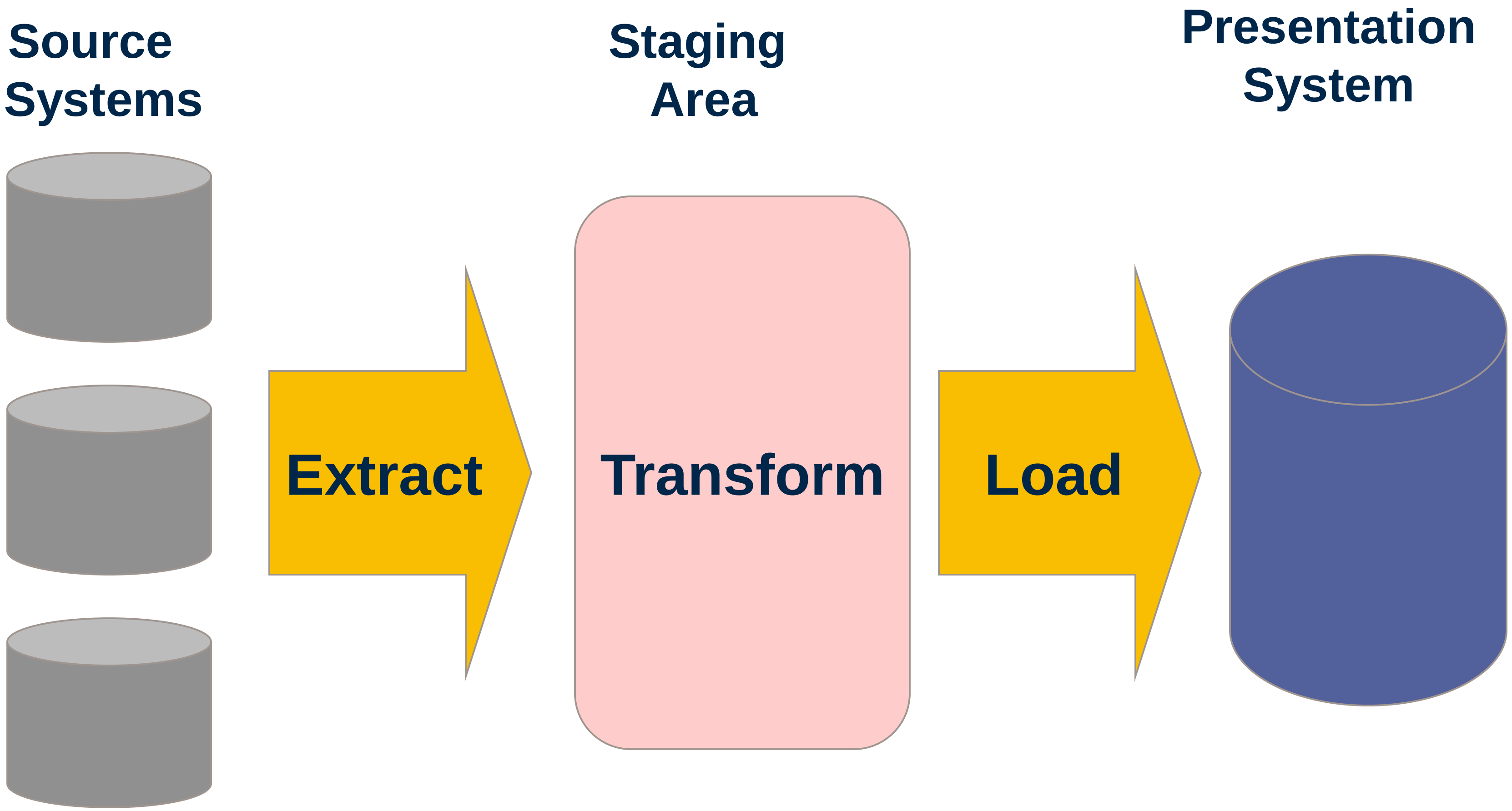
ETL Basics

- Lesson 2: ETL Process

Lesson Objectives

- On completion of this lesson on Data Modeling, you will be able to understand:
  - The ETL process
  - The steps in Data Cleansing

The ETL Process



Source Systems

Staging Area

Presentation System

Extract

Transform

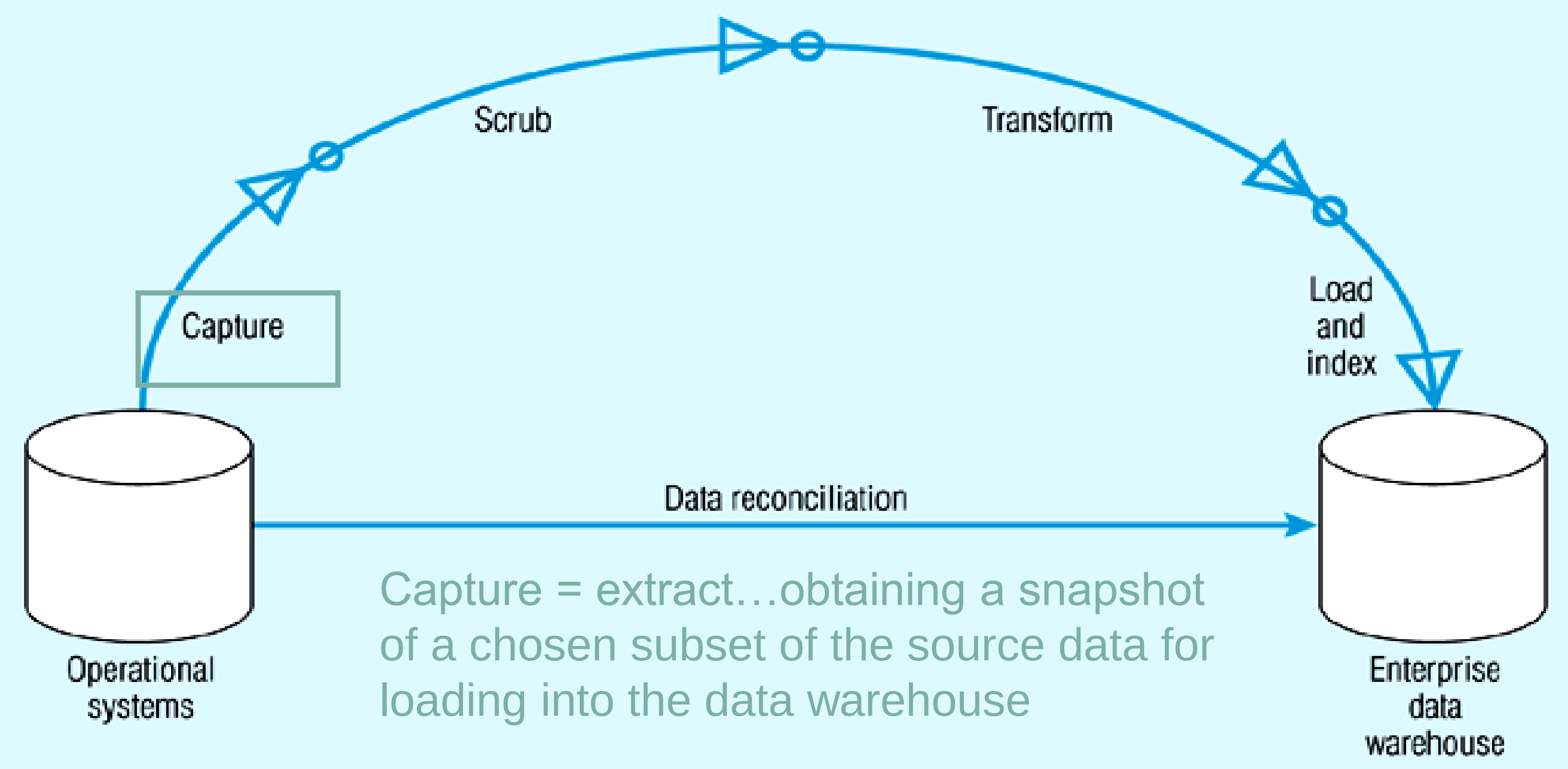
Load

The ETL Process

- Extract
  - Extract relevant data
- Transform
  - Transform data to DW format
  - Build keys, etc.
  - Cleansing of data
- Load
  - Load data into DW
  - Build aggregates, etc

EXTRACTION PHASE

ETL – DATA CAPTURE



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

Change Data Capture

- Data warehousing involves the extraction and transportation of data from one or more databases into a target system or systems for analysis.
- But this involves the extraction and transportation of huge volumes of data and is very expensive in both resources and time.
- The ability to capture only the changed source data and to move it from a source to a target system(s) in real time is known as Change Data Capture (CDC).

Change Data Capture

- CDC helps identify the data in the source system that has changed since the last extraction.

- Set of software design patterns used to determine the data that has changed in a database.

Change Data Capture

- Based on the Publisher/Subscriber model.
- Publisher
  - Identifies the source tables from which the change data needs to be captured
  - Captures the change data and stores it in specially created change tables
  - Allows the subscribers controlled access to the change data
- Subscriber
  - Subscriber needs to know what change data it is interested in
  - It creates a subscriber view to access the change data to which it has been granted access by the publisher

Data Staging

- Often used as an interim step between data extraction and later steps
- Accumulates data from asynchronous sources using native interfaces, flat files, FTP sessions, or other processes
- At a predefined cutoff time, data in the staging file is transformed and loaded to the warehouse
- There is usually no end user access to the staging file
- An operational data store may be used for data staging

Reasons for “Dirty” Data

- Dummy Values
- Absence of Data
- Multipurpose Fields
- Inappropriate Use of Address Lines
- Violation of Business Rules
- Reused Primary Keys,
- Non-Unique Identifiers
- Data Integration Problems

ETL – DATA Extraction

- The extraction process can be done either by hand coded method or by using tools.
- Advantages and disadvantages Of Custom-programmed)/Hand Coded Extraction (PL SQL Scripts) and Tool based extraction.
- Tools have Well Defined disciplined approach and Documentation.
- Tools provide an easier way to perform the extraction method by providing click, drag and drop features.
- Hand coded extraction techniques allow extraction in cost effective manner since the PL/SQL construct are available with the RDBMS.
- Hand coded extraction are used when the extraction is to be taken place where the programmer has clear data structure known.

ETL - Extraction Techniques

- Extraction Technique
- Bulk Extraction-
  - The entire data warehouse is refreshed periodically by extraction's from the source systems.
  - All applicable data are extracted from the source systems for loading into the warehouse.
  - This approach heavily uses the network connection for loading data from source to target databases, but such mechanism is easy to set up and maintain.

Data Extraction

- Capture of data from Source Systems
- Important to decide the frequency of Extraction
- Sometimes source data is copied to the target database using the replication capabilities of standard RDBMS (not recommended because of “dirty data” in the source systems)

Data Transformation

- Transforms the data in accordance with the business rules and standards that have been established

- Example include: format changes, de-duplication, splitting up fields, replacement of codes, derived values, and aggregates

Data Transformation

- Validating
  - Process of ensuring that the data captured is accurate and transformation process is correct
  - E.g. Date of Birth of a Customer should not be more than today’s date

Data Transformation

- Data Cleansing
  - Source systems contain “dirty data” that must be cleansed
  - ETL software contains rudimentary data cleansing capabilities
  - Specialized data cleansing software is often used.
  - Important for performing name and address correction and house holding functions
  - Leading data cleansing vendors include Vality (Integrity), Harte-Hanks (Trillium), and Firstlogic (i.d.Centric)

Data Transformation

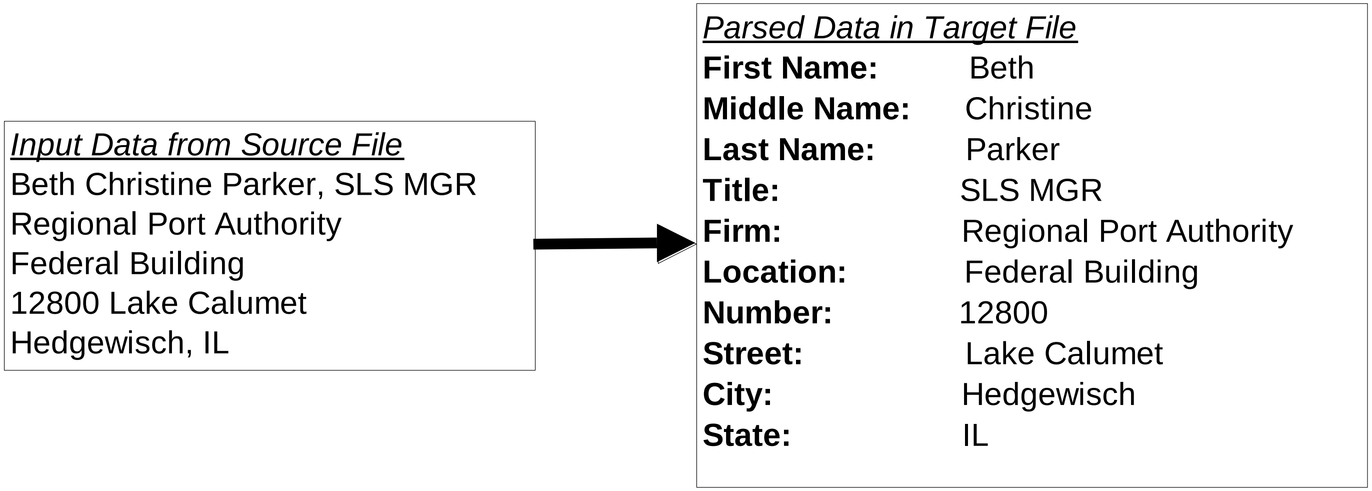
- Steps in Data Cleansing
  - Parsing
  - Correcting
  - Standardizing
  - Matching
  - Consolidating
  - Conditioning
  - Enrichment

Data Transformation

- Parsing
  - Parsing locates and identifies individual data elements in the source files and then isolates these data elements in the target files
  - Examples include :
    - parsing the first, middle, and last name;
    - street number and street name; and city and state

Data Transformation

- Parsing



Input Data from Source File

Beth Christine Parker, SLS MGR
Regional Port Authority
Federal Building
12800 Lake Calumet
Hedgewisch, IL

Parsed Data in Target File

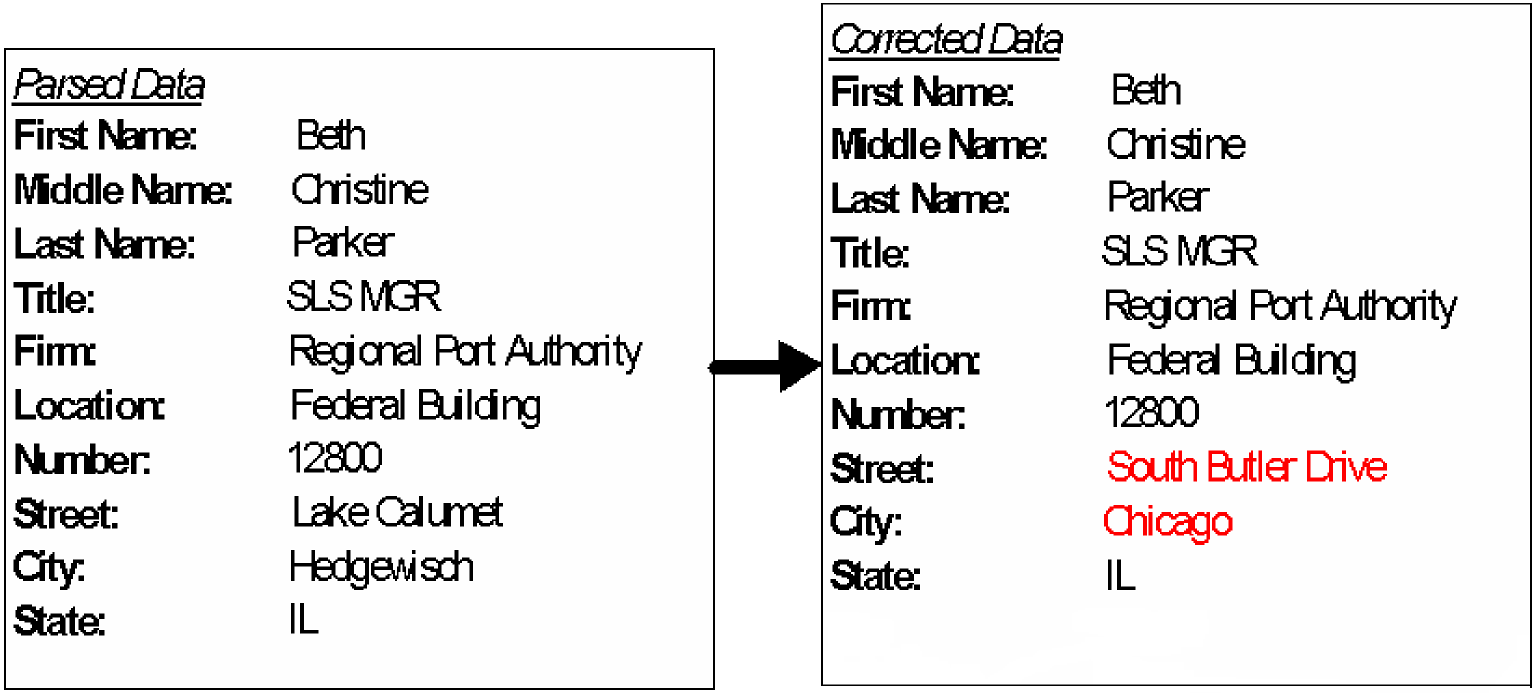
First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

Data Transformation

- Correcting
  - Corrects parsed individual data components using sophisticated data algorithms and secondary data sources.
  - Example include replacing a vanity address and adding a zip code.

Data Transformation

- Correcting



Parsed Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

Corrected Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL

Data Transformation

- Standardizing
  - Standardizing applies conversion routines to transform data into its preferred (and consistent) format using both standard and custom business rules.
  - Examples include adding a pre name, replacing a nickname, and using a preferred street name.

Data Transformation

- Standardizing

Corrected Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Corrected Data

Pre-name:	Ms.
First Name:	Beth
1st Name Match Standards:	Elizabeth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr.
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Data Transformation

- Matching
  - Searching and matching records within and across the parsed, corrected and standardized data based on predefined business rules to eliminate duplications.
  - Examples include identifying similar names and addresses.

Data Transformation

- Matching

Corrected Data (Data Source #1)

Pre-name:	Ms.
First Name:	Beth
1st Name Match Standards:	Elizabeth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr.
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

↔

Corrected Data (Data Source #2)

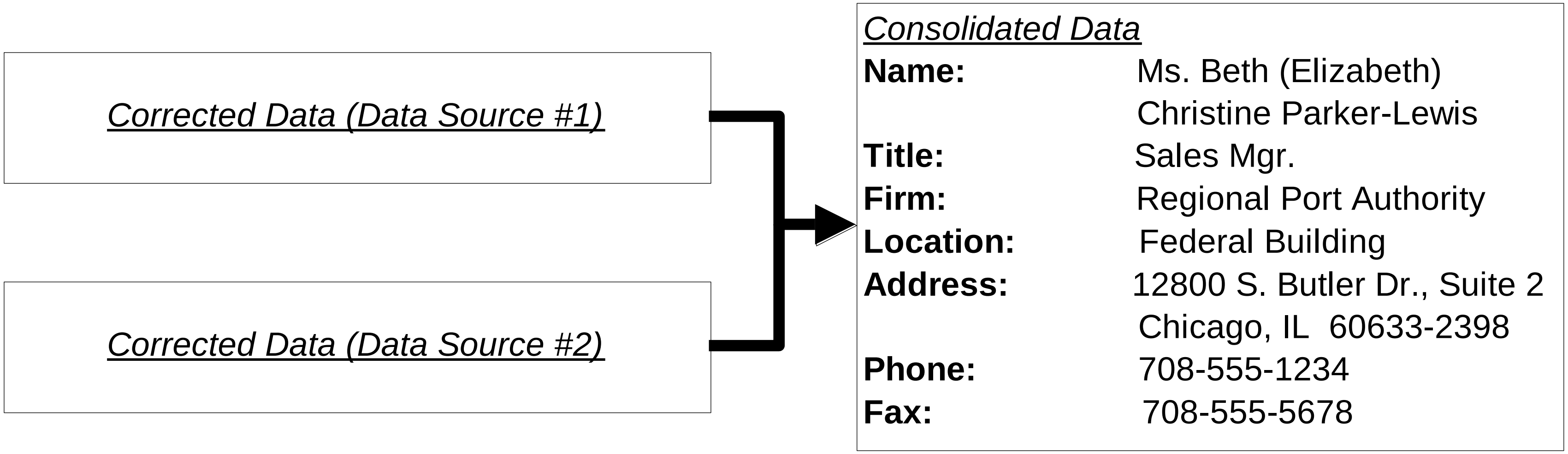
Pre-name:	Ms.
First Name:	Elizabeth
1st Name Match Standards:	Beth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker-Lewis
Title:	
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr., Suite 2
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398
Phone:	708-555-1234
Fax:	708-555-5678

Data Transformation

- Consolidating
- Analyzing and identifying relationships between matched records and consolidating/merging them into ONE representation.

Data Transformation

- Consolidating



Corrected Data (Data Source #1)

Corrected Data (Data Source #2)

Consolidated Data

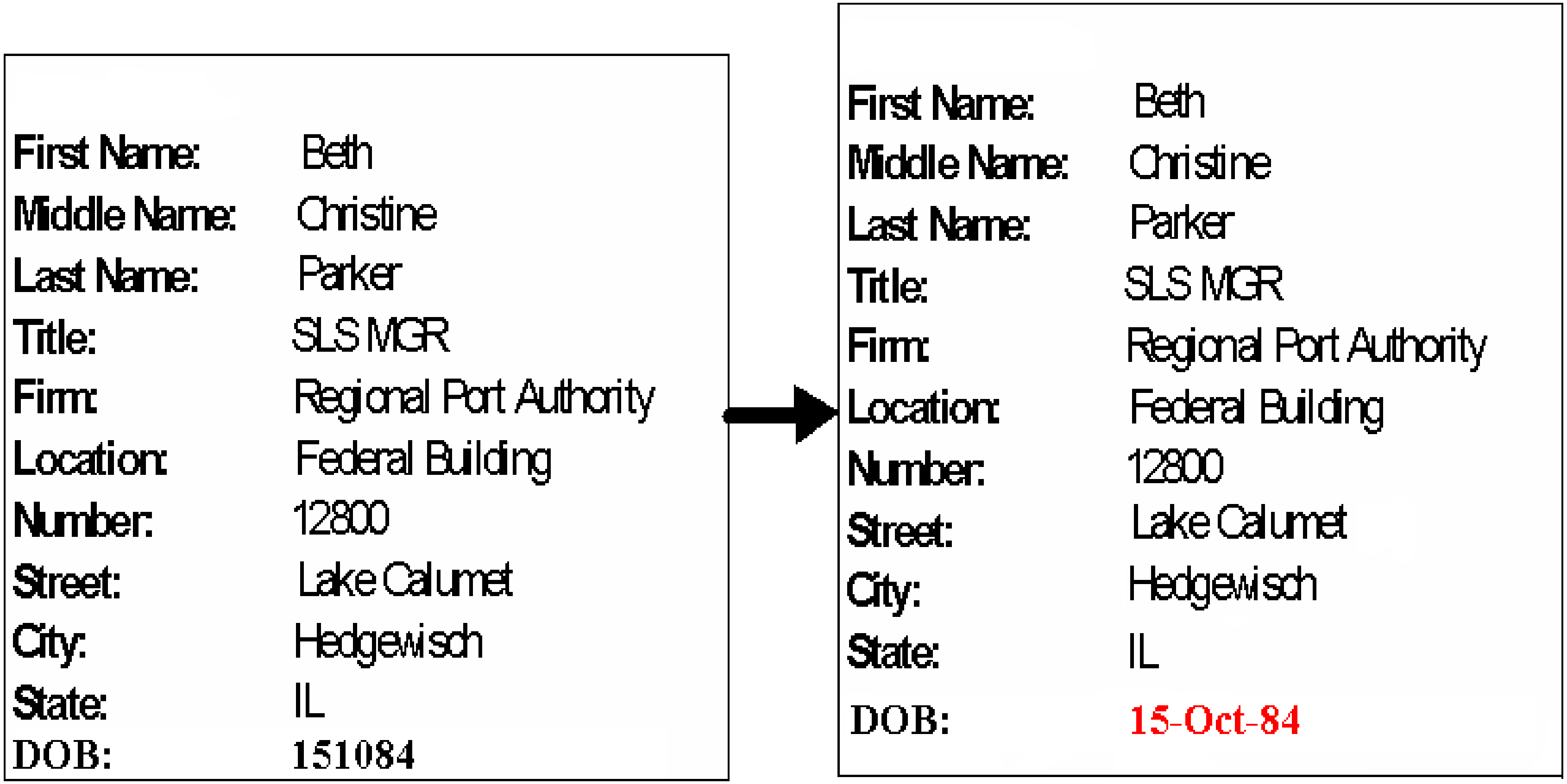
Name:	Ms. Beth (Elizabeth) Christine Parker-Lewis
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Address:	12800 S. Butler Dr., Suite 2 Chicago, IL 60633-2398
Phone:	708-555-1234
Fax:	708-555-5678

Data Transformation

- Conditioning
  - The conversion of data types from the source to the target data store (warehouse) -- always a relational database
  - Eg. OLTP Date stored as text (DDMMYY); DW format is Oracle Date type

Data Transformation

- Conditioning



First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL
DOB:	151084

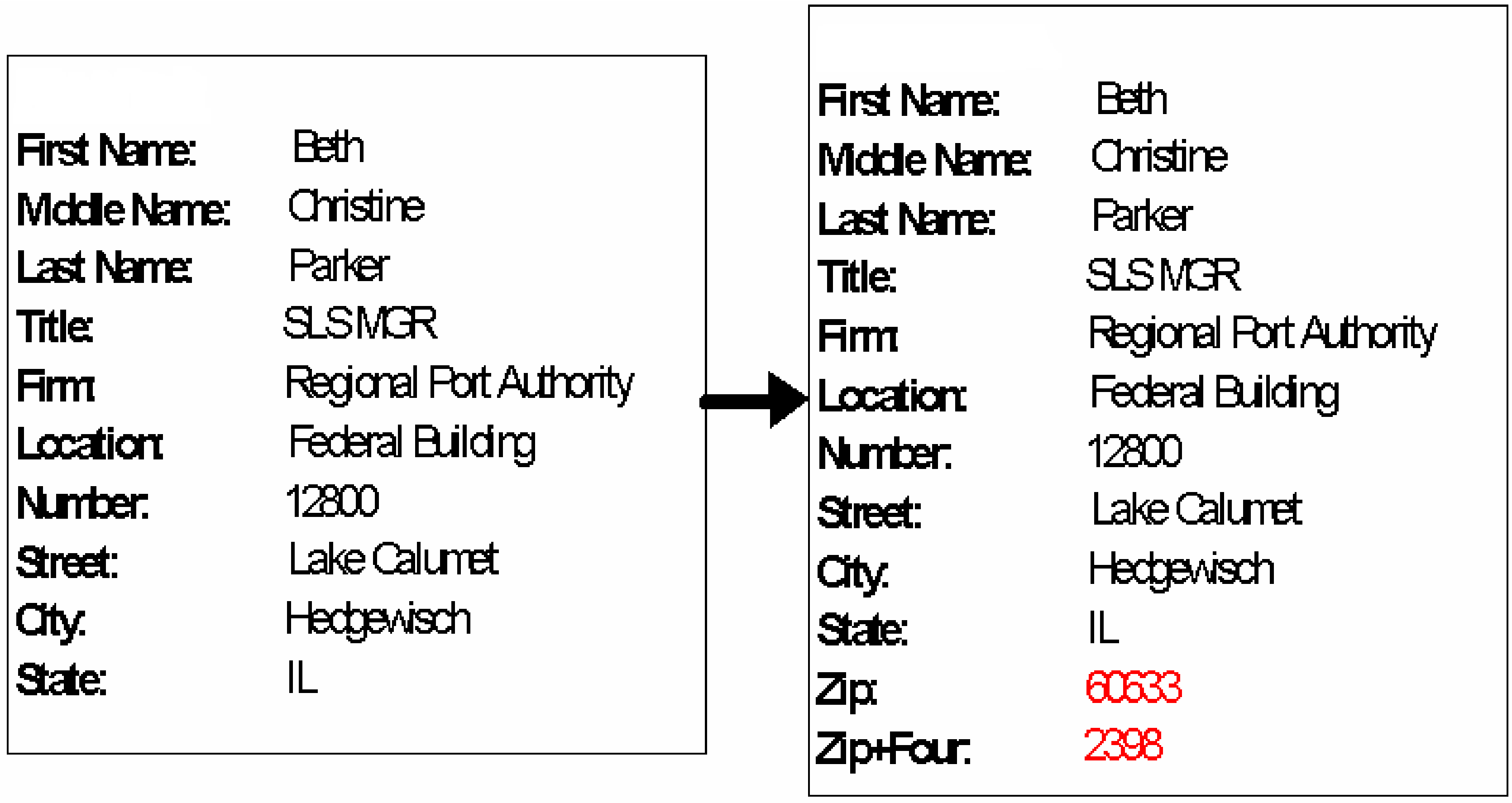
First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL
DOB:	15-Oct-84

Data Transformation

- Enrichment
  - Adding/combining external data values, rules to enrich the information already existing in the data
  - E.g. If we can get a list that provides a relationship between Zip Code, City and State, then if a address field has Zip code 06905 it be safely assumed and address can be enriched by doing a lookup on this table to get Zip Code 06905 –> City Stamford –> State CT

Data Transformation

- Enrichment



First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL
Zip:	60633
Zip+Four:	2398

Data Loading

- Data are physically moved to the data warehouse
- The loading takes place within a “load window”
- Loading the Extracted and Transformed data into the Staging Area or Data Warehouse.

Data Loading

- First time bulk load to get the historical data into the Data Warehouse
- Periodic Incremental loads to bring in modified data
- Design load strategy to using appropriate Slowly Changing Dimension type .
- The Loading window should be as small as possible
- Should be clubbed with strong Error Management process to capture the failures or rejections in the Loading process

Slowly Changing Dimension Types

- Three types of slowly changing dimensions
  - Type 1
    - Updates existing record with modifications
    - Does not maintain history
  - Type 2
    - Adds new record
    - Maintain history
    - Maintains old record
  - Type 3:
    - Keep old and new values in the existing row
    - Requires a design change

Meta Data

- Data about data
- Needed by both information technology personnel and users
- IT personnel need to know data sources and targets; database, table and column names; refresh schedules; data usage measures; etc.
- Users need to know entity/attribute definitions; reports/query tools available; report distribution information; help desk contact information, etc.

Metadata

- Metadata is more comprehensive and transcends the data.
  - Metadata provide the ***format and name*** of data items
  - It actually provides the ***context*** in which the data element exists.
  - provides information such as the ***domain*** of possible values;
  - the ***relation*** that data element has to others;
  - the data's ***business rules***,
  - and even the ***origin of the data***.

Importance of Metadata

- Metadata establish the context of the Warehouse data
- Metadata facilitate the Analysis Process
- Metadata are a form of Audit Trail for Data Transformation
- Metadata Improve or Maintain Data Quality

Feature of ETL Tools

- Support data extraction, cleansing, aggregation, reorganization, transformation, and load operations
- Generate and maintain centralized metadata
- Filter data, convert codes, calculate derived values, map source data fields to target data fields
- Automatic generation of ETL programs
- Closely integrated with RDBMS
- High speed loading of target data warehouses using Engine-driven ETL Tools

Advantages of using ETL Tools

- GUI based design of jobs – ease of development and maintenance
- Generation of directly executable code
- Engine driven technology is fast, efficient and multithreaded
- In-memory data streaming for high-speed data processing
- Products are easy to learn and require less training

Advantages of using ETL Tools

- Automatic generation and maintenance of open, extensible metadata
- Support for multiple data formats and platforms
- Large number of vendor supplied data transformation objects

Example of ETL requirements

- Integration of masters across different systems
  - E.g. State code AP could mean Andhra Pradesh in one system while it could mean Arunachal Pradesh in another
- De-duplication of data from different systems
  - E.g. State Karnataka could be represented as KA in one system and KN in another system
- Mapping of old codes to Data Warehouse codes
- Data Cleansing - Changing to upper case, assigning defaults to unavailable data elements

Summary

- In this module, you learned about the following:
  - ETL process
  - Cleansing steps