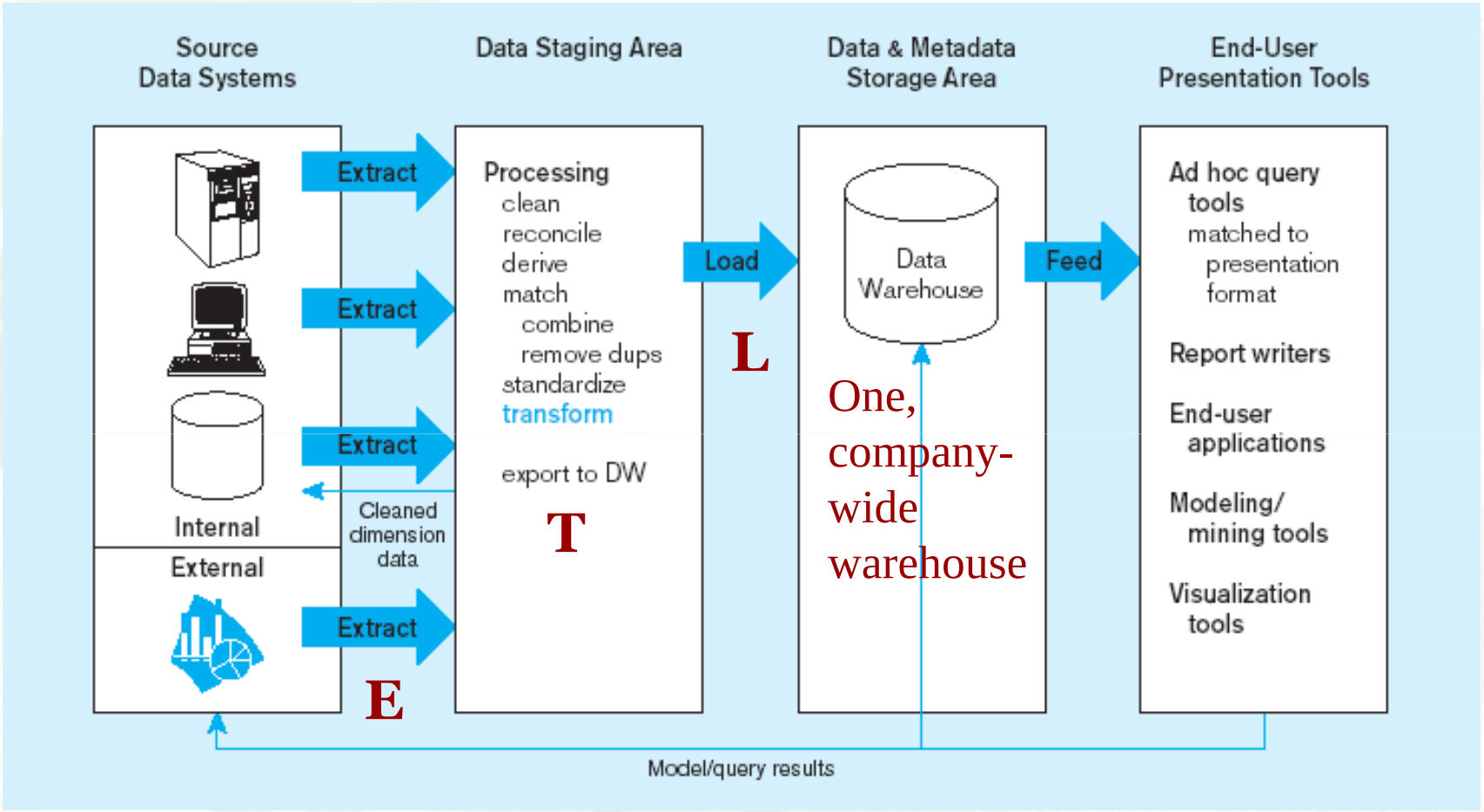
- Extraction Transformation Loading – ETL
- Get data out of sources and load into the DW
  - Data is extracted from OLTP database, transformed to match the DW schema and loaded into the data warehouse database
  - May also incorporate data from non-OLTP systems such as text files, legacy systems, and spreadsheets



Introduction

- ETL is a complex combination of process and technology
  - The **most underestimated process in DW development**
  - The **most time-consuming process in DW development**
    - Often, 80% of development time is spent on ETL

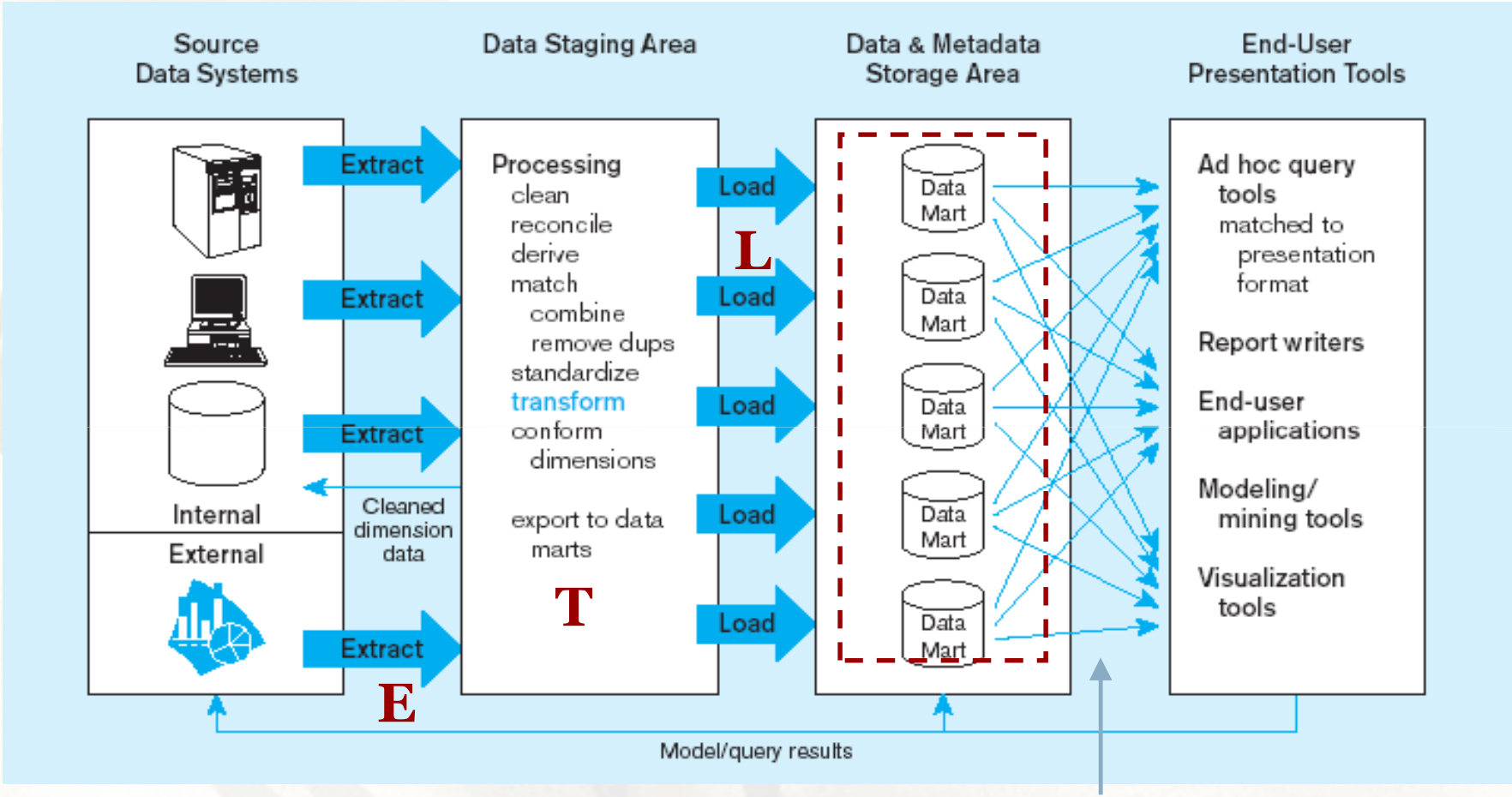
ETL for generic DW architecture



Periodic extraction ➔ data is not completely fresh in warehouse

Independent data marts

Data marts:
Mini-warehouses, limited in scope

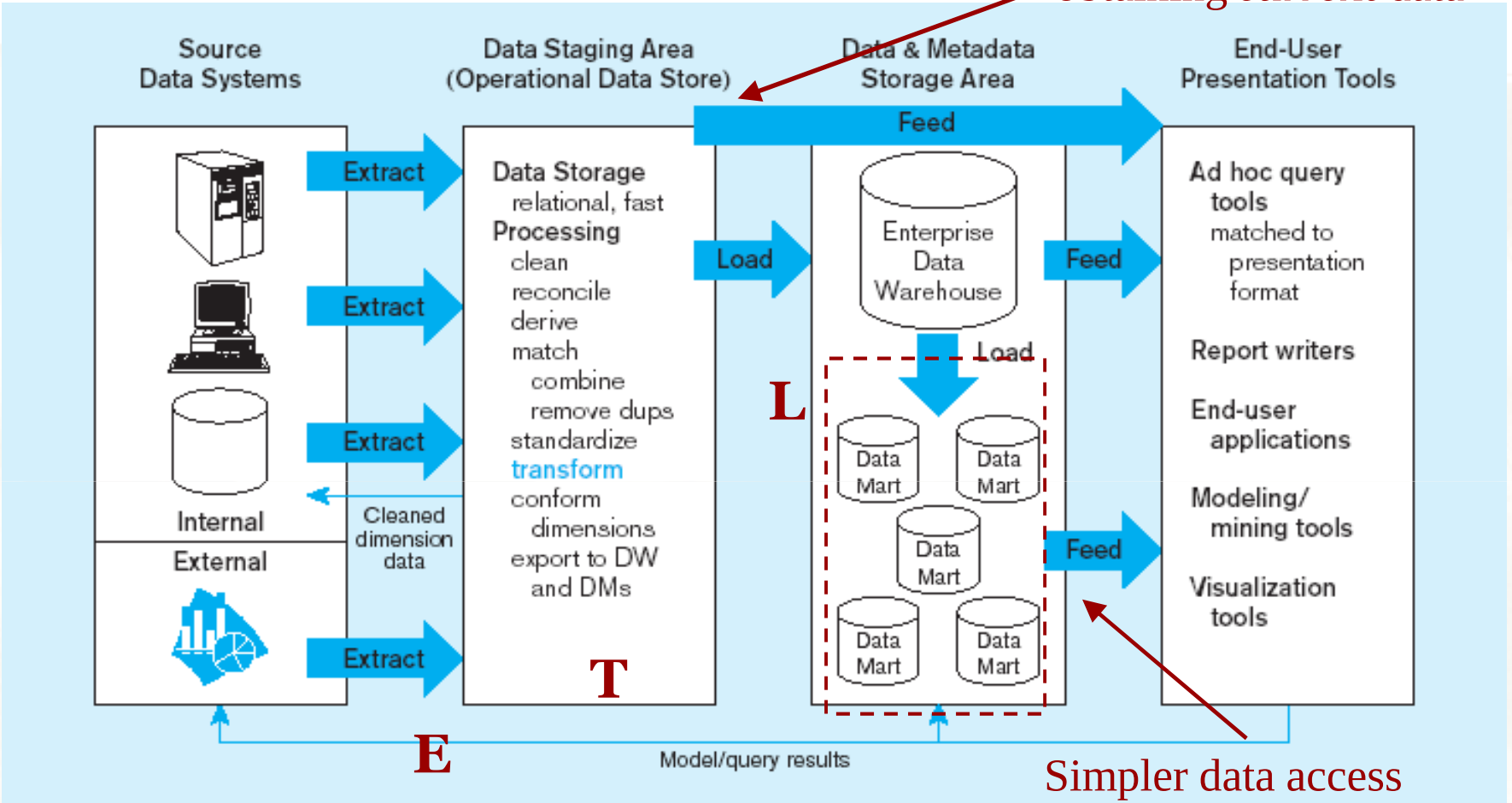


Separate ETL for each ***independent*** data mart

Data access complexity due to ***multiple*** data marts

Dependent data marts

ODS provides option for obtaining ***current*** data



Single ETL for ***enterprise data warehouse (EDW)***

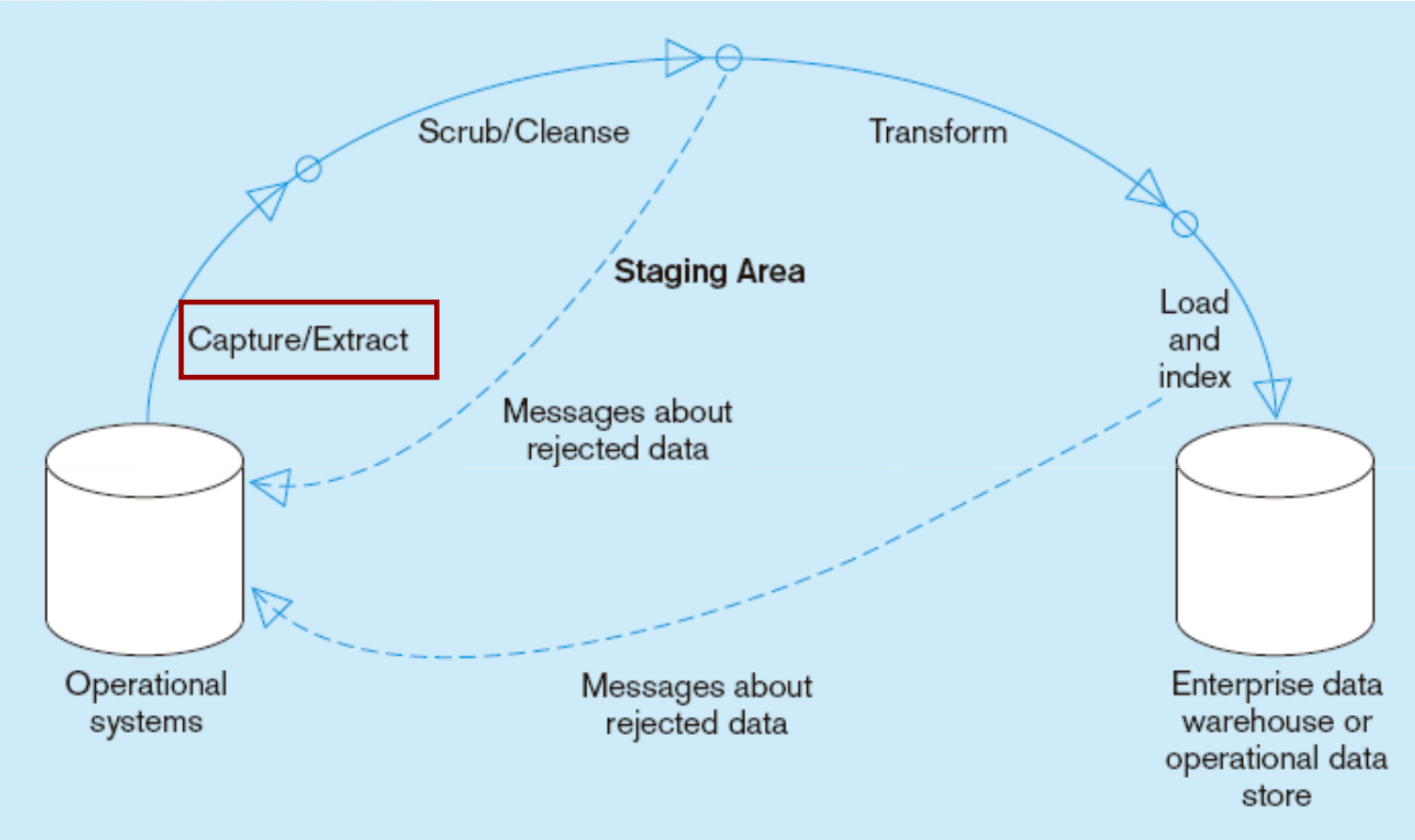
Dependent data marts loaded from EDW

Data Staging Area (DSA)

- Transit storage for data underway in the ETL process
  - Creates a logical and physical separation between the source systems and the data warehouse
  - Transformations/cleansing done here
- No user queries (some do it)
- Sequential operations (few) on large data volumes
  - Performed by central ETL logic
  - Easily restarted
  - RDBMS or flat files? (DBMS have become better)
- Allows centralized backup/recovery
  - Often too time consuming to initial load all data marts by failure
  - Thus, backup/recovery facilities needed

Extract

Capture/Extract…obtaining a snapshot of a chosen subset of the source data for loading into the data warehouse



Static extract = capturing a snapshot of the source data at a point in time

Incremental extract = capturing changes that have occurred since the last static extract

Types of Data Sources

- Extraction is very dependent on the source types
- Non-cooperative sources
  - Snapshot sources –provides only full copy of source
  - Specific sources –each is different, e.g., legacy systems
  - Logged sources –writes change log (DB log)
  - Queryable sources –provides query interface, e.g., SQL
- Cooperative sources
  - Replicated sources –publish/subscribe mechanism
  - Call back sources –calls external code (ETL) when changes occur
  - Internal action sources –only internal actions when changes occur
    - DB triggers is an example

Incremental extraction

- Information changes at the sources
  - Need to update the DW
  - Cannot apply Extract/ETL from scratch
    - Extract from source systems can take a long time (days/weeks)
    - Drain on the operational systems and on DW
- Incremental Extract/ETL
  - Extract only changes since last load
  - A number of methods can be used

Timestamps

- Put update timestamp on all rows
  - Updated by DB trigger
  - Extract only where ”timestamp > time for last extract”
- Evaluation:
  - \+ Reduces extract time
  - \+ Less operational overhead
  - \- Source system must be changed

Messages

- Applications insert messages in a ”queue” at updates

- Evaluation:
  - + Works for all types of updates and systems
  - - Operational applications must be changed and incurred operational overhead

DB triggers

- Triggers execute actions on INSERT, UPDATE, DELETE

- Evaluation:
  - + Operational applications need not be changed (but need to set the triggers -)
  - + Enables real-time update of DW
  - - Operational overhead

DB log

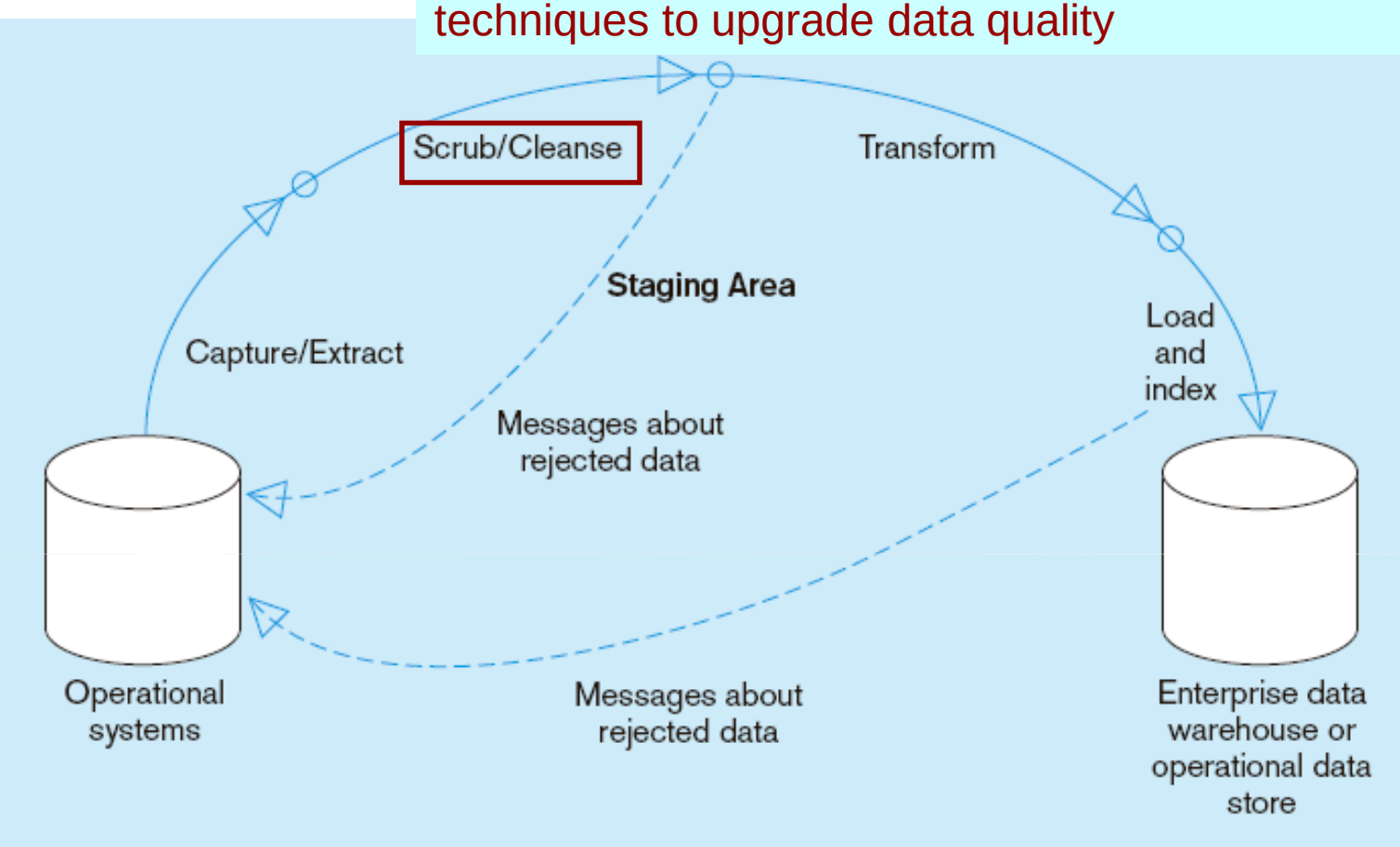
- Replication based on DB log
  - Find changes directly in DB log which is written anyway
- Evaluation:
  - \+ Operational applications need **not be changed**
  - \+ No operational overhead
  - \- Not possible in some DBMS (SQL Server, Oracle, DB2 can do it)

Transform

- Main step where the ETL adds value
- Changes data to be inserted in DW
- 3 major steps
  - Data Cleansing
  - Data Integration
  - Other Transformations (includes replacement of codes, derived values, calculating aggregates)

Cleanse

Scrub/Cleanse…uses pattern recognition and AI techniques to upgrade data quality



Fixing errors: misspellings, erroneous dates, incorrect field usage, mismatched addresses, missing data, duplicate data, inconsistencies

Also: decoding, reformatting, time stamping, conversion, key generation, merging, error detection/logging, locating missing data

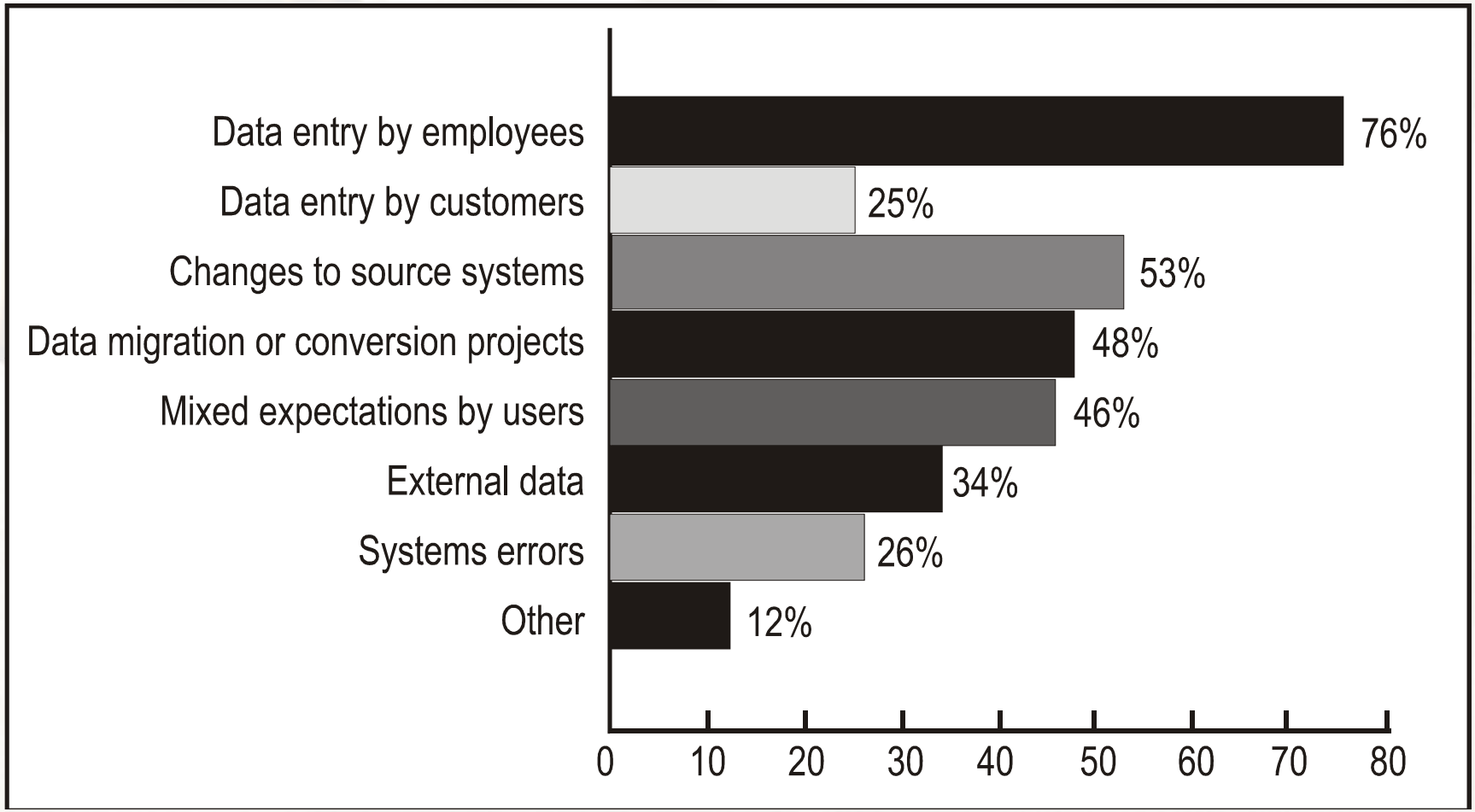
Examples of "Dirty" Data

- Dummy Values
  - a clerk enters 999-99-9999 as a SSN rather than asking the customer, customers 235 years old
- Absence of Data
  - NULL for customer age
- Cryptic Data
  - 0 for male, 1 for female
- Contradicting Data
  - customers with different addresses
- Violation of Business Rules
  - NULL for customer name on an invoice
- Reused Primary Keys, Non-Unique Identifiers
  - a branch bank is closed. Several years later, a new branch is opened, and the old identifier is used again
- Data Integration Problems
  - Multipurpose Fields, Synonyms

What Causes Poor Data Quality?

- Factors contributing to poor data quality:
  - Business rules do not exist or there are no standards for data capture.
  - Standards may exist but are not enforced at the point of data capture.
  - Inconsistent data entry (incorrect spelling, use of nicknames, middle names, or aliases) occurs.
  - Data entry mistakes (character transposition, misspellings, and so on) happen.
  - Integration of data from systems with different data standards is present.
  - Data quality issues are perceived as time-consuming and expensive to fix.

Primary Sources of Data Quality Problems



Source: *The Data Warehousing Institute, Data Quality and the Bottom Line, 2002*

Actions for Cleansing

- "Data stewards" responsible for data quality
  - A given steward has the responsibility for certain tables
    - Includes manual inspections and corrections!
- Cleansing tools:
  - Special-purpose cleansing (addresses, names, emails, etc)
  - AI and data mining cleansing tools
    - Rule-based cleansing (if-then style)
    - Automatic rules
      - Guess missing sales person based on customer and item
- Do not fix all problems with data quality
  - Allow management to see "weird" data in their reports?

Typical Cleansing Tasks

- Standardization
  - changing data elements to a common format defined by the data steward
- Gender Analysis
  - returns a person's gender based on the elements of their name (returns M(ale), F(emale) or U(nknown))
- Entity Analysis
  - returns an entity type based on the elements in a name (i.e. customer names) (returns O(rganization), I(ndividual) or U(nknown))
- De-duplication
  - uses match coding to discard records that represent the same person, company or entity
- Matching and Merging
  - matches records based on a given set of criteria and at a given sensitivity level. Used to link records across multiple data sources.

Typical Cleansing Tasks

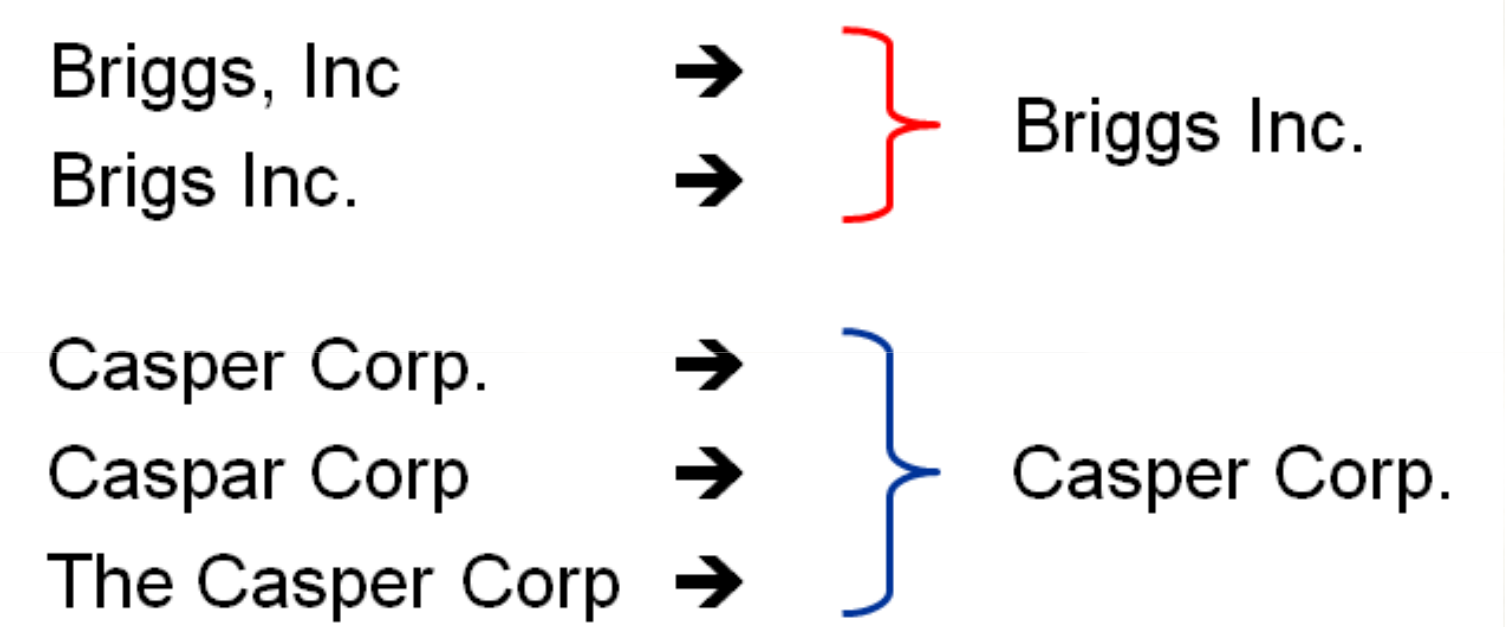
- Address Verification
  - verifies that if you send a piece of mail to a given address, it will be delivered. Does not verify occupants, just valid addresses.
- Householding/Consolidation
  - used to identify links between records, such as customers living in the same household or companies belonging to the same parent company.
- Parsing
  - identifies individual data elements and separates them into unique fields
- Casing
  - provides the proper capitalization to data elements that have been entered in all lowercase or all uppercase.

Analysis and Standardization Example

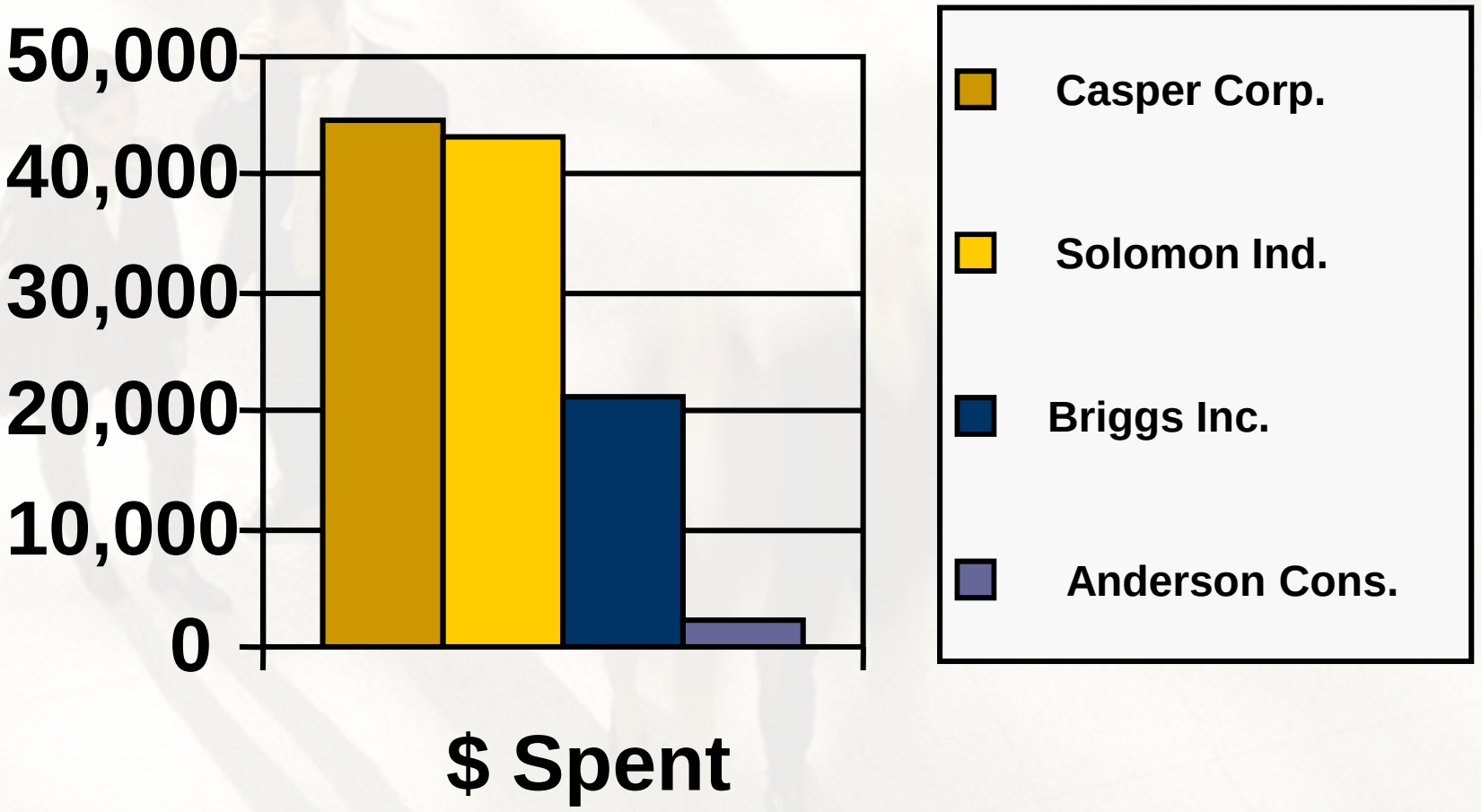
Who is the biggest supplier?

Anderson Construction	$ 2,333.50
Briggs,Inc	$ 8,200.10
Brigs Inc.	$12,900.79
Casper Corp.	$27,191.05
Caspar Corp	$ 6,000.00
Solomon Industries	$43,150.00
The Casper Corp	$11,500.00

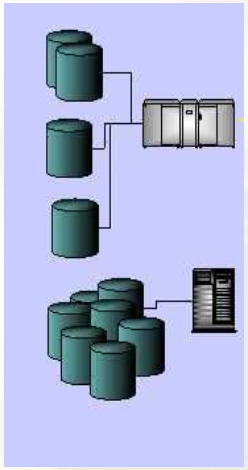
Analysis and Standardization Example



Analysis and Standardization Example



Data Matching Example



Operational System of Records

Mark Carver
SAS
SAS Campus Drive
Cary, N.C.

Mark W. Craver
Mark.Craver@sas.com

Mark Craver
Systems Engineer
SAS

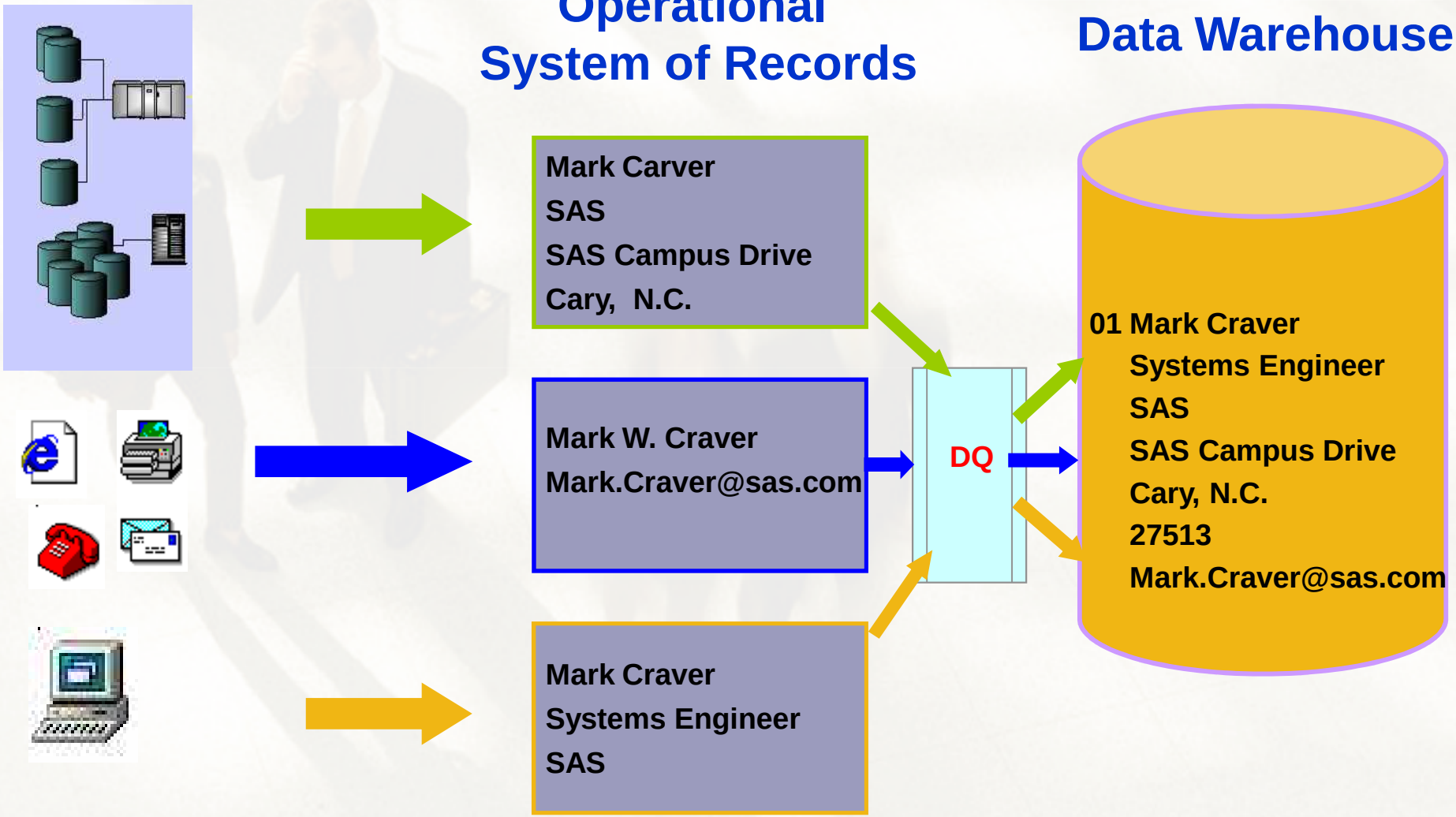
Data Warehouse

01 Mark Carver
SAS
SAS Campus Drive
Cary, N.C.

02 Mark W. Craver
Mark.Craver@sas.com

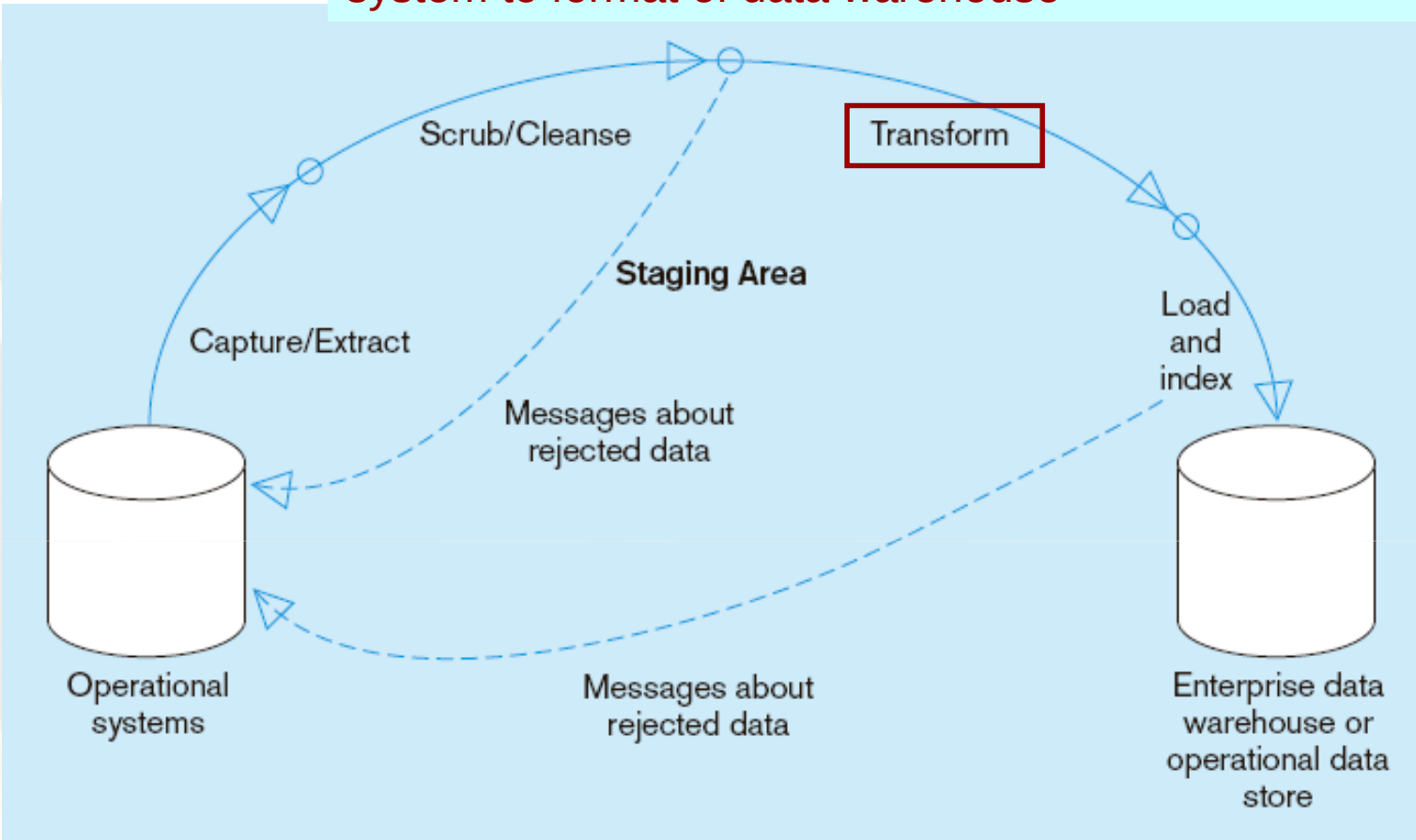
03 Mark Craver
Systems Engineer
SAS

Data Matching Example



Transform

Transform = convert data from format of operational system to format of data warehouse



Record-level:

Selection–data partitioning
Joining–data combining
Aggregation–data summarization

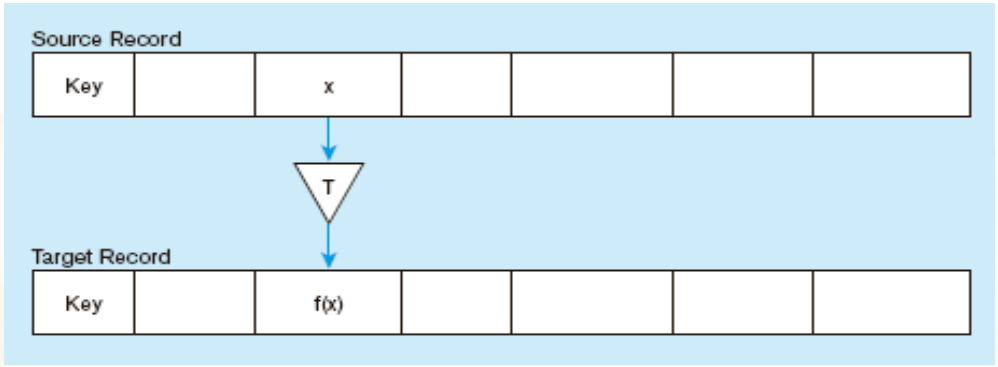
Field-level:

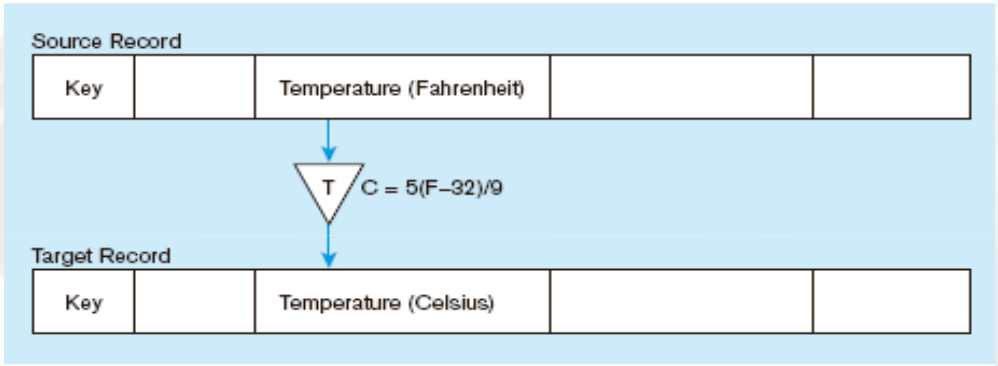
single-field–from one field to one field
multi-field–from many fields to one, or one field to many

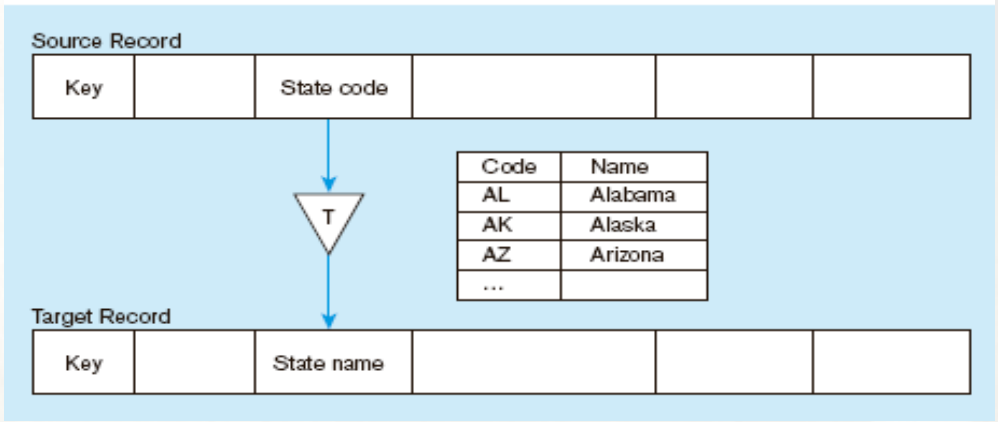
Data Transformation

- Transforms the data in accordance with the business rules and standards that have been established
- Example include: splitting up fields, replacement of codes, derived values, and aggregates

Single-field transformation





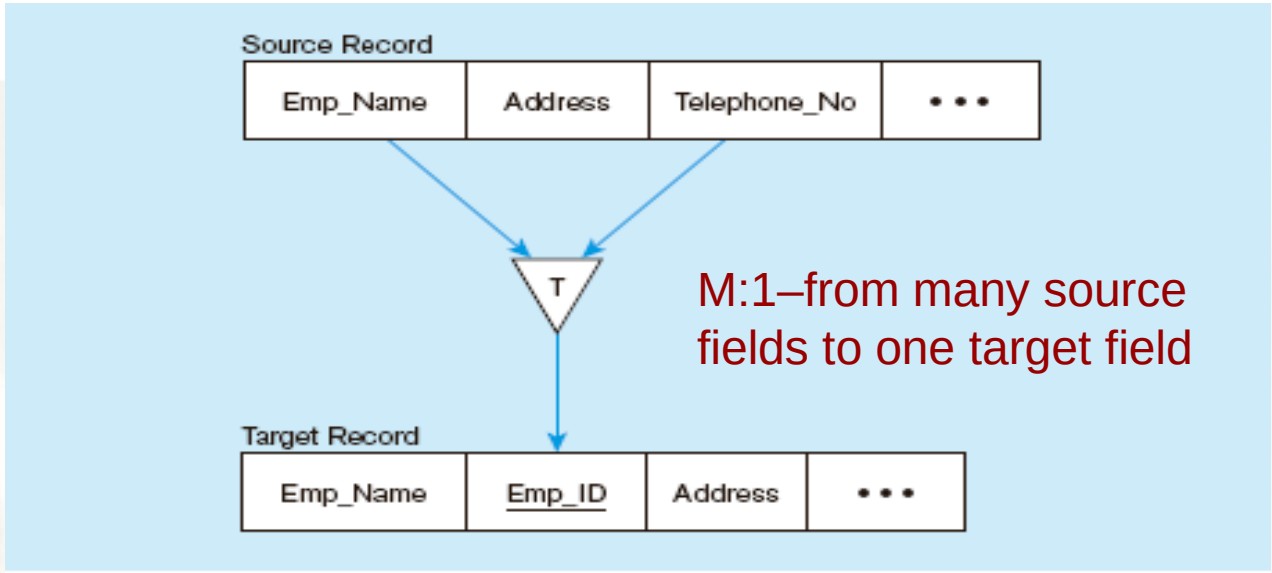


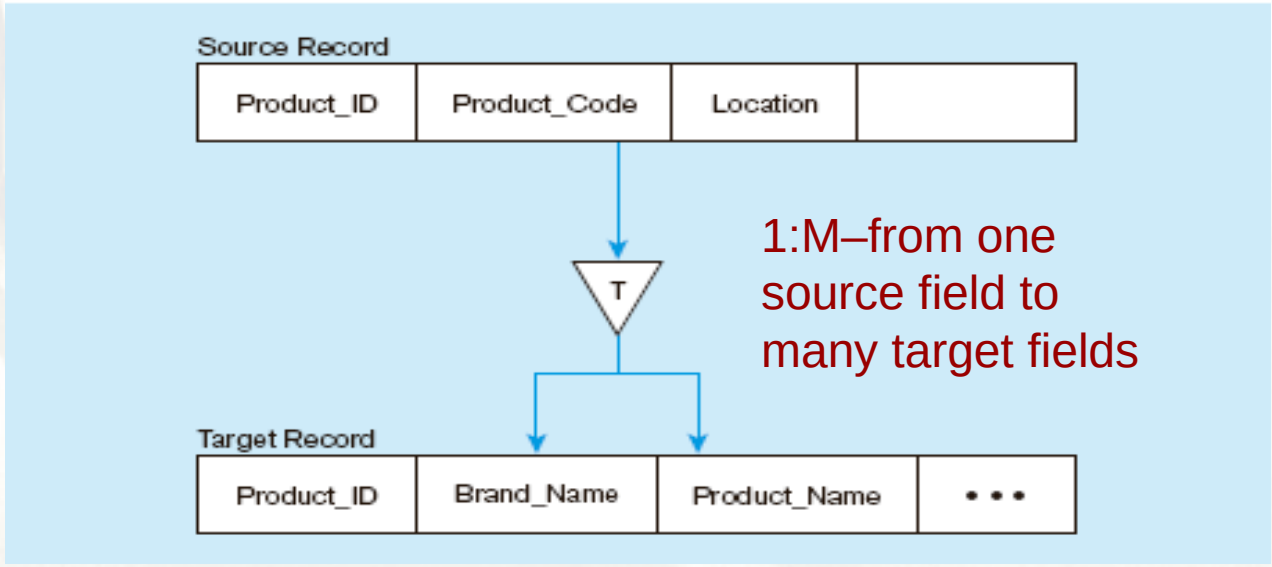
In general–some transformation function translates data from old form to new form

Algorithmic transformation uses a formula or logical expression

Table lookup–another approach, uses a separate table keyed by source record code

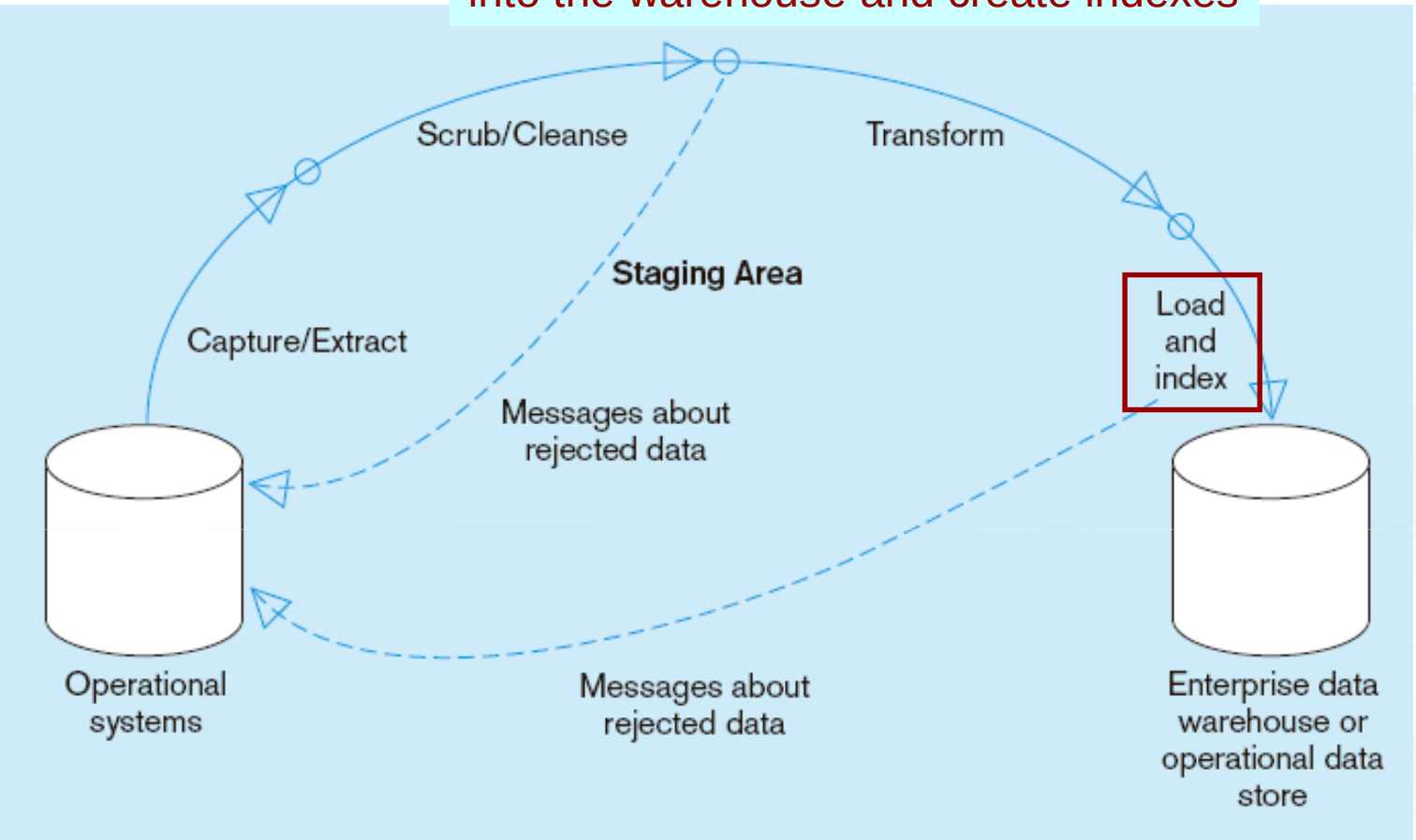
Multifield transformation





Loading

Load/Index= place transformed data into the warehouse and create indexes



Refresh mode: bulk rewriting of target data at periodic intervals

Update mode: only changes in source data are written to data warehouse

Loading

- Data are physically moved to the data warehouse
- Most critical operation in any warehouse
  - affects the DW itself (not the stage area)
- Loading can be broken down into 2 different types:
  - Initial Load
  - Continuous Load (loading over time)

Loading tools

- Backup and Disaster Recovery
  - Restart logic, Error detection
- Continuous loads during a fixed batch window (usually overnight)
  - Maximize system resources to load data efficiently in allotted time
- Parallelization
  - Several tenths GB per hour

ETL tools

- Tools vs. program-from-scratch
  - No tools:
    - Start Immediately, flexibility
    - Complexity, Maintenance
  - Tools:
    - Reduces development costs, easy maintenance
    - Limited flexibility, financial cost

ETL Tools

- ETL tools from the big vendors
  - Oracle Warehouse Builder
  - IBM DB2 Warehouse Manager
  - Microsoft Integration Services
- 100's others:
  - http://en.wikipedia.org/wiki/Extract,_transform,_load#Tools

MS Integration Services