Information Policy

Data Warehouse

D05. ETT

- Code: 164323-03
- Course: Information Policy
- Period: Spring 2013
- Professor: Sync Sangwon Lee, Ph. D

Contents

01. ETT

- ETT(Extraction, Transformation, Transportation)
  - The whole process to load data from source systems to DW including drawing and cleaning
  - Its methodologies differ according to type of source system, extraction period, data volume, loading velocity, quality of source data, type of old data, user's requirements, and so on.
  - Final tables needed in DW:
    - Fact table and summary table

01. ETT

- Types of ETT: Who makes fact/summary tables?
  - Ready-made source system
    - Impossible ← Source systems are mission-critical systems.
  - Semi on-line
    - Extracting data by analyzing log files of DB
    - Reorganizing data at DW
    - Making fact/summary tables at DW
  - On-line
    - Loading linking DB of source systems directly to DW
    - Making fact/summary tables at DW
  - Off-line:
    - Periodically loading SAM files from DB of source systems
    - Delivering SAM files to DW
    - Making fact or summary tables at DW

01. ETT

- Importance of ETT
  - ETT is a CSF in implementing DW.
  - Why is ETT difficult?
    - Heterogeneity of source systems of each company
      - ➔ Absence of a standard solution for ETT
    - Inferior hardware for huge data
    - A lot of errors of source data
    - Long time required for coding source programs when there are many source systems
    - Difficult to update mapping tables for time-variant data
    - Time-variant data: organization code, product code, …

01. ETT

- ODS (Operational Data Store)
  - A data store in the middle phase to fact tables of DW
  - Storing extracted and refined source data
  - Taking charge of ETT at DW
  - Satisfying user's requirements
  - Also called as Integrated DB or Staging DB
  - ODS has raw data ➔ speedy response to user's requirements
  - ODS is recoverable fast when problems of fact tables arise.

02. Initial Data vs. Periodic Data

- Initial data
  - Data for the past several years
  - Generally stored on magnetic tapes ➔ offline
  - Real situation
    - The record layouts of the old data cannot be stored.
    - The source programs cannot be backed up.
      - ➔ It's impossible to understand the structure of source data.
    - Most of ISD(IS Department) tends to ignore data management rather than system management.

02. Initial Data vs. Periodic Data

- Periodic data
  - Data periodically delivered from source systems to DW while the systems operate.
  - The offline is desirable when the volume of data are large.
    - ← The speed is important.

03. Data Extraction

- Correctness of Extracted Data
  - It's not easy to guarantee data correctness in implementing DW.
  - Verification of correctness
    - Checking the number of data extracted from source systems
    - Summarizing the value of specific fields
    - Verifying the number and summary value of fields, after loading on DW
    - Rectifying error data

04. Data Transformation

- Data Transformation
  - It is impossible to use the data of source systems at DW.
    - The fact tables of DW are specially formatted.
    - The user's requirements are complex.
      - It is inevitable to associate several source systems.
      - Ex: When referring to total sales and credit sales in sales subject, it is necessary to extract data from sales management systems and credit sales systems respectively and associate them.
    - Code discordance of source systems
      - Ex: Product code discordance b/t product management systems and credit sales systems

04. Data Transformation

- Data Transformation
  - Source data go through many transformation processes until loading on DW.
    - There are many discordance b/t values of DW and those of source systems. ➔ transformation errors

04. Data Transformation

- Data Refinement
  - It is meaningless that the data loaded on DW are not correct.
  - Ex.
    - Seoul branch code: ‘001’, while that of source: ‘0001’
    - September 31

04. Data Transformation

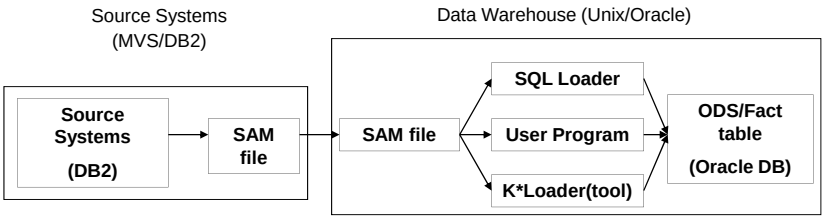
- Data Refinement of Huge Source
  - It is unreasonable to refine them by SQL.
  - A tool for refinement while loading is needed.
    - Ex. Oracle K*Loader (most widely used)
  - Why does DW reply on user's programs as ETT?
    - Infinite variety of data quality
      - ➔ Impossible to map by use of existing tools
    - High Cost

04. Data Transformation

- Data Refinement of Huge Source
  - Functions of Oracle K*Loader
    - Extraction
      - Extracting specific records satisfying specific conditions
      - Extracting specific fields
    - Transformation
      - Loading specific records after mapping according to conversion metadata
    - Validity verification
      - Verifying validity for numbers/characters and loading them
    - Correctness verification
      - Checking specific number fields after loading
    - Logging
      - Storing error records at files while verifying validity (primary key, error field name, error field value)

05. Data Transportation

- Offline Method
  - Data are delivered from source systems in the form of files.
  - Structure



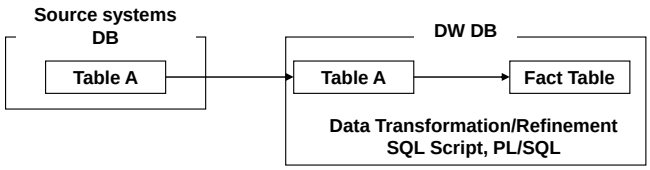
- K*Loader loads data concurrently with refining.
- ODS manipulates data at RDB by use of temporary users.
- SQL Loader loads data directly to fact tables.

05. Data Transportation

- Offline Method
  - Characteristics
    - Creating SAM files from source systems by use of DB2 utility
    - Delivering SAM files to DW server by use of MT and FTP
    - If there are many SAM files or Biz logics are needed to be integrated, user programs are needed.

05. Data Transportation

- Online Method
  - Data are delivered from source systems directly to DW when tables are identically formatted at both source systems and DW
  - Structure



  - Data transformation and refinement: by SQL Script , PL/SQL

05. Data Transportation

- Online Method
  - Open gateway
    - Extracting data from operation systems ➔ delivering them to DW
    - Ex. Oracle distributed option
      - DBLINK, SNAPSHOT

05. Data Transportation

- File Sharing Method
  - Generally used at banks for real-time processing
  - Writing log-files as soon as updated at source systems
  - Delivering data to DW by checking log-files periodically
    - Not perfect real-time processing
    - But almost real-time processing according to periodicity