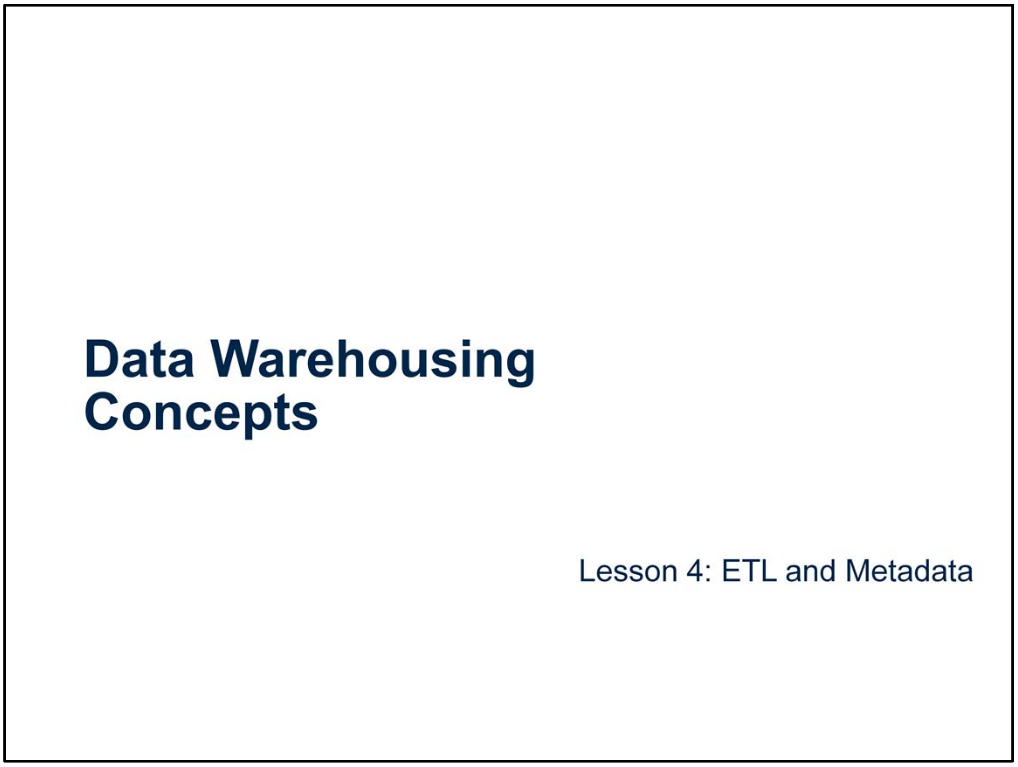
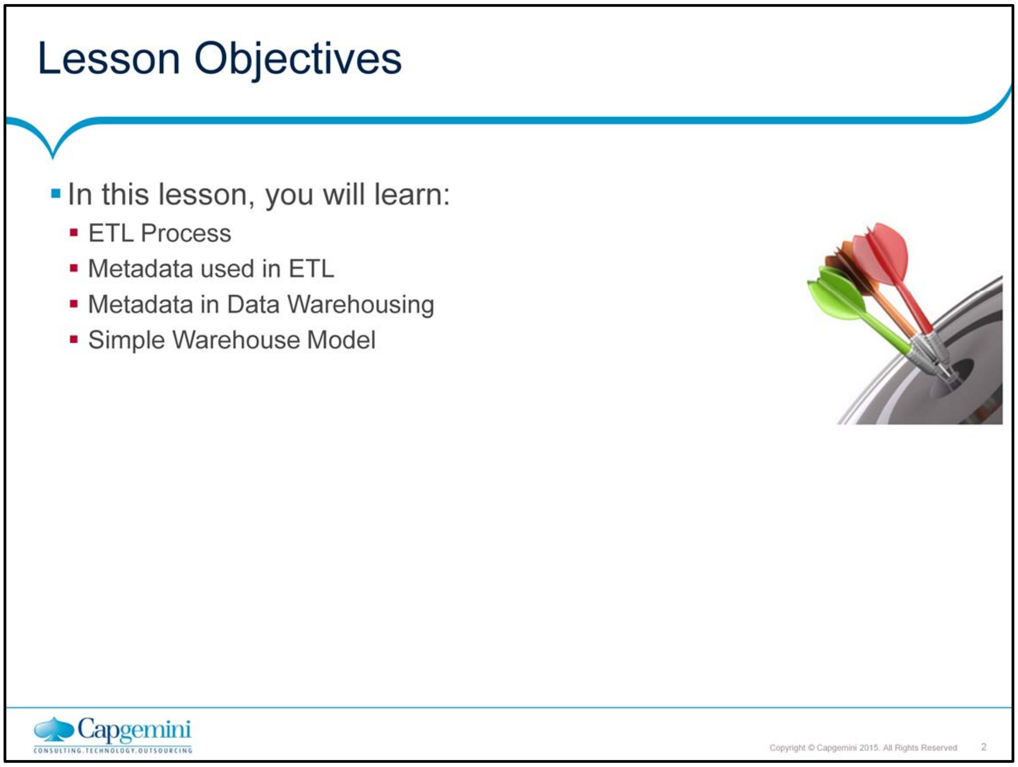


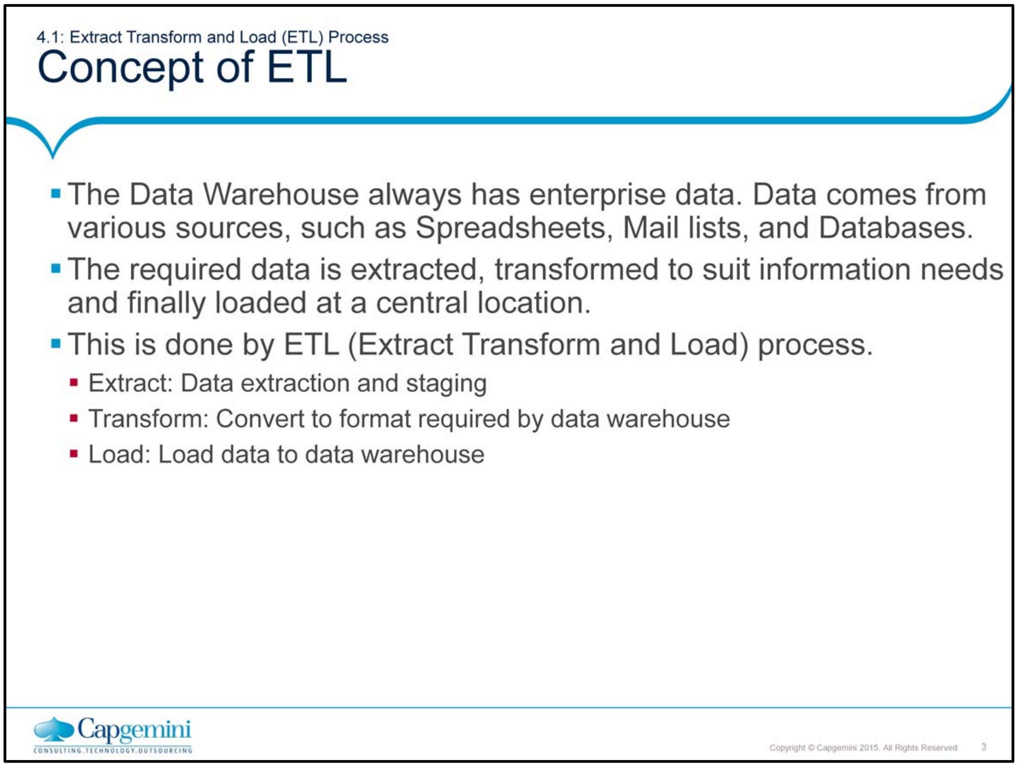
Data Warehousing Concepts

Lesson 4: ETL and Metadata



Lesson Objectives

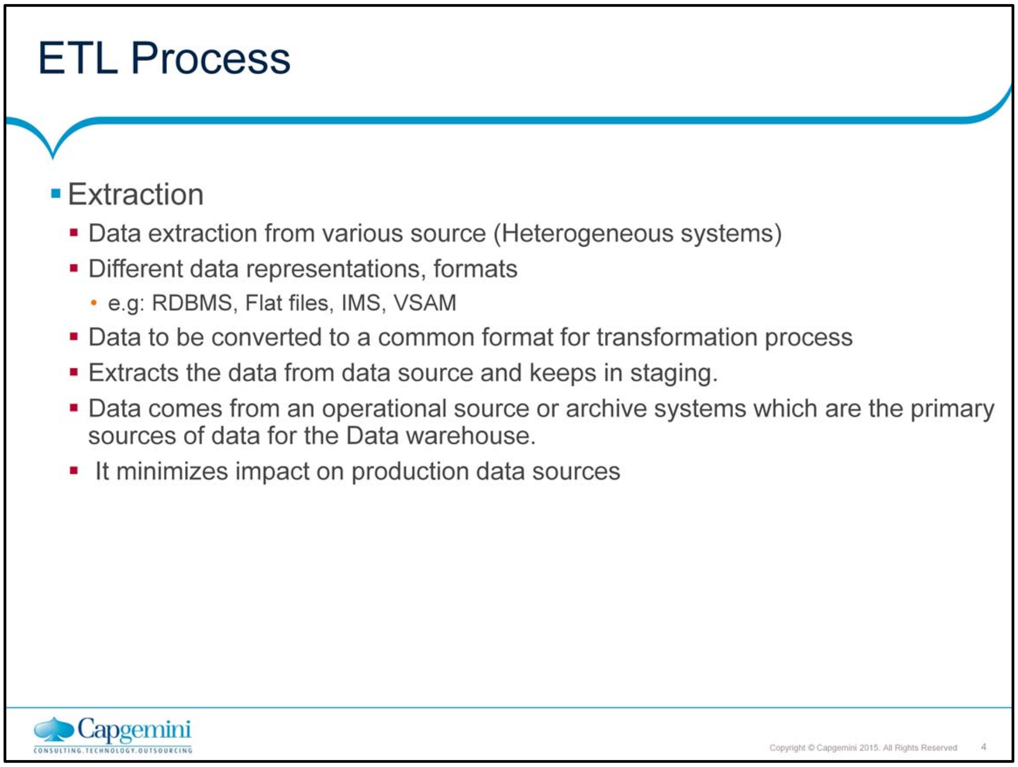
- In this lesson, you will learn:
  - ETL Process
  - Metadata used in ETL
  - Metadata in Data Warehousing
  - Simple Warehouse Model



4.1: Extract Transform and Load (ETL) Process

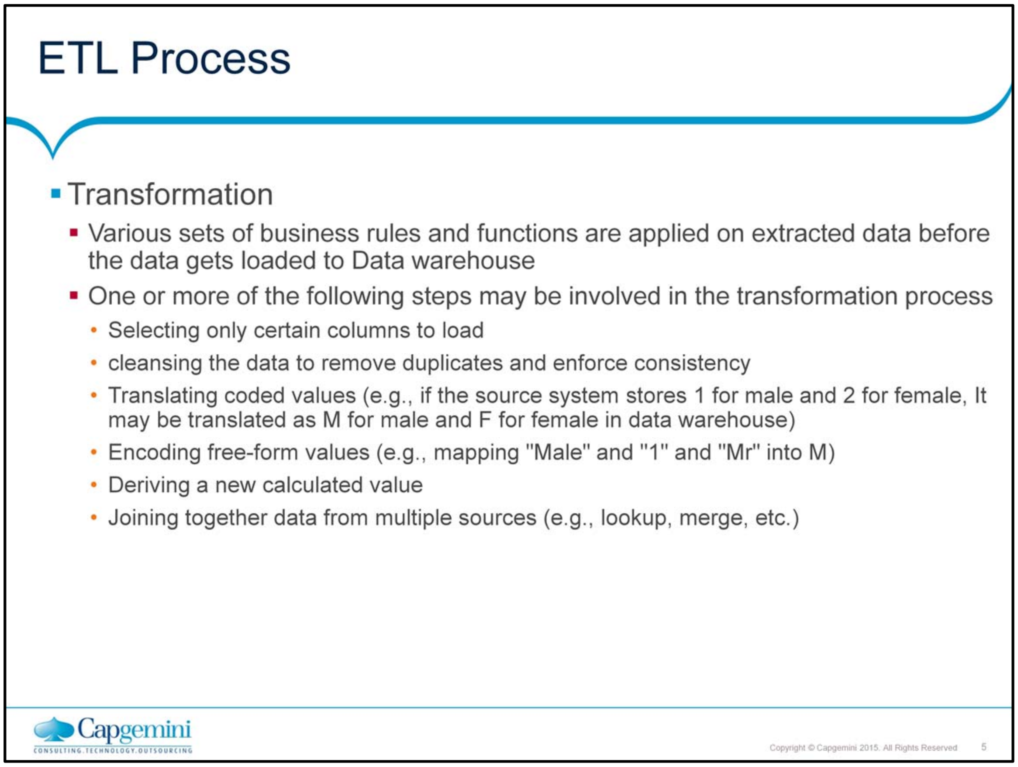
Concept of ETL

- The Data Warehouse always has enterprise data. Data comes from various sources, such as Spreadsheets, Mail lists, and Databases.
- The required data is extracted, transformed to suit information needs and finally loaded at a central location.
- This is done by ETL (Extract Transform and Load) process.
  - Extract: Data extraction and staging
  - Transform: Convert to format required by data warehouse
  - Load: Load data to data warehouse



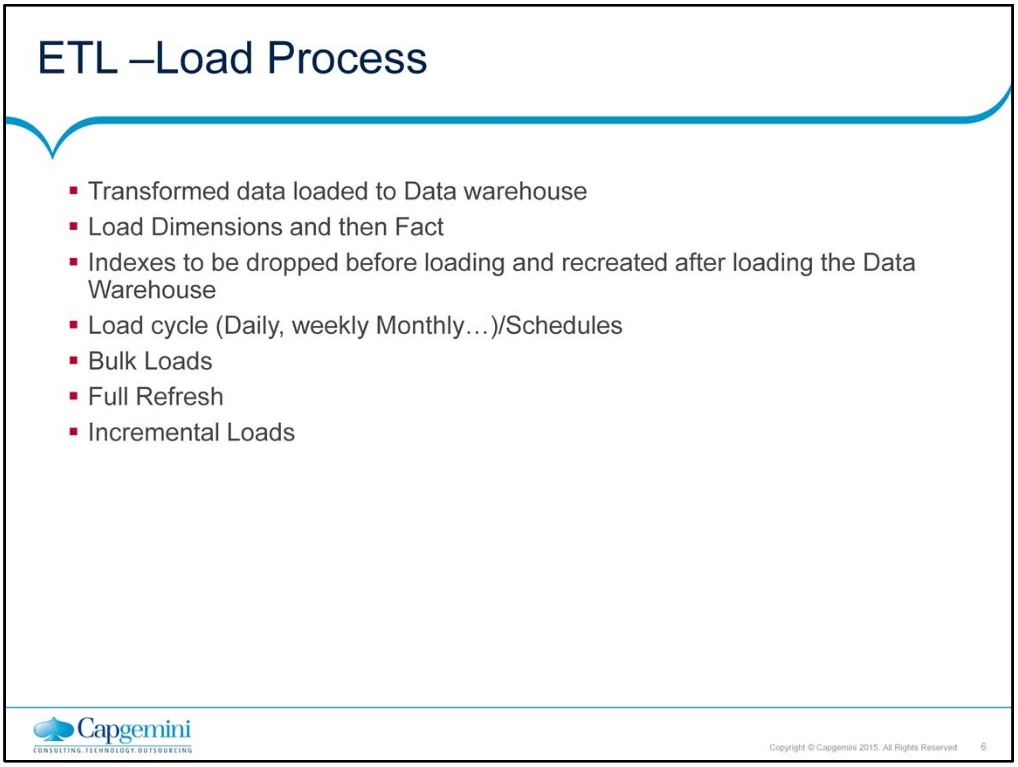
ETL Process

- Extraction
  - Data extraction from various source (Heterogeneous systems)
  - Different data representations, formats
    - e.g: RDBMS, Flat files, IMS, VSAM
  - Data to be converted to a common format for transformation process
  - Extracts the data from data source and keeps in staging.
  - Data comes from an operational source or archive systems which are the primary sources of data for the Data warehouse.
  - It minimizes impact on production data sources



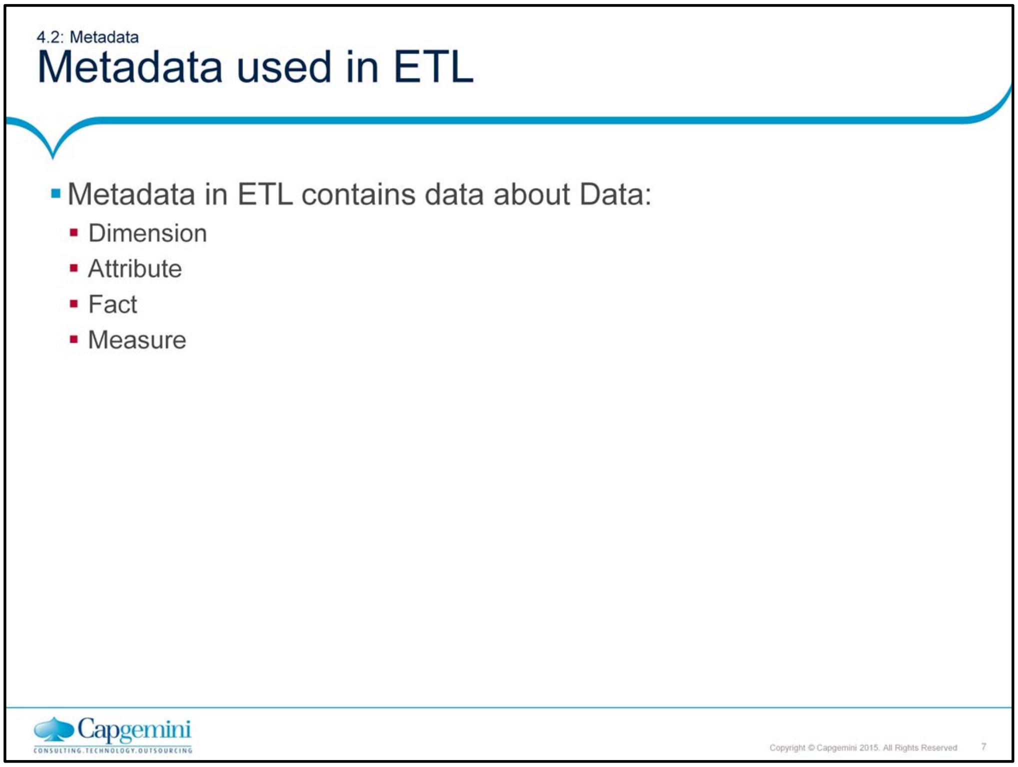
ETL Process

- Transformation
  - Various sets of business rules and functions are applied on extracted data before the data gets loaded to Data warehouse
  - One or more of the following steps may be involved in the transformation process
    - Selecting only certain columns to load
    - cleansing the data to remove duplicates and enforce consistency
    - Translating coded values (e.g., if the source system stores 1 for male and 2 for female, It may be translated as M for male and F for female in data warehouse)
    - Encoding free-form values (e.g., mapping "Male" and "1" and "Mr" into M)
    - Deriving a new calculated value
    - Joining together data from multiple sources (e.g., lookup, merge, etc.)



ETL –Load Process

- Transformed data loaded to Data warehouse
- Load Dimensions and then Fact
- Indexes to be dropped before loading and recreated after loading the Data Warehouse
- Load cycle (Daily, weekly Monthly…)/Schedules
- Bulk Loads
- Full Refresh
- Incremental Loads



4.2: Metadata

Metadata used in ETL

- Metadata in ETL contains data about Data:
  - Dimension
  - Attribute
  - Fact
  - Measure

7

Metadata:

Metadata is the data about Data.

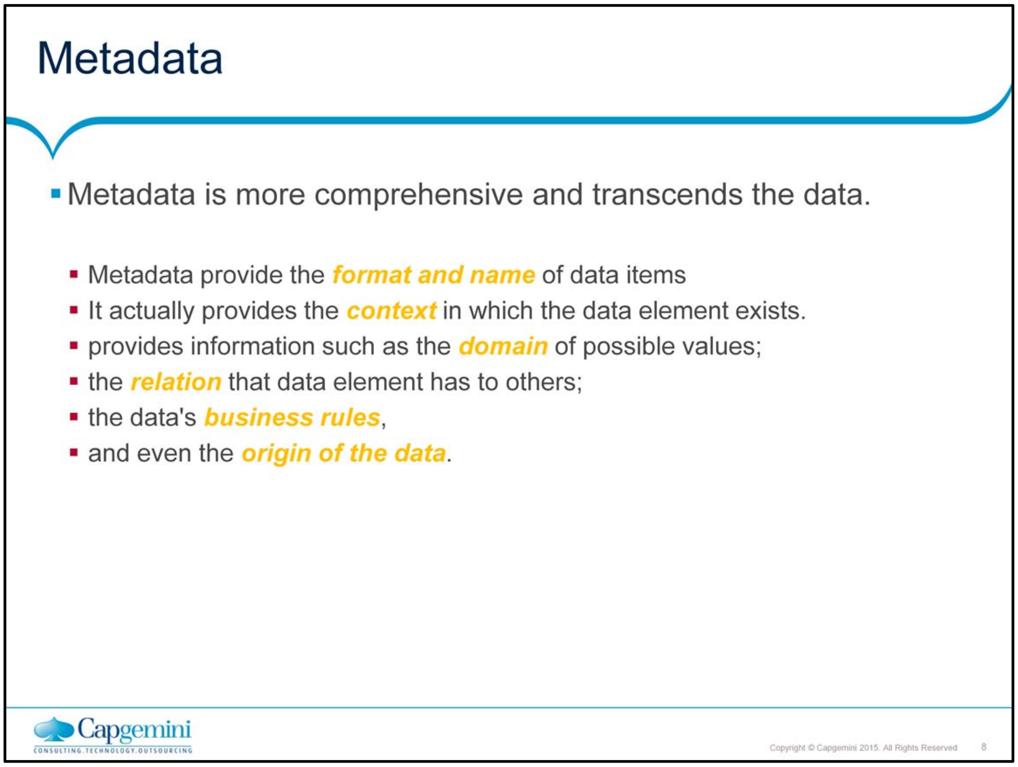
- **Dimension:** It is a perspective that can be used to analyze the data. Dimensions become more useful when there are many descriptive attributes that can be used for analyzing the data.
- **Attribute:** It is often used to describe the extended Dimension.

  Example: Customer, Item, Date, Fact
- **Fact:** It is the raw enumerable piece of information about the transaction. It is always a numeric value (usually aggregatable) about the transaction.

  Examples: Quantity, Unit Price, Count
- **Measure:** Measure can be the product of one or more fact tables. Measure can be the result of any formula which is derived from Relational Database or Business Intelligences Tool analytical engine. It is the product of one or more Facts.

  Examples: Quantity, Unit Price, Count, Quantity * Unit Price, Average (Unit Price) and Minimum (Quantity).

  For example, if a customer is an attribute from your databases table, then customer metadata can give the information about the customer like name, address will be the dimensions, whereas telephone number can act as fact, etc. Age can considered as measure, since it can be calculated from DOB-Current date.

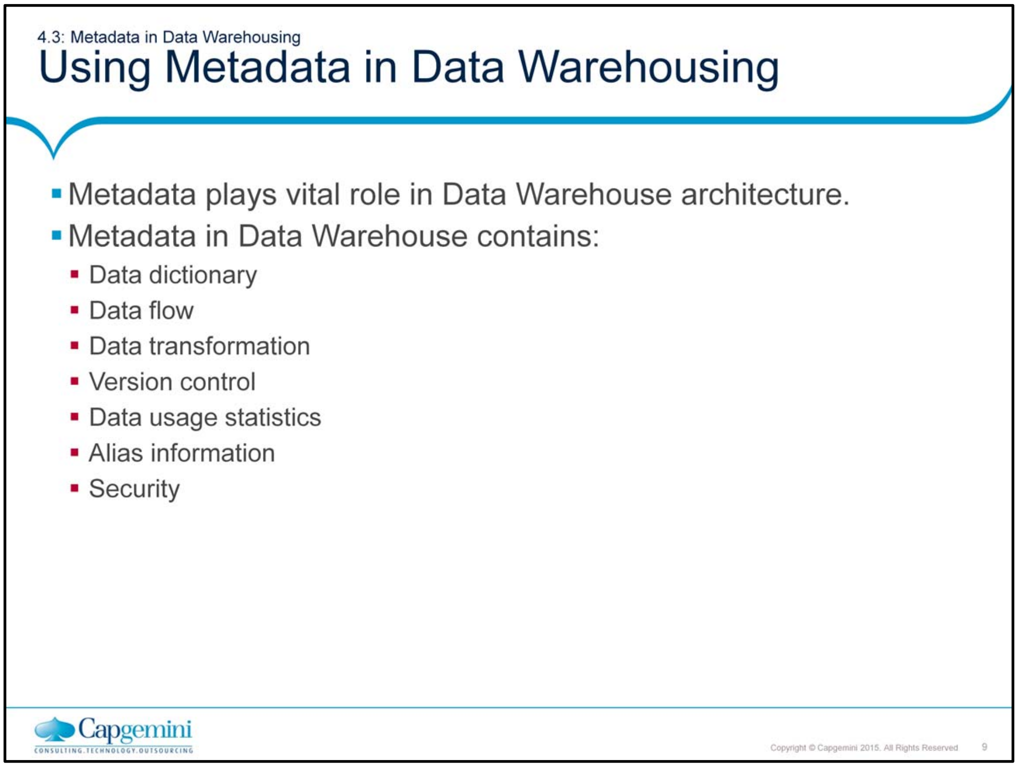


Metadata

- 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***.

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.



4.3: Metadata in Data Warehousing

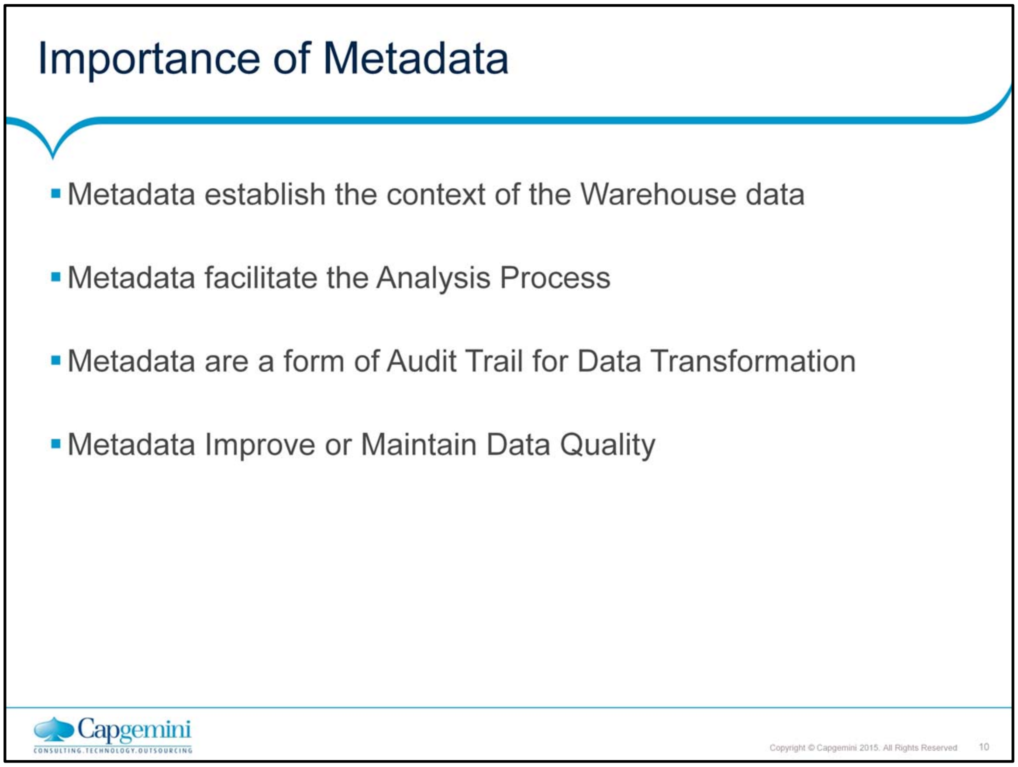
Using Metadata in Data Warehousing

- Metadata plays vital role in Data Warehouse architecture.
- Metadata in Data Warehouse contains:
  - Data dictionary
  - Data flow
  - Data transformation
  - Version control
  - Data usage statistics
  - Alias information
  - Security

Capgemini CONSULTING.TECHNOLOGY.OUTSOURCING

Metadata in Data Warehousing:

- ➢ Metadata is the blood of the Data Warehouse. It is the information that describes the system. Metadata plays a vital role in Data Warehouse architecture. It provides the information to the application to control warehouse activities. A single change in the metadata repository affects the entire architecture.
- ➢ Metadata in Data Warehousing:
  - **Data dictionary:** It contains definitions of the databases and relationship between data elements.
  - **Data flow:** It contains direction and frequency of data feed.
  - **Data transformation:** It contains transformations required when data is moved.
  - **Version control:** It records changes to stored metadata.
  - **Data usage statistics:** It is a profile of data in the warehouse.
  - **Alias information:** It contains alias names for a field.
  - **Security:** It contains the names of the data access authorized people.



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

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.

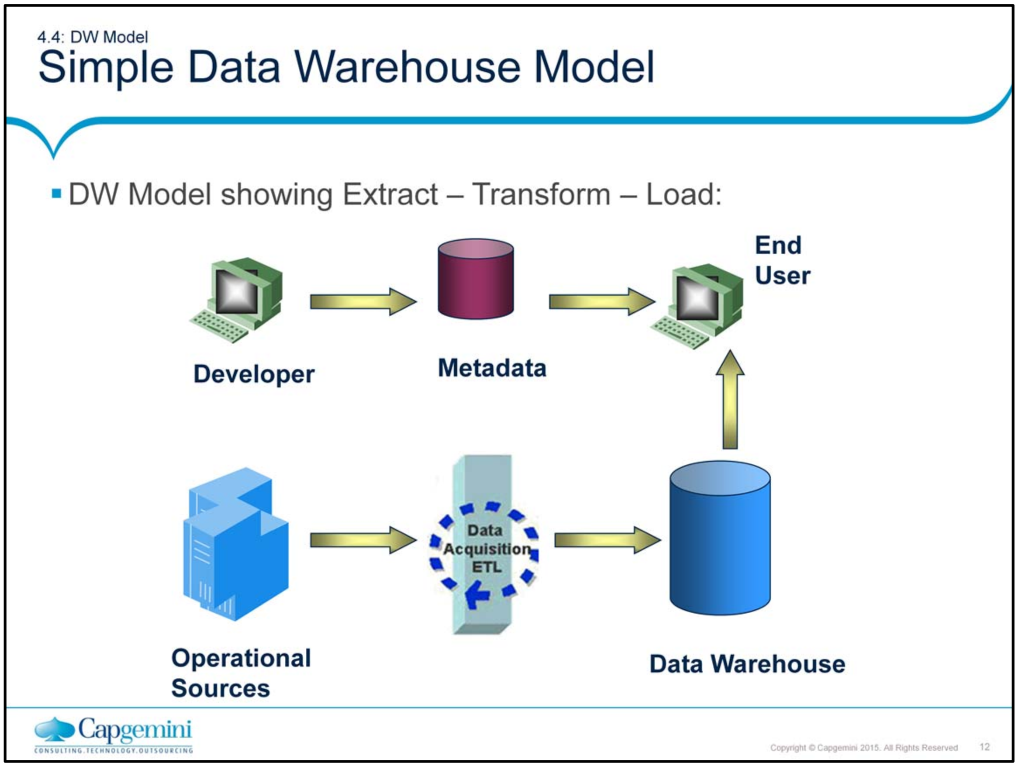
Metadata Improve or Maintain Data Quality
Metadata can improve or maintain warehouse data quality through the definition of valid values for individual warehouse data items. Using a data quality tool prior to actual loading into the warehouse, the warehouse load images can be reviewed to check for compliance 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.



4.4: DW Model

Simple Data Warehouse Model

- DW Model showing Extract – Transform – Load:

Developer

Metadata

End User

Operational Sources

Data Acquisition ETL

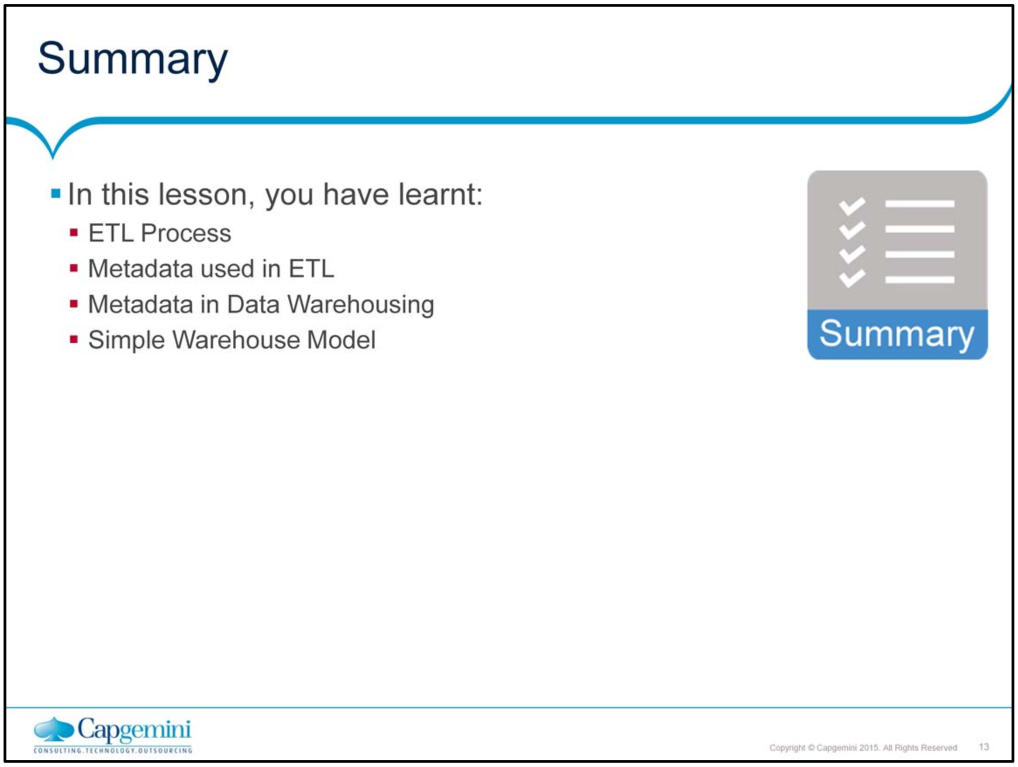
Data Warehouse

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

12

DW Model:

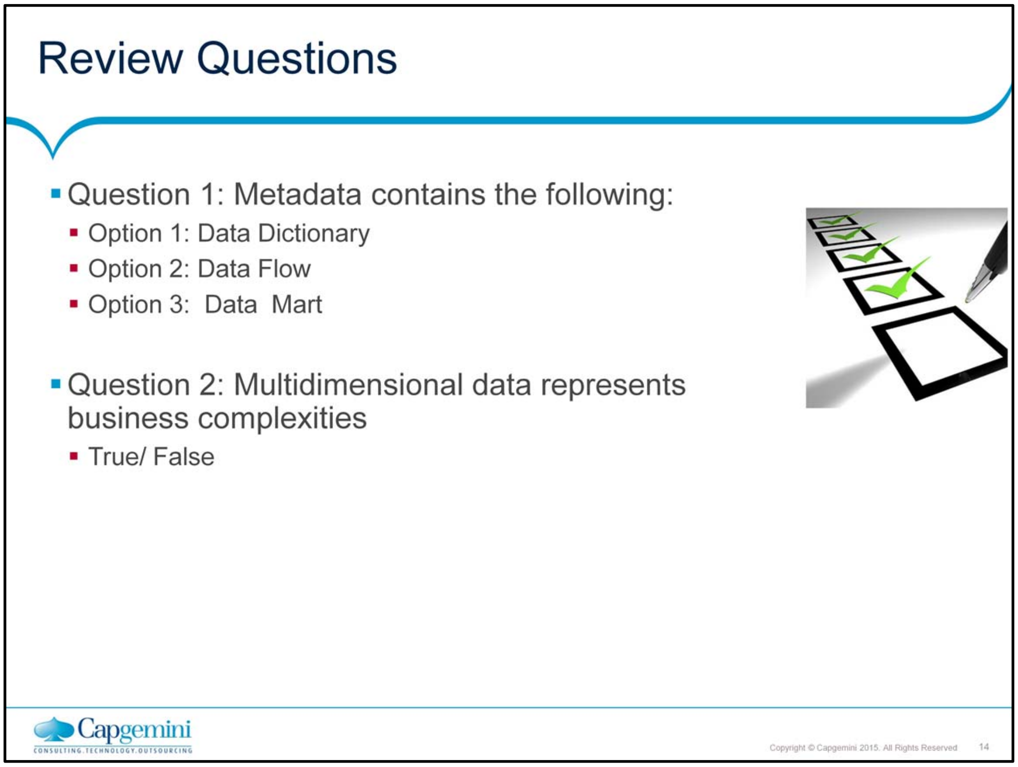
- A Data Warehouse setup typically comprises of the following end points:
  - **Developer:** The developer puts business rules for data transformation into the metadata repository.
  - **Metadata:** It indicates about the data is available in the warehouse and where the data is located.
  - **Data Warehouse:** Data Warehouse integrates and aggregates data from various operational and external databases maintained by different Business Units.
  - **Operational Sources:** It can comprise of Customer Database, Sales Database, and Product Database.
  - **End User:** High performance is achieved by pre-planning the requirement for joins, summations, and periodic reports by end users.



Summary

- In this lesson, you have learnt:
  - ETL Process
  - Metadata used in ETL
  - Metadata in Data Warehousing
  - Simple Warehouse Model

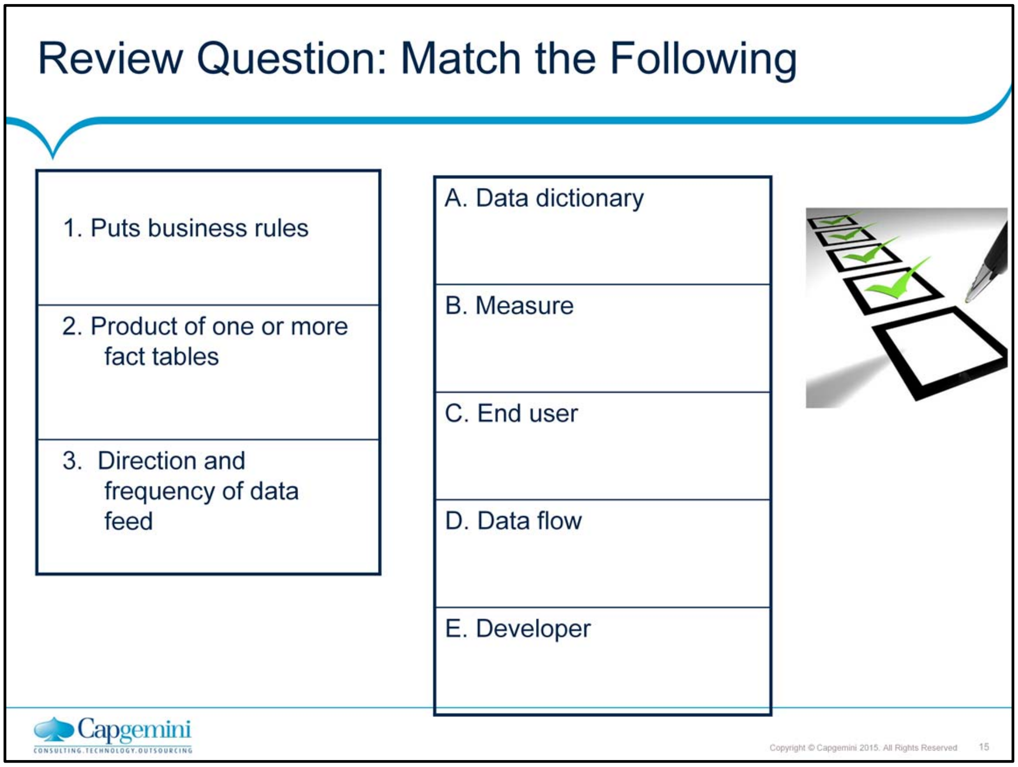
13



Review Questions

- Question 1: Metadata contains the following:
  - Option 1: Data Dictionary
  - Option 2: Data Flow
  - Option 3: Data Mart
- Question 2: Multidimensional data represents business complexities
  - True/ False

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING



Review Question: Match the Following

1. Puts business rules

2. Product of one or more fact tables

3. Direction and frequency of data feed

A. Data dictionary

B. Measure

C. End user

D. Data flow

E. Developer