


Data Warehousing Concepts

Lesson 4: ETL and Metadata

June 15, 2014

Proprietary and Confidential

- 1 -


Speed. Agility. Integration.

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

June 12, 2014

Proprietary and Confidential

- 3 -

IGATE
Speed. Agility. Integration.

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

June 15, 2014

Proprietary and Confidential

- 4 -

IGATE
Speed. Agility. Integration.

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.)

June 12, 2014

Proprietary and Confidential

- 5 -

IGATE
Speed. Agility. Innovation.

ETL Process

➤ Loading

- 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...)

June 12, 2014

Proprietary and Confidential

- 6 -

IGATE
Speed. Agility. Innovation.

4.2: Metadata

Metadata used in ETL

➤ Metadata in ETL contains data about Data:

— Dimension

— Attribute

— Fact

— Measure

June 10, 2014

Proprietary and Confidential

• 7 •

IGATE
Speed. Agility. Integration.

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 be considered as measure, since it can be calculated from DOB-Current date.

4.3: Metadata in Data Warehousing
Using Metadata in Data Warehousing

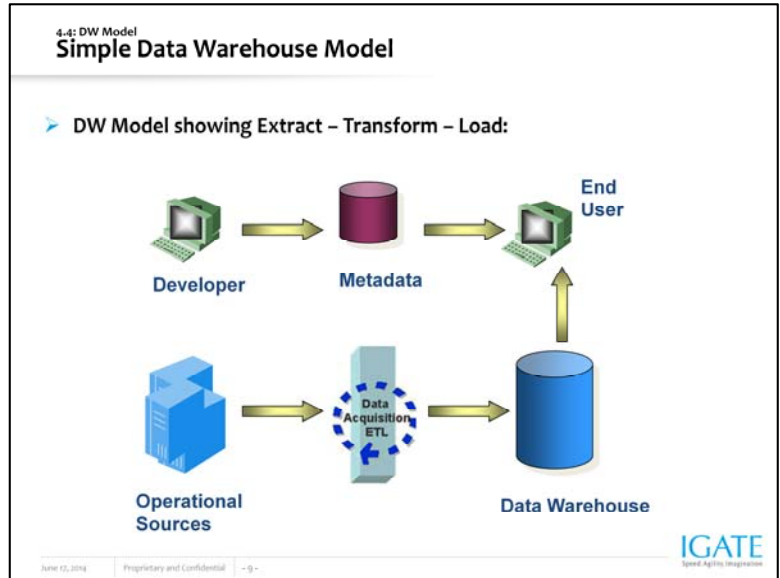
- **Metadata plays vital role in Data Warehouse architecture.**
- **Metadata in Data Warehouse contains:**
 - Data dictionary
 - Data flow
 - Data transformation
 - Version control
 - Data usage statistics
 - Alias information
 - Security

June 10, 2014 Proprietary and Confidential - 8 -

IGATE
Speed. Agility. Integration.

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.

**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



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



Review Question: Match the Following

1. Puts business rules	A. Data dictionary
2. Product of one or more fact tables	B. Measure
3. Direction and frequency of data feed	C. End user
	D. Data flow
	E. Developer

