ETL

Dr. Yashvardhan Sharma

Department of Computer Science & Information Systems

BITS, Pilani

L9 (Lec 14)

Topics

- Requirements
- Build or Buy?
- ETL Data Structures
- Data Flow
  - Extract
  - Clean & Conform
  - Deliver
    - Dimension Tables
    - Fact tables
- Implementation & Operations

ETL

"A Properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions… ETL makes or breaks the data warehouse**…"** ***Ralph Kimball***

Requirements

- Business Needs
  - Information requirements of the end user.
  - Captured by interview with users, independent investigations about the possible sources by the ETL team.
- Compliance Requirements
  - Sarbanes-Oxley Act 2002 (Deals with the regulation of corporate governance) (more on ***http://www.soxlaw.com***)
  - Proof of complete transaction flow that changed any data.

Requirements (contd…)

- Data Profiling
  - Systematic examination of quality, scope and the context of a data source
  - Helps ETL team determine how much data cleaning activity require.
  - "[Data Profiling] employs analytic methods for looking at data for the purpose of developing a thorough understanding of the content, structure and the quality of the data. A good data profiling [system] can process very large amounts of data, with the skills of analyst, uncover all sorts of issues that need to be addressed" ***Jack Olson***

Requirements (contd…)

- Security Requirements
  - ETL team have complete read/ write access to the entire corporate data.
  - ETL workstations on the company intranet, A major threat. Keep it in a separate subnet with packet filtering gateway.
  - Secure Backups, as well.
- Data Integration
  - Identified as conform step
  - Conform Dimensions & Conform Facts

Requirements (contd…)

- Data Latency
  - How quickly the data can be delivered to the users?
  - ETL architecture has direct impact on it.
- Archiving & Lineage
  - Stage data after each major transformations, Not just after all the four steps viz extract, clean, conform & deliver.
  - Each archived/ staged data set should have accompanying metadata.
  - Tracking this lineage is explicitly required be certain compliance requirements

Requirements (contd…)

- End user delivery Interfaces
  - ETL team is responsible for the content and structure of data, making the end user applications fast.
- Available Skills
  - Expertise in building ETL system around vendors tool
  - Decision between hand coded or vendors package of ETL tools
- Legacy Licenses
  - Managements insistence to use legacy licenses

Choice of Architecture →
Tool Based ETL

Simpler, Cheaper & Faster development

- **People with business skills & not much technical skills can use it.**
- **Automatically generate Metadata**
- **Automatically generates data Lineage & data dependency analysis**
- **Offers in-line encryption & compression capabilities**
- **Manage complex load balancing across servers**

Choice of Architecture →

Hand-Coded ETL

- **Quality of tool by exhaustive unit testing**
- **Better metadata**
- **Requirement may be just file based processes not database-stored procedures**
- **Use of existing legacy routines**
- **Use of in-house programmers**
- **Unlimited flexibility**

To stage or not to stage

- **Decision to store data in physical staging area versus processing it in memory is ultimately the choice of the ETL architect**

To stage or not to stage

- A conflict between
  - getting the data from the operational systems as fast as possible
  - having the ability to restart without repeating the process from the beginning
- Reasons for staging
  - **Recoverability**: stage the data as soon as it has been extracted from the source systems and immediately after major processing (cleaning, transformation, etc).
  - **Backup**: can reload the data warehouse from the staging tables without going to the sources
  - **Auditing**: lineage between the source data and the underlying transformations before the load to the data warehouse

Designing the staging area

- The staging area is owned by the ETL team
  - no indexes, no aggregations, no presentation access, no querying, no service level agreements
- Users are not allowed in the staging area for any reason
  - staging is a “construction” site
- Reports cannot access data in the staging area
  - tables can be added, or dropped without notifying the user community
  - Controlled environment

Designing the staging area (contd…)

- **Only ETL processes can read/write the staging area (ETL developers must capture table names, update strategies, load frequency, ETL jobs, expected growth and other details about the staging area)**
- **The staging area consists of both RDBMS tables and data files**

Staging Tables Volumetric Worksheet

- **Lists each table in the staging area with the following information:**
  - **Table Name: name or table or file in the DSA. One row in the WS for each staging table**
  - **Update Strategy: Indicates how a table is maintained. For persistent tables it will have data appended, updated, or perhaps deleted. Transient tables are truncated and reloaded with each process**
  - **Load Frequency: How often the table is loaded or changed by the ETL process. In real-time environment – continuously**
  - **ETL Jobs**
  - **Initial Row count:**

Staging Area data Structures in the ETL System

- **Flat files**
  - **fast to write, append to, sort and filter (grep) but slow to update, access or join**
  - **enables restart without going to the sources**
- **XML Data Sets**
  - **Used as a medium of data transfer between incompatible data sources**
  - **Gives enough information to create tables using CREATE TABLE**
- **Relational Tables**
  - **Metadata, SQL interface, DBA support**

Staging Area data Structures in the ETL System (contd…)

- Dimensional Model Constructs: Facts, Dimensions, Atomic Facts tables, Aggregate Fact Tables (OLAP Cubes)

- Surrogate Key Mapping Tables
  - map natural keys from the OLTP systems to the surrogate key from the DW
  - can be stored in files or the RDBMS

Logical data map

- Represented as a table with the following attributes
  - Target table and column, table type (Dimension, Fact)
  - Slow-changing dimension type per target column of each dimensions
    - Type 1, overwrite (Customer first name)
    - Type 2, retain history (Customer last name)
    - Type 3, retain multiple valid alternative values
  - Source database, table, column
  - Transformations

Logical Map development

- Have a plan !
  - Logical map is provided by Datawarehouse architect to ETL Team & serves as the specification of ETL processes.
  - Identify data lineage between the data source & target
- Identity source candidates
- Analyze source systems with a data profiling tool
  - Detect data anomaly, identify appropriate actions & document it.
  - Identify the quality

Logical Map development (contd…)

- Receive walk-through of data lineage and business rules (from the DW architect and business analyst to the ETL developer)
  - data alterations during data cleansing, calculations and formulas
  - standard conformance to dimensions and numerical facts
- Receive walk-through of the dimensional model
  - The objective of ETL team is to deliver data to the dimensional model in a more effective way & hence the understanding of dimensional model is helpful
- Validate calculations & Formulas used in ETL.

Logical Map development (contd…)

- Complete logical map cannot exist until all the source systems are identified & analyzed.
- Analysis of source
  - Data discovery phase
    1. Collecting & documenting Source systems
    2. Keeping track of source systems

       Identify ownership, responsible for the content and its storage & usage statistics
    3. Determine the system-of-record (source of data)

       Identify the source, when redundant sources coexist
    4. Analyze the source systems for any relationship between tables
  - Anomaly detection phase (Data content Analysis)
    1. NULL values
    2. Dates in Nondate fields

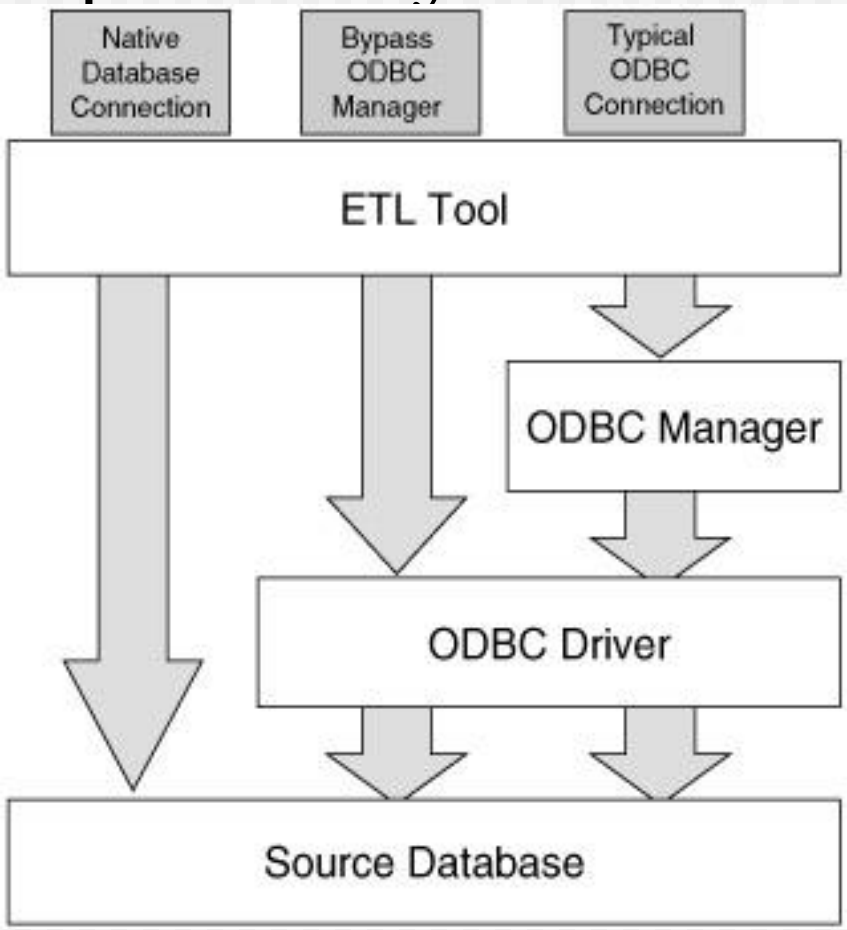
Some good rules

- Analyze your source system
  - get a ER-model for the system or reverse engineering one (develop one by looking at the metadata of the system)
  - reverse engineering is not the same as “forward engineering”, i.e., given the ER-models of the source systems derive the dimensional schema of the data warehouse
- Reverse engineering helps understanding
  - unique identifiers and natural keys
  - data types
  - relationships between tables

Of the source systems

Extract data from disparate systems

- ODBC Manager:
  - Accepts SQL from ETL applications & routes it through appropriate ODBC driver
- ODBC can provide a common gateway to diverse sources



Different sources

1. Mainframe Sources
2. Flat Files
3. XML sources
4. Web Log sources
5. ERP system sources

Tips for Extracting

- Constrain on indexed columns
- Retrieve only the data you need
  - Do not retrieve entire table & select from that.
- Use DISTINCT & Set operations sparingly
  - Try if the slower DISTINCT, Set operations UNION, MINUS and INTERSECT operations can be avoided
  - Do the best effort to avoid NOT operation, which normally scans the entire database
- Avoid functions in the WHERE clause
  - Difficult to avoid
  - Try different techniques before using the functions, at least.
- Avoid subqueries

Do it also !

- When a dimension is populated by several distinct systems, it is important to include the unique identifier from each of those systems in the target dimension in the data warehouse. Those identifiers should be viewable by end users to ensure peace of mind that the dimension reflects *their* data that they can tie back to in their transaction system.

Transformation Tools

- **Transform extracted data into the appropriate format, data structure, and values that are required by the DW**
- **Features provided:**
  - **Field splitting & consolidation**
  - **Standardization**
    - **Abbreviations, date formats, data types, character formats, etc.**
  - **Deduplication**

Source System	Type of transformation	DW
Address Field: #123 ABC Street XYZ City 1000 Republic of MN	Field Splitting	No: 123 Street: ABC City: XYZ Country: Republic of MN Postal Code: 1000
System A Customer title: President System B Customer title: CEO	Field Consolidation	Customer title: President & CEO
Order Date:05 August 1998 Order Date: 08/08/98	Standardization	Order Date: 05 August 1998 Order Date: 08 August 1998
System A Customer Name: John W. Smith System B Customer Name: John William Smith	Deduplication	Customer Name: John William Smith

4. Cleaning & Conforming

Cleaning and Conforming

- While the Extracting and Loading part of an ETL process simply moves data, the cleaning and conforming part , the transformation part that truly adds value
- How do we deal with dirty data?
  - Data Profiling report
  - The Error Event fact table
  - Audit Dimension
- Challenges
  - Completeness Vs Speed
  - Corrective Vs Transparent
    - Too corrective system hides/obscures the operational deficiencies & slows organizational progress

Defining Data Quality

- Basic definition of data quality is data accuracy and that means
  - Correct: the values of the data are valid, e.g., my resident state is PA
  - Unambiguous: The values of the data can mean only one thing, e.g., there is only one PA
  - Consistent: the values of the data use the same format, e.g., PA and not Penn, or Pennsylvania
  - Complete: data are not null, and aggregates do not lose data record somewhere in the information flow

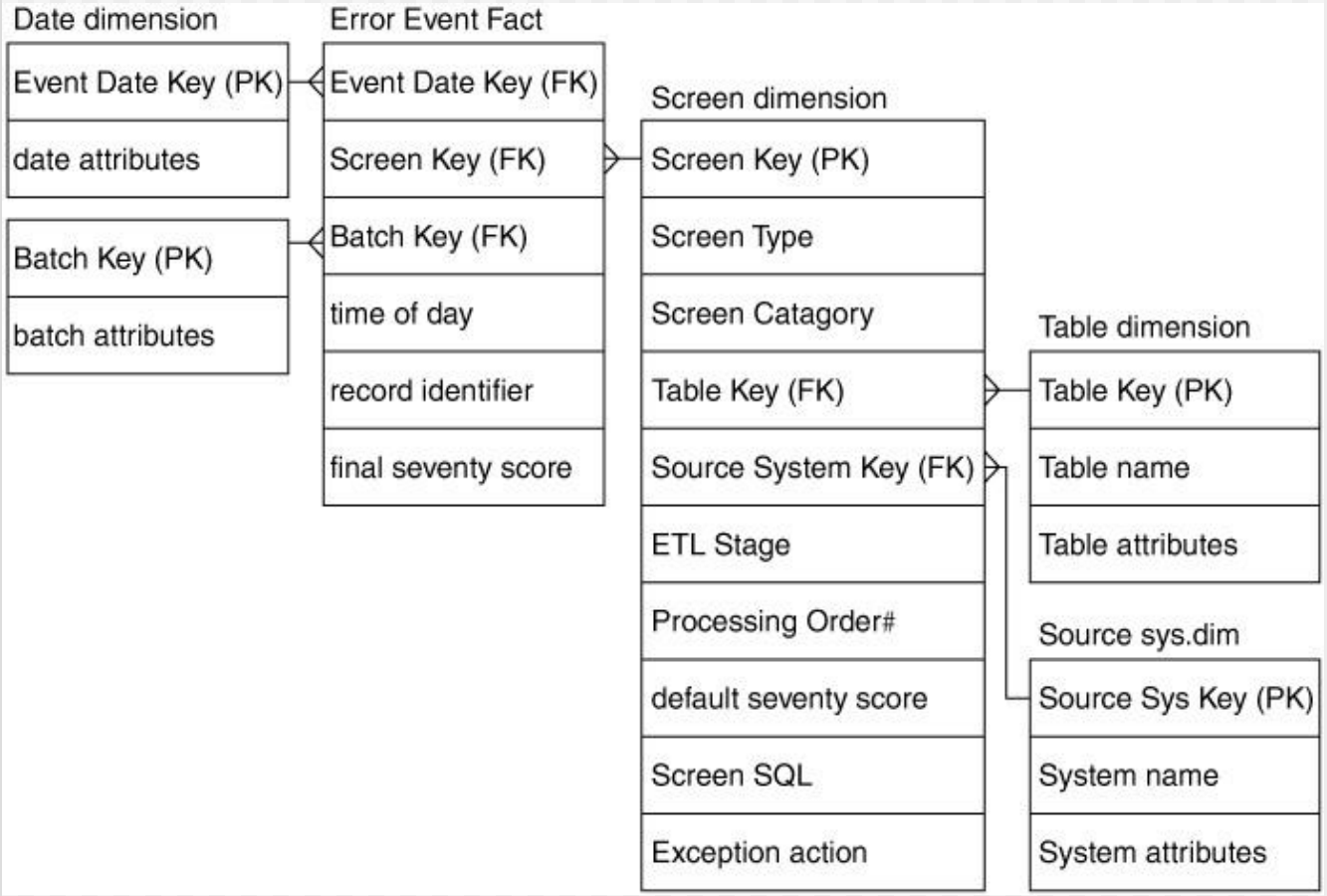
Cleaning Deliverables

- Keep accurate records of the types of data quality problems you look for, when you look, what you look at, etc
  - Is data quality getting better or worse?
  - Which source systems generate the most data quality errors?
  - Is there any correlation between data quality levels and the performance of the organization as a whole?

Data Profiling Deliverable

- Start before building the ETL system
- Data profiling analysis including
  - Schema definitions
  - Business objects
  - Domains
  - Data Sources
  - Table definitions
  - Synonyms
  - Data rules
  - Value rules
  - Issues that need to be addressed

Error Event Table Deliverable



- Built as a star schema
- Each data quality error or issue is added to the table

Conforming

- Integration of data
- A conformed product dimension is the enterprise's agreed upon master list of products, including all attributes. It can be considered as a master table of products with clean surrogate product key with all relevant attributes.
- Processes of conforming
  - Standardizing
  - Matching & deduplication
  - Surviving

DATA LOADING

DATA LOADING

- Data loading takes the prepared data, applies it to the data warehouse, and stores it in the database
- Terminology:
  - **Initial Load** — populating all the data warehouse tables for the very first time
  - **Incremental Load** — applying ongoing changes as necessary in a periodic manner
  - **Full Refresh** — completely erasing the contents of one or more tables and reloading with fresh data (initial load is a refresh of all the tables)

Applying Data: Techniques and Processes

- load,
- append,
- destructive merge,
- constructive merge.

Load

- If the target table to be loaded already exists and data exists in the table, the load process wipes out the existing data and applies the data from the incoming file.
- If the table is already empty before loading, the load process simply applies the data from the incoming file.

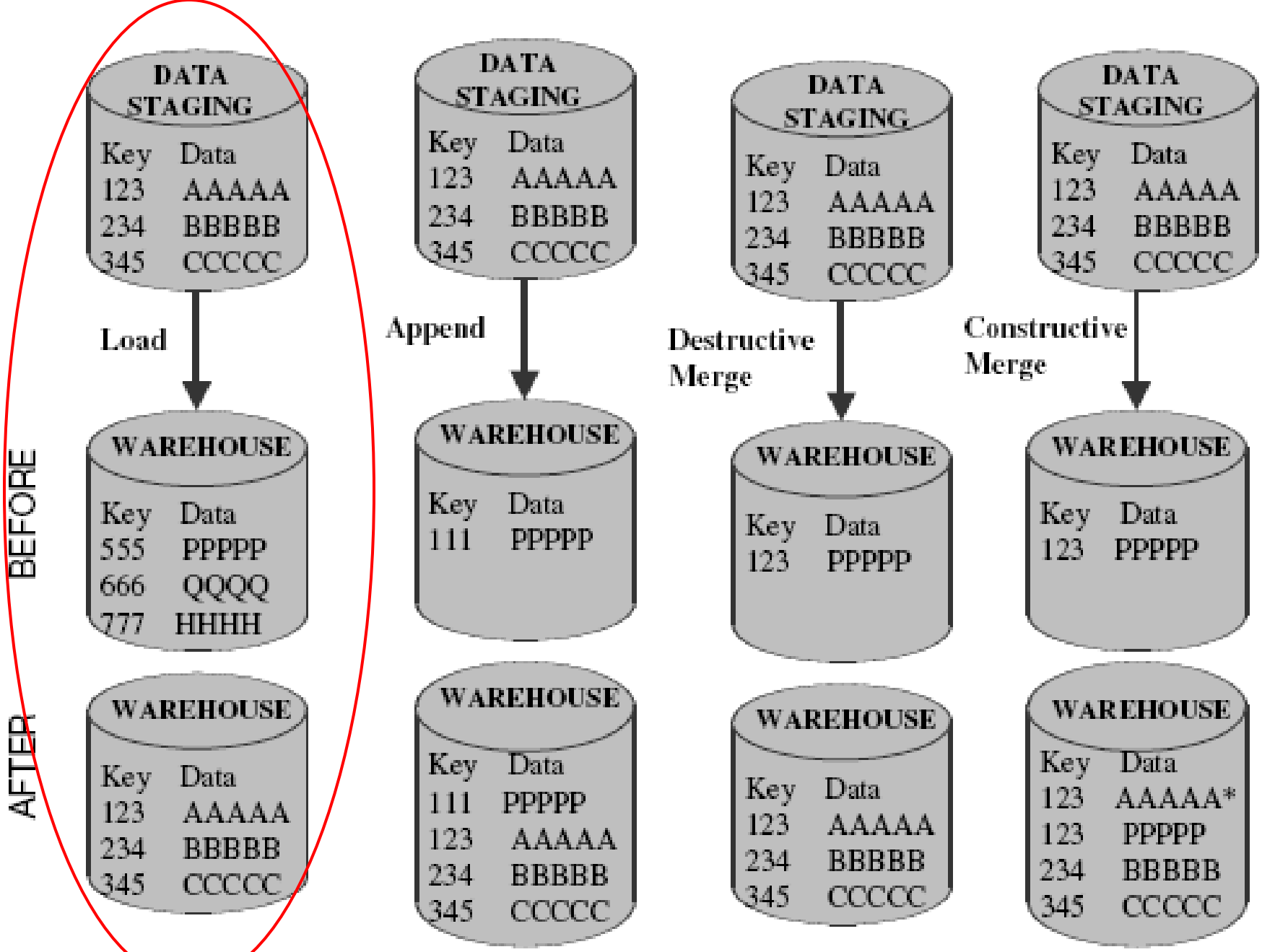


Figure 12-11 Modes of applying data.

Append

- extension of the load.
- If data already exists in the table, the append process unconditionally adds the incoming data, preserving the existing data in the target table.
- When an incoming record is a duplicate of an already existing record, you may define how to handle an incoming duplicate:
  - The incoming record may be allowed to be added as a duplicate.
  - In the other option, the incoming duplicate record may be rejected during the append process.

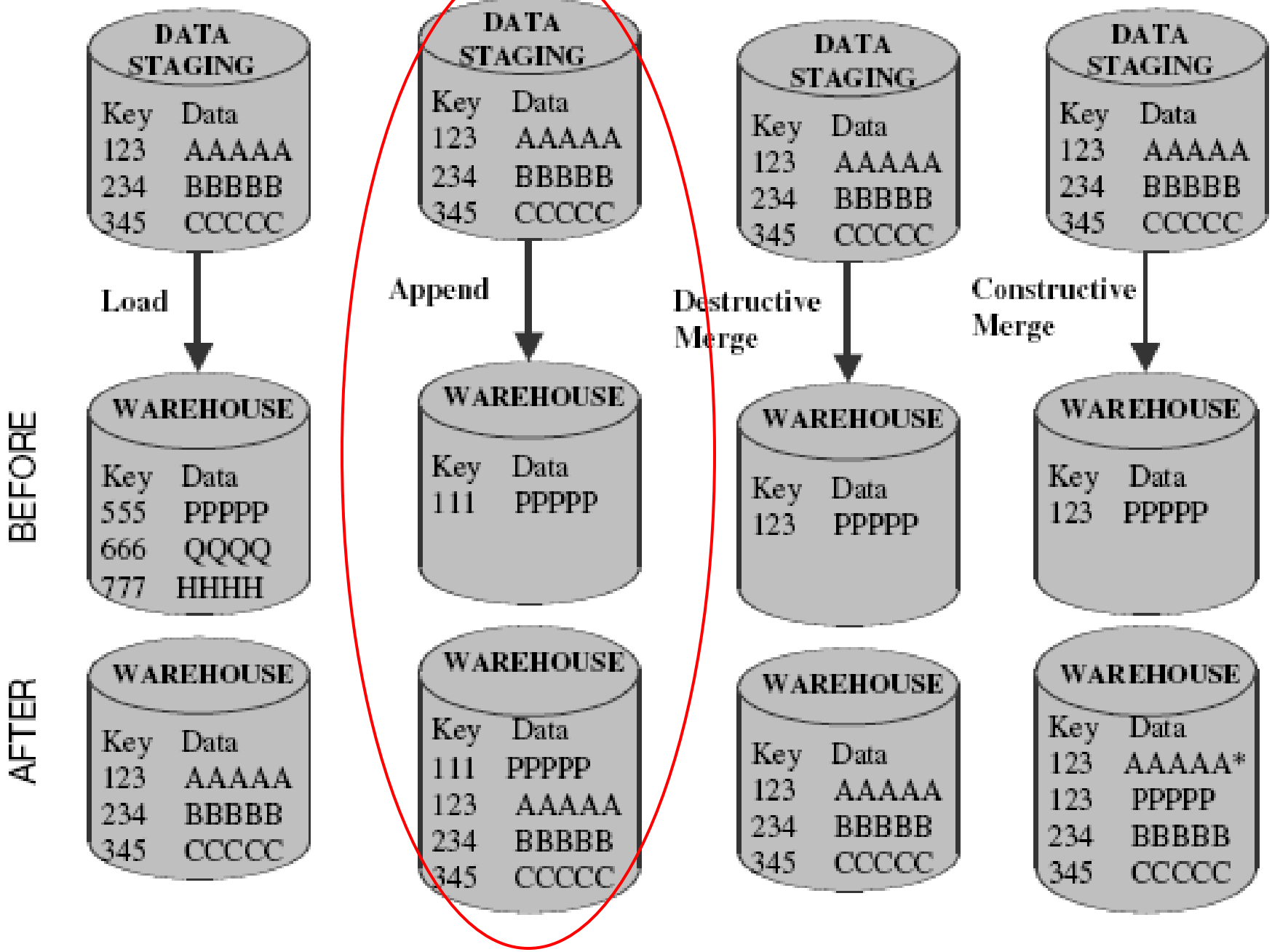


Figure 12-11 Modes of applying data.

Destructive Merge

- Applies incoming data to the target data.
- If the primary key of an incoming record matches with the key of an existing record, update the matching target record.
- If the incoming record is a new record without a match with any existing record, add the incoming record to the target table.

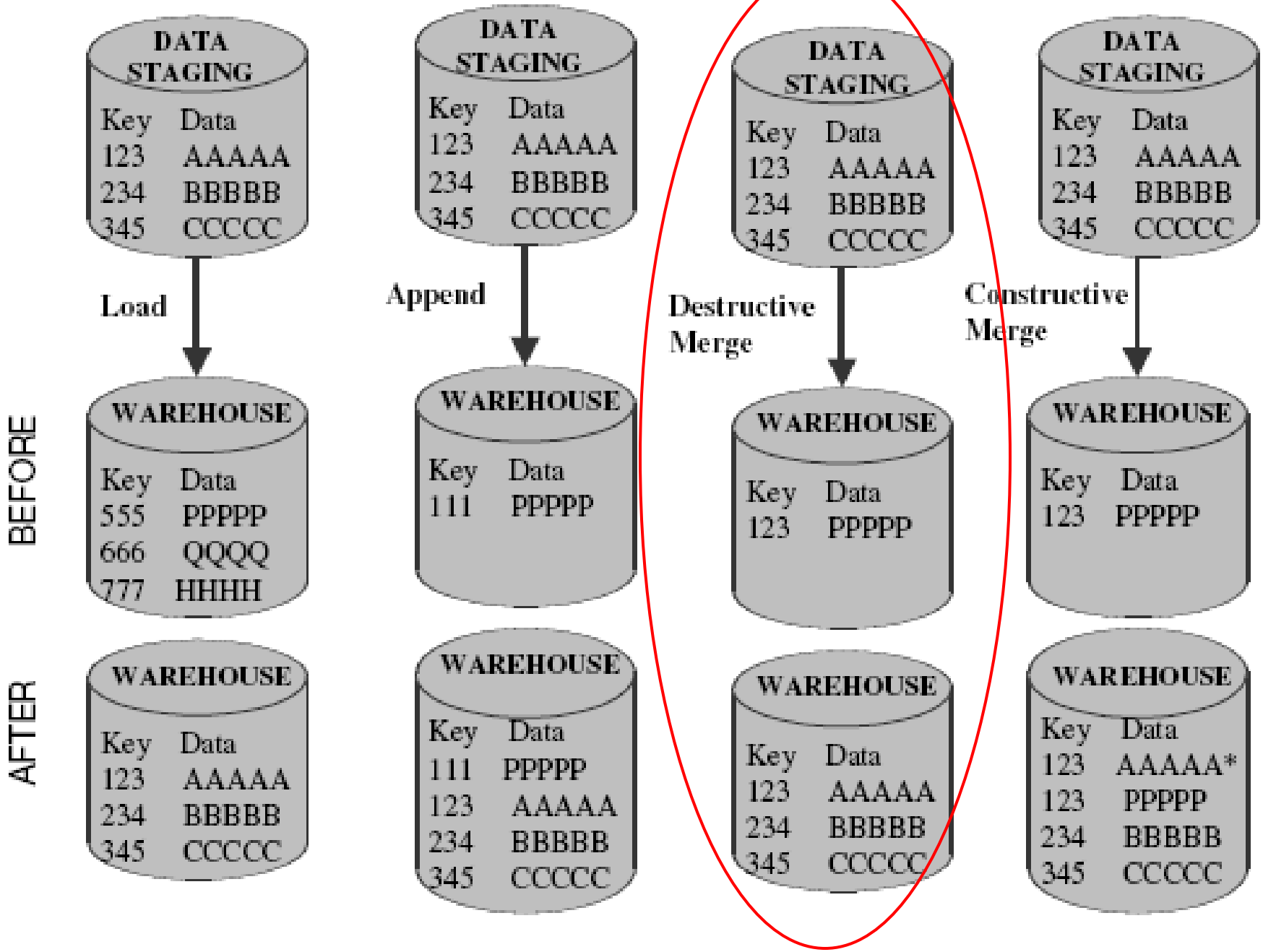


Figure 12-11 Modes of applying data.

Constructive Merge

- Slightly different from the destructive merge.
- If the primary key of an incoming record matches with the key of an existing record, leave the existing record, add the incoming record, and mark the added record as superceding the old record.

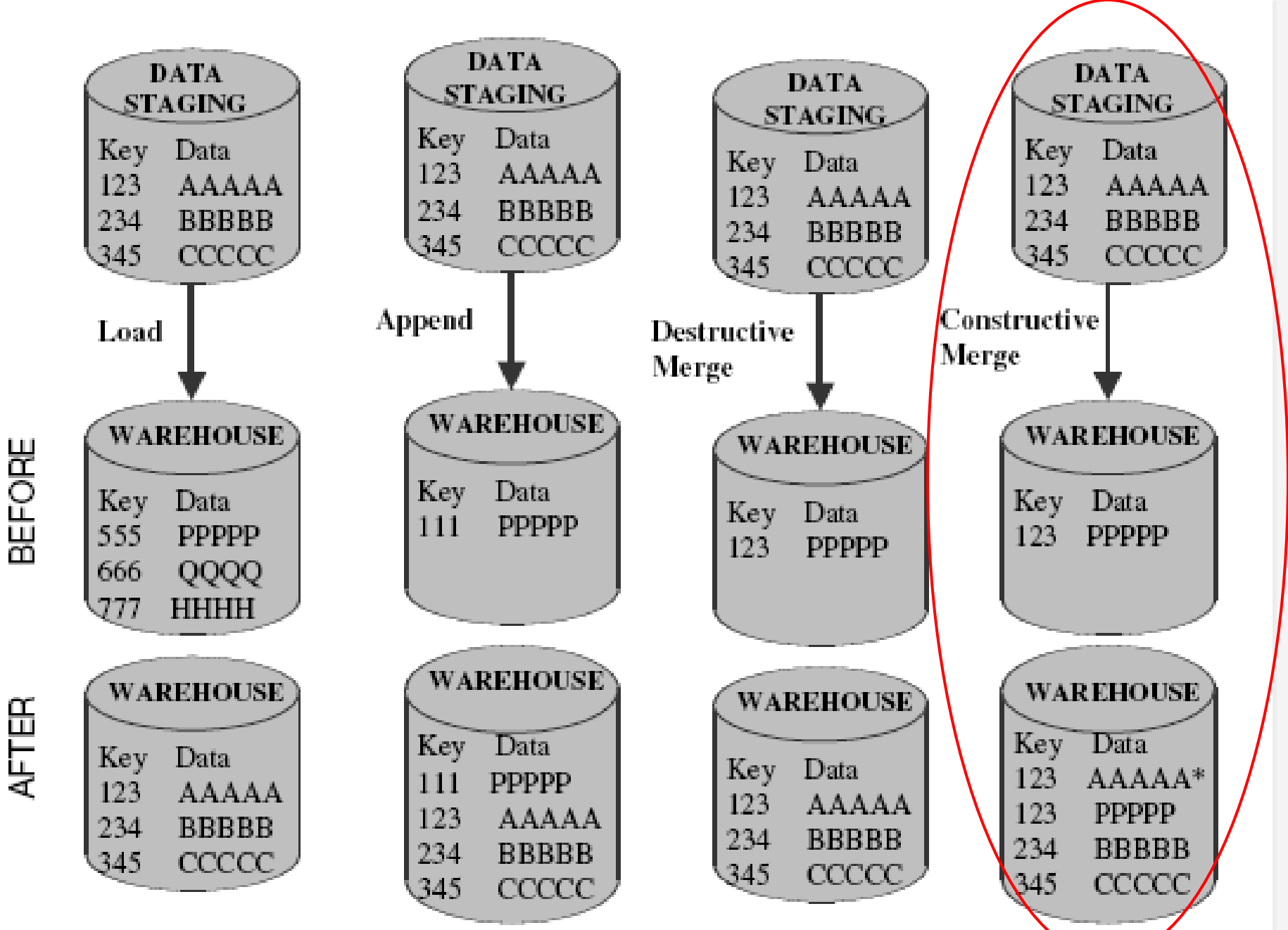


Figure 12-11 Modes of applying data.

ETL Tools Options

- Data transformation engines
- Data capture through replication
- Code generators

Data transformation engines

- Consist of dynamic and sophisticated data manipulation algorithms.
- The tool suite captures data from a designated set of source systems at user-defined intervals, performs elaborate data transformations, sends the results to a target environment, and applies the data to target files.
- These tools provide maximum flexibility for pointing to various source systems, to select the appropriate data transformation methods, and to apply full loads and incremental loads.
- The functionality of these tools sweeps the full range of the ETL process.

Data capture through replication

- Most of these tools use the transaction recovery logs maintained by the DBMS.
- The changes to the source systems captured in the transaction logs are replicated in near real time to the data staging area for further processing.
- Some of the tools provide the ability to replicate data through the use of database triggers. These specialized stored procedures in the database signal the replication agent to capture and transport the changes.

Code generators

- Tools that directly deal with the extraction, transformation, and loading of data.
- The tools enable the process by generating program code to perform these functions.
- Code generators create 3GL/4GL data extraction and transformation programs.
- The tools generate most of the program code in some of the common programming languages.
- Own program code can be addedd, also.
- The code automatically generated by the tool has exits at which points you may add your code to handle special conditions.

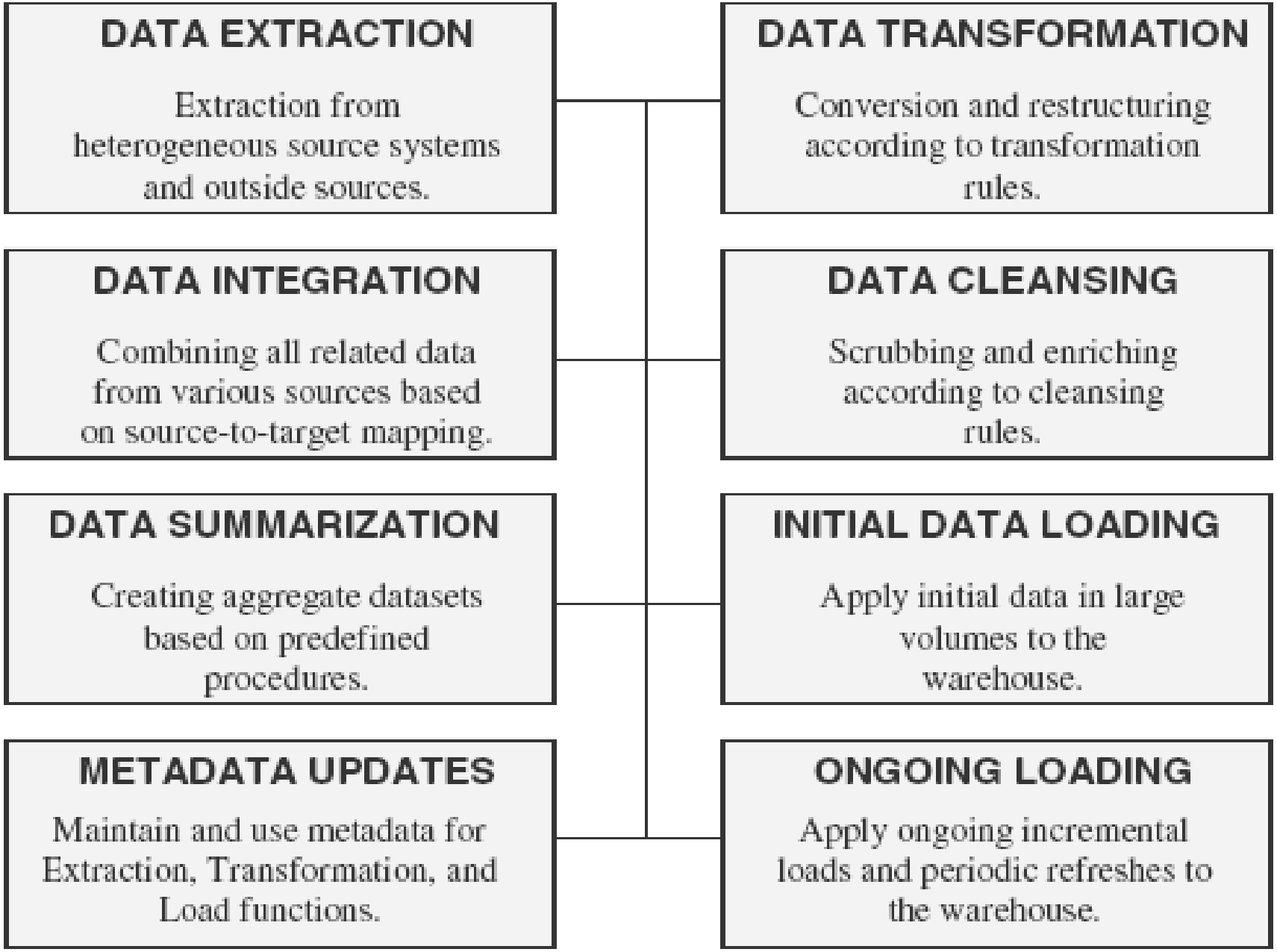


Figure 12-14 ETL summary.