ETL PROCESS

31.03.2014
(Muscat, Oman)

OUTLINE

- ETL Definition
- Extraction Process
- Transformation Steps
- Loading into Data Warehouse
- ETL Tools
- Case Studies

DEFINITION

•**Component of BI**

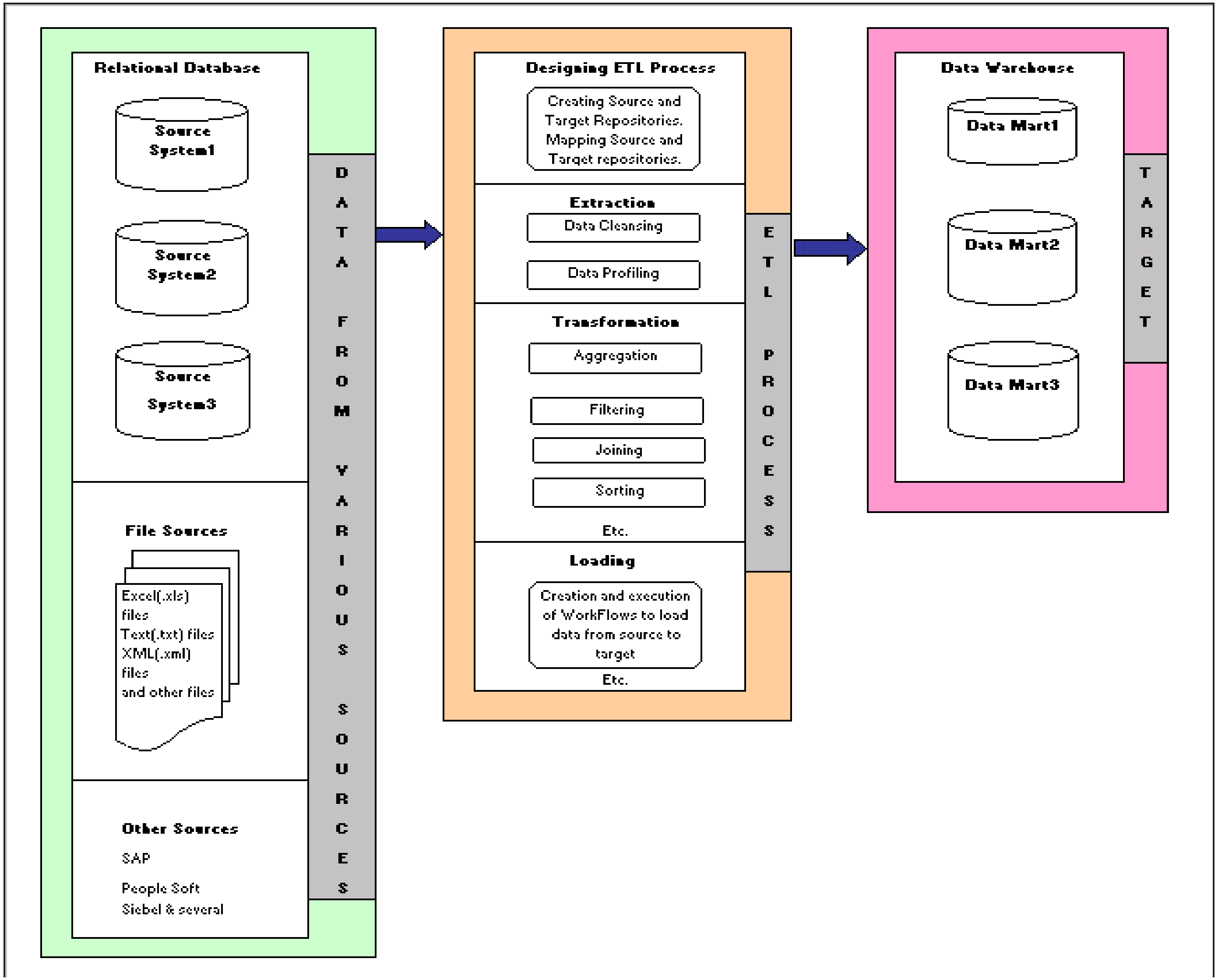
ETL is the set of process that includes
extraction, transformation and loading data

Data warehouses are supplied by ETL processing

Data are moved from sources to target databases

It is A very costly and time consuming part of data warehousing

ELT : Extract – Load – Transform
Difference from ETL is semantic.



Relational Database

Source System1

Source System2

Source System3

File Sources

Excel(.xls) files
Text(.txt) files
XML(.xml) files
and other files

Other Sources

SAP

People Soft

Siebel & several

DATA FROM VARIOUS SOURCES

Designing ETL Process

Creating Source and Target Repositories. Mapping Source and Target repositories.

Extraction

Data Cleansing

Data Profiling

Transformation

Aggregation

Filtering

Joining

Sorting

Etc.

Loading

Creation and execution of WorkFlows to load data from source to target

Etc.

ETL PROCESS

Data Warehouse

Data Mart1

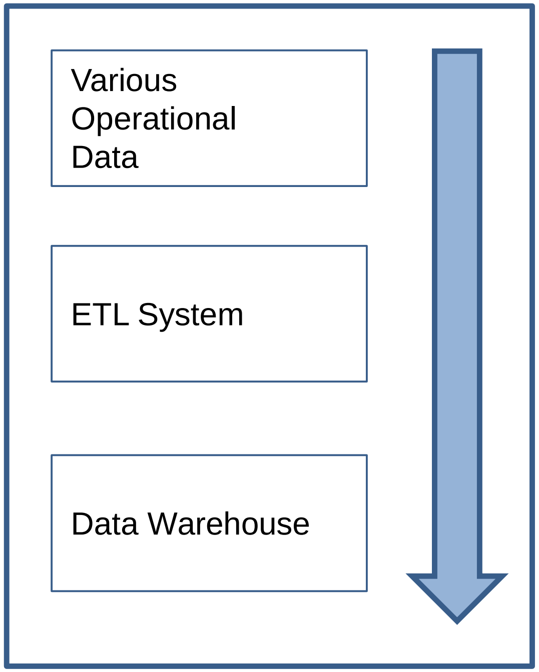
Data Mart2

Data Mart3

TARGET

EXTRACTION

Extraction



Various Operational Data

ETL System

Data Warehouse

- Data is extracted from data sources
- Different systems may exist

  DBMS
  Operating Systems
  Hardware
  Communication protocols

- Needs logical data map document that includes

  rules, sources and target tables definitons

TRANSFORMATION

Transformation is the **main step** where the ETL adds value

Actually changes data and provides guidance for its intended purposes

İncludes : aggregation/disaggregation, sorting, cleaning dirty data,

Checking for data integrity and applying business rules

Could be done by SQL codes or ETL tools

Dirty Data:

What is the dirty data ?

1) Lack of Standardization

- Multiple encoding, locales, languages..
- Different abbreviations (Mareşal Fevzi Çakmak Streeet , Mar. Fev. Ç. Street)
- Semantic equivalence (Gaziantep, Antep, G. Antep)
- Multiple standards : 1.6 miles is the same as 1 kilometer

Because OLTP s are different cities or different countries, different abbreviations or standards may be used)

Dirty Data (Cont’d)

2) Missing, incorrect and duplicate data

- Missing age field for an employee
- Incorrectly entered values
- Duplication of datasets across OLTP units
- Semantic duplication

what is the correlation the age of the staff between their performance ? We need to know age .

3) Inconsistencies

- Incorrect use of codes

  (M/F is used somewhere, 0/1 is used in others for gender)

there could be inconsistent data..

- Referential inconsistency (for example there is 24 as department id, although there isn't a department which id is 24)

Transformation – Data Cleaning

•Dirty data is cleaned in transformation processs.

•While cleaning data,

standardizing is important. So, Companies decide on the standards.

•Data cleaning is not a simple issue

It is not automatically and required considerable knowledge
Complexity increases with increasing data sources

For example ; there is one km in a record; and there is 1 mile in a record.
Or there may be abbreviation of a street
You should know geography, metrics, so on to compare and clean data.

Data Cleaning Steps

1. Data Analysis : Analyse data set to obtain meta-data and detect dirty data
2. Definiton of transformation rules
3. Rule Verification : Verification of the transformation rules on test data sets
4. Transformation : Execution of transformation rules on dataset
5. Back flow: Re-populating data sources with cleaned data

Transformation – Confirming DB Structure

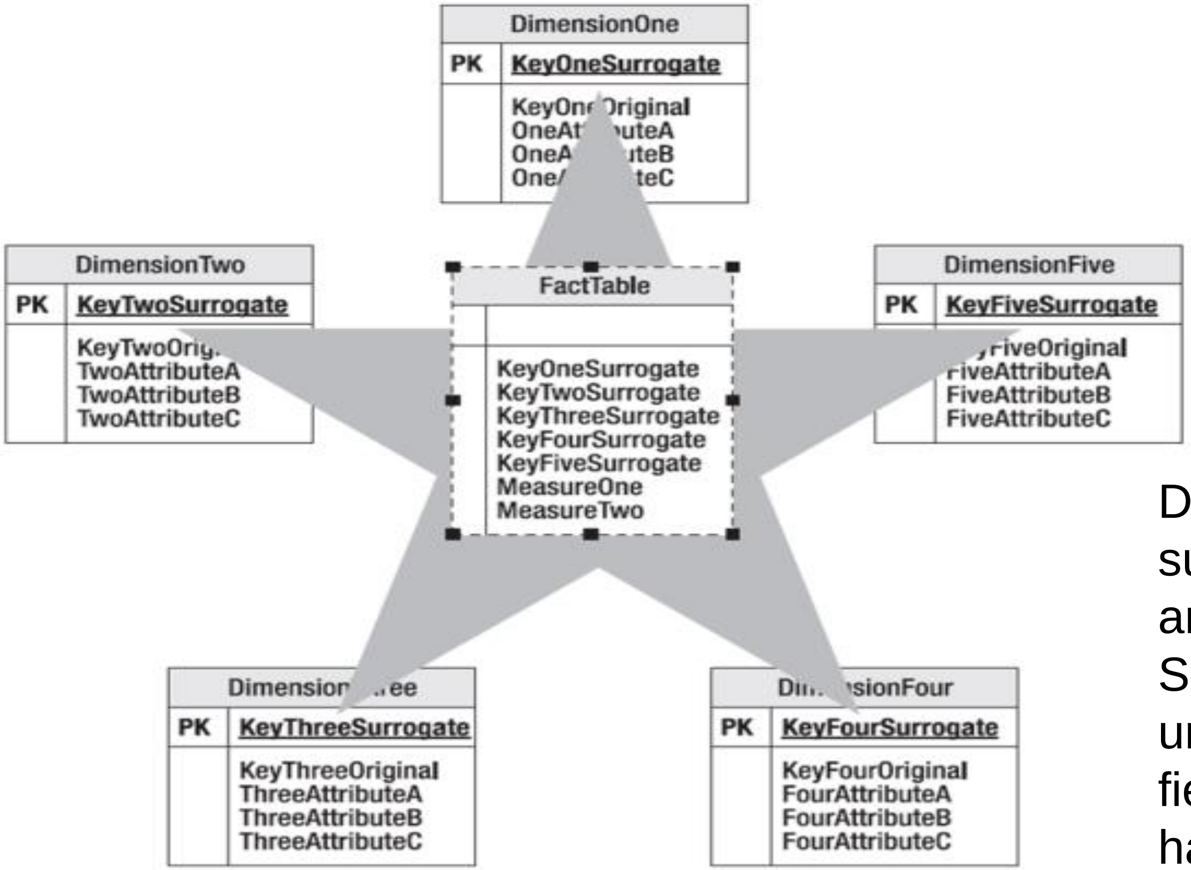
Tables should :

- have proper primary and foreign keys
- obey referential integrity
- simple business rules
- Provide logical data checks

LOADING

- Loading is the third part of the ETL process
- Loading is performing transformed data to the target Data Warehouse
- Data is physically moved to Data Warehouse
- Data Warehouse is a star-schema
- To provide data integrity : first dimension tables and then fact tables are loaded
- Generally only change data is loaded

Loading to DW



DimensionOne

PK **KeyOneSurrogate**

KeyOneOriginal
OneAttributeA
OneAttributeB
OneAttributeC

DimensionTwo

PK **KeyTwoSurrogate**

KeyTwoOriginal
TwoAttributeA
TwoAttributeB
TwoAttributeC

FactTable

KeyOneSurrogate
KeyTwoSurrogate
KeyThreeSurrogate
KeyFourSurrogate
KeyFiveSurrogate
MeasureOne
MeasureTwo

DimensionFive

PK **KeyFiveSurrogate**

KeyFiveOriginal
FiveAttributeA
FiveAttributeB
FiveAttributeC

DimensionThree

PK **KeyThreeSurrogate**

KeyThreeOriginal
ThreeAttributeA
ThreeAttributeB
ThreeAttributeC

DimensionFour

PK **KeyFourSurrogate**

KeyFourOriginal
FourAttributeA
FourAttributeB
FourAttributeC

Dimension tables have a surrogate key, Normal key and attributes,
Surrogate key should be a unique integer, a single field While normal key may have multiple field.

Slowly Changing Dimensions

The "Slowly Changing Dimension" problem is a common one particular to data warehousing.

Briefly, this applies to cases where the attribute for a record changes over time.

We give an example:

Customer Id	Name	City
1001	Ahmet ak	Ankara

Later he moved to İstanbul.
How should company modify its customer table to reflect this change?
This is the "Slowly Changing Dimension" problem.

There are three ways to solve this type of problem, and they are categorized as follows:

Type 1: The new record replaces the original record. No trace of the old record exists.

Type 2: A new record is added into the customer dimension table. Therefore, the customer is treated essentially as two people.

Type 3: The original record is modified, but the old value is also kept.

Type 1 Dimension

The new information simply overwrites the original information. In other words, no history is kept.

New record will be:

Customer Id	Name	City
1001	Ahmet ak	İstanbul

This is the easiest way to handle the Slowly Changing Dimension problem. But the historical data is lost. Company doesn’t know the previous city of the Customer.

This type can be used when history of data is not important.

Type 2 Dimension

A new record is added to the table to represent the new information. Therefore, both the original and the new record will be present. The new record gets its own primary key.

Result will be :

Customer Id	Name	City
1001	Ahmet ak	Ankara
1002	Ahmet ak	İstanbul

In this approach, history of data is kept. But, the size of the table will increase. And ETL process will be complicated.

When tracking the changes is important, this type of dimension could be used.

Type 3 Dimension

In Type 3 Slowly Changing Dimension,
There will be two columns to indicate the particular attribute of interest, one indicating the original value, and one indicating the current value.
There will also be a column that indicates when the current value becomes active (date value)

After Customer moves to İstanbul, the table will be shown as below:

customer_id	name	original_city	current_city	effective_date
1001	Ahmet Ak	Ankara	İstanbul	24.03.2014

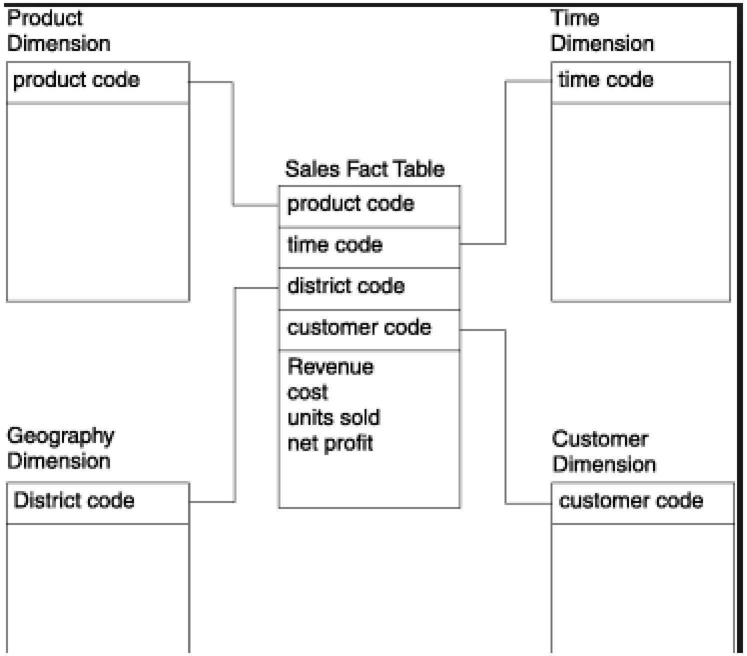
Keeps some historical data, (only 1 historical data, here)
size is not increasing more.
This type is rarely used.

Loading Fact Tables

Fact table consists of

•Dimensions’ keys

•Measures



Metadata Repository

Metada defines warehouse objects

It stores:

Description of the structure of the data warehouse

Operational meta-data

The algorithms used for summarization

The mapping from operational environment to the data warehouse

Data related to system performance

Business data, business rules

Most Popular ETL Tools are:

- Informatica
- Power CenterIBM
- SAP - BusinessObjects Data Integrator
- IBM - Cognos Data Manager
- Microsoft - SQL Server Integration Services
- Oracle - Data Integrator
- SAS - Data Integration Studio
- Oracle - Warehouse Builder

- **Open Source Examples**: Pentaho, Talend,..

Advantages of ETL Tools

The single greatest advantage of an ETL tool is that it provides a visual flow of the system's logic.

It also provides attractive, self documentation.

These tools provide monitoring the ETL system.

Manual coded is also useable in ETL tools.

Finding data dependencies will be easier if needed after or before any change.

These tools have cleanining functionality.

Performance may be better in some situations.

Using an ETL tool will be easier for inexperienced developer.

Disadvantages of ETL Tools

Software licensing cost is the most important disadvantage.

There is uncertainty in many ETL teams. They may use only few features of the Tools. Finding experienced ETL team is difficult.

Sometimes, flexibility is limited. They have some limits and experience is needed.

Developers may resist to use a new tool

So, Should we use an ETL tool ? ..

Because, each company has different data size, business rules, etc.
,, tool necessity should be evaluated according to their criterias.

So decision is **changed** for each company.

Key Points:

✓ETL is often a complex combination of process and technology

✓ETL is not a one-time event

✓It should be performed periodically (monthly, daily, hourly)

✓Should be automated, well **documented** and easily changeable

✓Steps are; data cleaning, integrating and loading

✓First dimensions, then fact table are loaded

✓There are some useful ETL tools but they are not always necessary

Thank you & Questions ?