

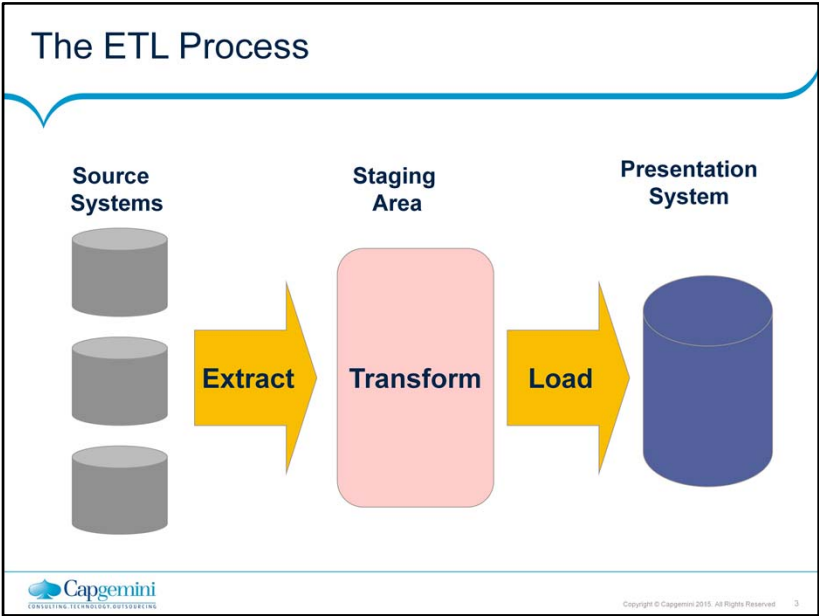
ETL Basics

Lesson 2: ETL Process

Lesson Objectives

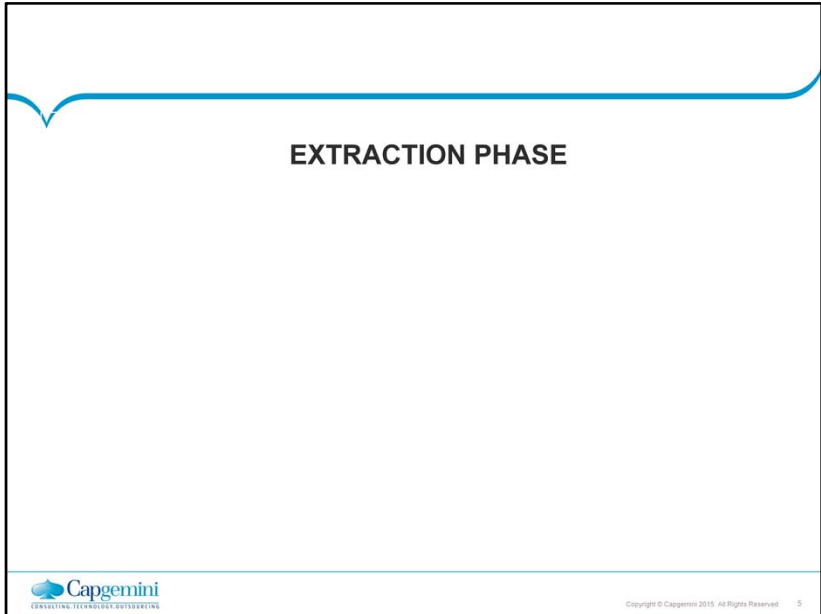
- On completion of this lesson on Data Modeling, you will be able to understand:
 - The ETL process
 - The steps in Data Cleansing



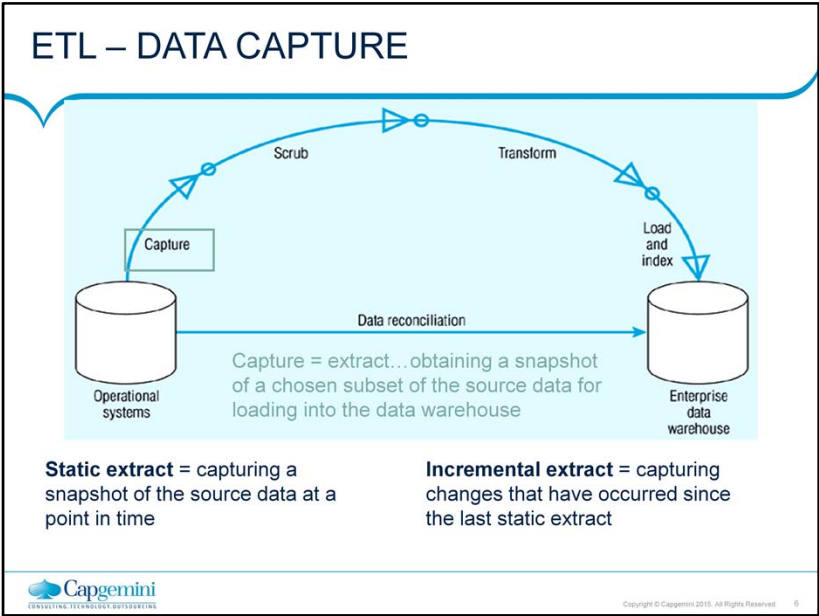


The ETL Process

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



Now Let' s go through that how Transforming data will take place in the Data Warehousing environment



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

Data staging is used in cleansing, transforming, and integrating the data.

Reasons for “Dirty” Data

- Dummy Values
- Absence of Data
- Multipurpose Fields
- Cryptic Data
- Contradicting Data
- 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.

Though the extraction process can be done in either of the methods i.e either by hand coded methods or by using the tools. Tool based extraction have a well defined approach with a better documentation and it also makes the extraction process easier by a simple click, drag and drop features that are more user-friendly to the programmers.

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.

Bulk extraction needs the entire data warehouse to be refreshed periodically in which the entire data which is there in the data warehouse and the data to be loaded in to the warehouse are loaded once again in to the warehouse which uses heavy network traffic. But this mechanism is much easier 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

Aggregates, such as sales totals, are often precalculated and stored in the warehouse to speed queries that require summary totals.

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 cleansing is critical to customer relationship management initiatives.

Data Transformation

■ Steps in Data Cleansing

- Parsing
- Correcting
- Standardizing
- Matching
- Consolidating
- Conditioning
- Enrichment

A good example to use is cleansing customer data. Most students can identify with receiving multiple copies of the same catalog because the company is not doing a good data cleansing job.

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

The record is broken down into atomic data elements.

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

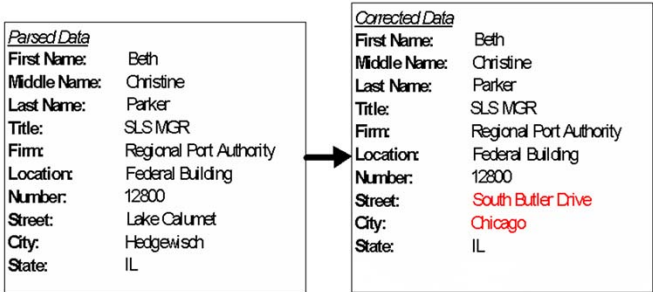
Data Transformation

- 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.

External data, such as census data, is often used in this process.

Data Transformation

■ Correcting



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.

Companies decide on the standards that they want to use.

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.

Commercial data cleansing software often uses AI techniques to match records.

Data Transformation

■ Matching

<u>Corrected Data (Data Source #1)</u>	
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



<u>Corrected Data (Data Source #2)</u>	
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

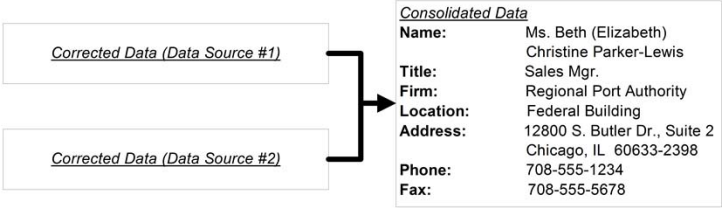
Data Transformation

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

All of the data are now combined in a standard format.

Data Transformation

■ Consolidating



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: SLSMGR
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: SLSMGR
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.

Most loads involve only change data rather than a bulk reloading of all of the data in the 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.

The importance of meta data is now realized, even though creating it is not glamorous work.

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**.



Copyright © Capgemini 2010. All Rights Reserved. 37

Metadata is the high level core internal document of the source code which runs as the lifeblood for a data warehouse.

Metadata not only describe the format and name but it provides details about the context I,e what is the need of the data item and what are the values that the data item can have, the relationship between the data elements ie whether the data element is found on other locations and how they are inter-linked to each other. Apart from the technical details It also holds the business rule. The origin of the data is so critical that the end user might like to trace back to the origin of the data which end user sees through the OLAP tools.

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



Copyright © Capgemini 2010. All Rights Reserved. 38

Importance of Metadata

Metadata establish the context of the Warehouse data

Metadata helps data warehouse administrators and users locate and understand data items, both in the source systems and in the warehouse data structures.

E.g.: The date 02/05/2010 could mean either May 2, 2010 or February 5, 2010 depending on the date convention used. Metadata describing the format of this date field could help determine the definite and unambiguous meaning of the data item.

Metadata facilitate the Analysis Process

Metadata must provide data warehouse end-users with the information they need to easily perform the analysis steps. It should thus allow users to quickly locate data that are in the warehouse.

Metadata should allow analysts to interpret data correctly by providing information about data formats and data definitions.

Metadata are a form of Audit Trail for Data Transformation

Metadata document the transformation of source data into warehouse data. Hence warehouse metadata must be capable of explaining how a particular piece of warehouse data was derived from the operational systems.

All business rules governing the transformation of data to new values or new formats are also documented as metadata.

Metadata Improve or Maintain Data Quality

Metadata can improve or maintain warehouse data quality through the definition of valid values for individual warehouse data items. Using a data quality tool prior to actual loading into the warehouse, the warehouse load images can be reviewed to check for compliance with valid values for key data items. Data errors are quickly highlighted for correction.

Metadata can be used as the basis for any error-correction processing that should be done if a data error is found. Error-correction rules are documented in the metadata repository and executed by program code on an as needed basis.

All business rules governing the transformation of data to new values or new formats are also documented as metadata.

Metadata Improve or Maintain Data Quality

Metadata can improve or maintain warehouse data quality through the definition of valid values for individual warehouse data items. Using a data quality tool prior to actual loading into the warehouse, the warehouse load images can be reviewed to check for compliance with valid values for key data items. Data errors are quickly highlighted for correction.

Metadata can be used as the basis for any error-correction processing that should be done if a data error is found. Error-correction rules are documented in the metadata repository and executed by program code on a need basis.

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


✓

✓

✓

✓

Summary



Capgemini
TRANSFORMING TECHNOLOGY INTO BUSINESS

Copyright © Capgemini 2019. All Rights Reserved. 46

Add the notes here.