University of South-Eastern Norway

Lecture 4: The ETL Process

Veralia Sánchez
Associate Professor

Outline

- Data warehouse and data marts
- Start Schema Designing Principles
  - Tutorial: A college star schema
- ETL
  - Extraction
  - Transformation
  - Loading
- Tutorials in Spoon
- Assignment

Why do we need a Data Warehouse?

- All information is in one place
- Up-to-date information
- Quick access
- No size limits
- All history available
- Easy to understand
- Clear and uniform definitions
- Santardized data

Data Marts

- To meet the specific needs of an organisation, a data mart may cover only a particular process and be limited to the boundaries of that process.
  - You won’t find employee absence information in a slaes data mart, because a sales nalayst doesn’t need that information
- However, there is no limitation to the amount or type of data that may be included in a data mart.

Start Schema Designing Principles

Surrogate keys

- Surrogate keys
  - There is always only a single column key for each dimension table
  - Intenger indexes are usually a lot faster than character or datetime indexes
  - Enable the storage of multiple versions of an item where the item retains its original source key but is allotted a new surrogate key
  - Allow for dealing with optional relations, unknown values and irrelevant data.

Naming and Type Conventions

- All tables get a prefix
  - STG_ for staging tables
  - HIS_for historical archive tables
  - DIM_ for dimension tables
  - FCT_for fact tables
  - AGG_for aggretate tables
  - LKP_for lookup tables
- Use meaningful names for the column
- Avoid the use of reserved words for database objects as tables.

Granularity

Granularity: The level of detail at which the data is stored in the data warehouse.

Golden rule: **Store the data at the lowest level of detail possible**.

- Example 1: for a retail company, this means individual sales transaction level

- Example 2: for a mobile operator: it is the call detail record level

Fact table

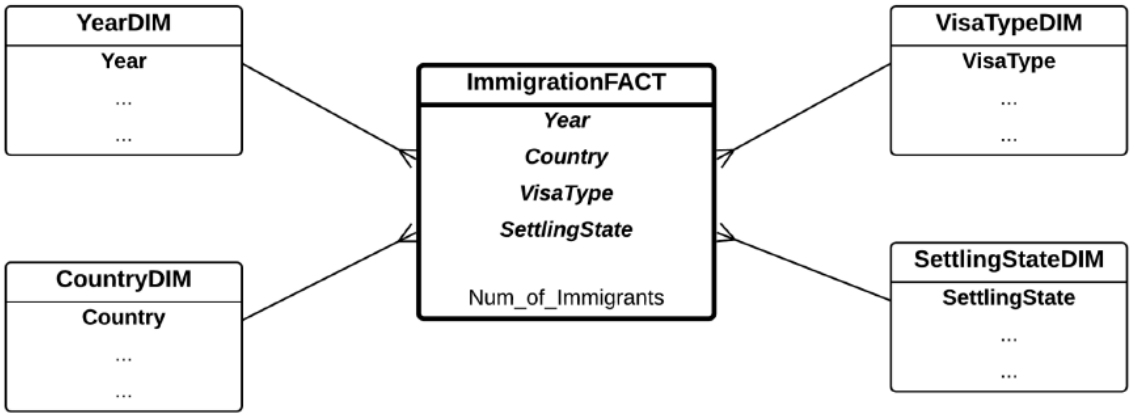
1) Identify the business process for analysis
   - Sales,
   - Order processing
2) Declare the grain (Level of analysis)
   - Transaction,
   - Order,
   - Order lines,
   - Daily,
   - Daily + location
3) Identify dimensions that are relevant
   - What, when, where and why
   - Time, location, products, customers,
   - Filtering, grouping
4) Identify facts for measurements

Two-Column Table Methodology

- To check the correctness of a star schema.
  - *Imaginary Table* of our view to the fact measure from one particular dimension angle.
- First column represents category or dimension
- Second column represents fact
- Consists of two types:
  - One Fact Measure
  - Multiple Fact Measure

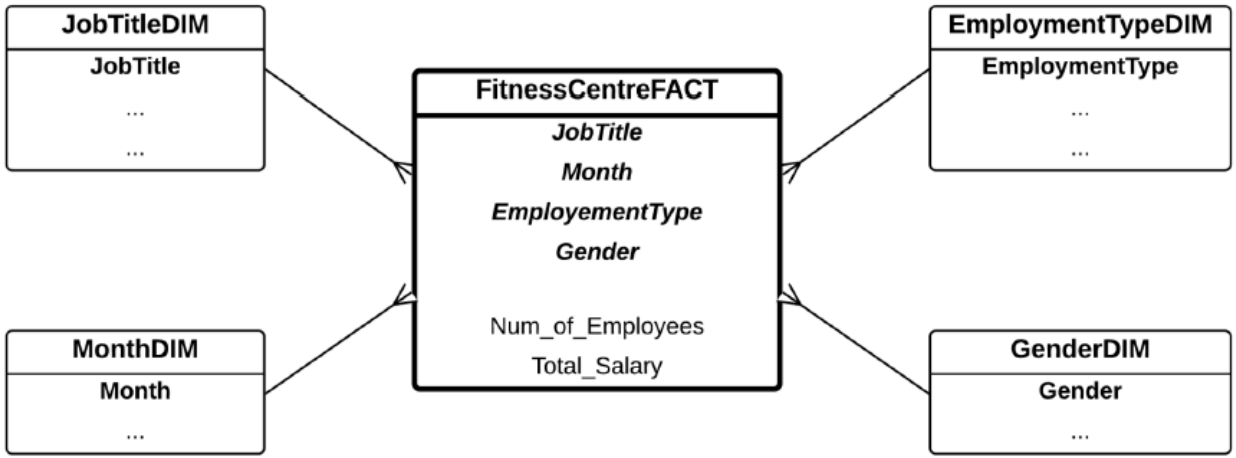
One fact measure

- First column contains a category
- Second column contains a statistical numerical figure



Multiple- fact measure

- Second column contains multiple facts F={F1,F2,F3…}
- All Fs must exist in all tables



A College Star Schema- Tutorial

E/R Diagram

- Used as operational system to support operational procedures

Example:

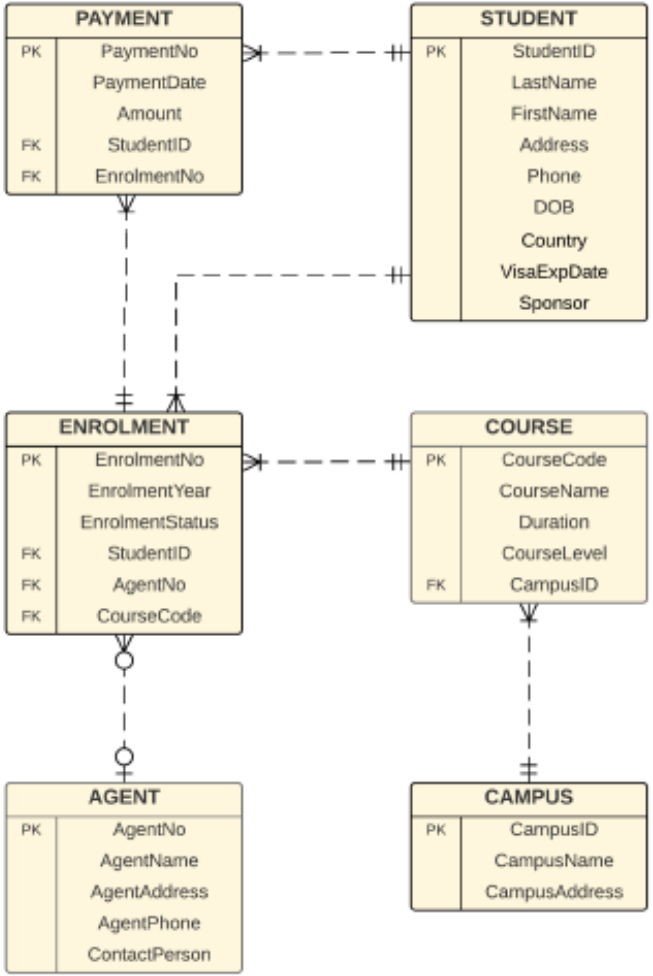
- As the College is a multi-campus university, some courses are offered in a different campus. The admission office handles international students of all campuses.

Star Schema

Used for analysis purposes.

Example:

- How many students come from certain countries?
- What is the total income for certain postgraduate courses?
- How many students are handled by certain agents?
- How the number of enrolment of courses fluctuates across the year?


PAYMENT
PK PaymentNo
PaymentDate
Amount
FK StudentID
FK EnrolmentNo
STUDENT
PK StudentID
LastName
FirstName
Address
Phone
DOB
Country
VisaExpDate
Sponsor
ENROLMENT
PK EnrolmentNo
EnrolmentYear
EnrolmentStatus
FK StudentID
FK AgentNo
FK CourseCode
COURSE
PK CourseCode
CourseName
Duration
CourseLevel
FK CampusID
AGENT
PK AgentNo
AgentName
AgentAddress
AgentPhone
ContactPerson
CAMPUS
PK CampusID
CampusName
CampusAddress

Case Study Summary

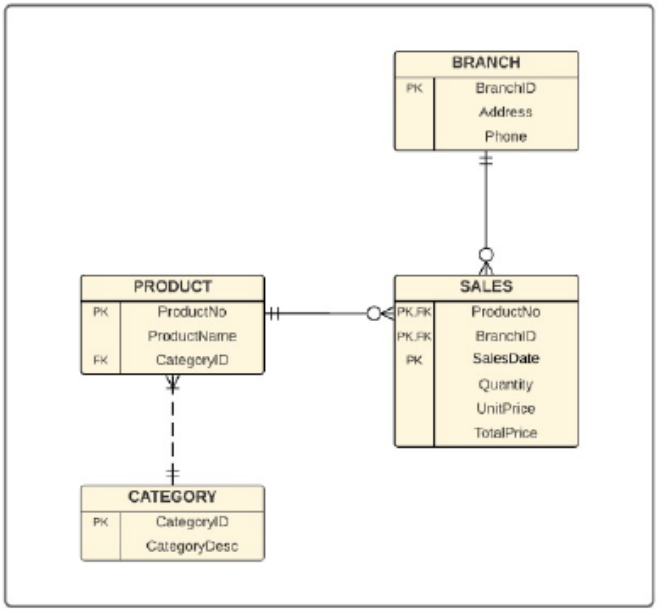
- Three ways to create dimension tables:
  - Use create table as select *
    which is direct copying from the table in the operational database
- Choose selected attributes from the table in the operational database
- Create the dimension table manually, followed by insert into to insert new records into the table

Summary on facts and dimensions

- Fact: Fact is numerical and aggregated value
- Dimensions: Point of view
- Creating Dimension Tables:
  - Direct Copy
  - Extracting some relevant attributes
  - Manually created
- Creating Fact Tables:
  - Direct retrieval from the tables in the operational database
- To validate, the two-column method can be used

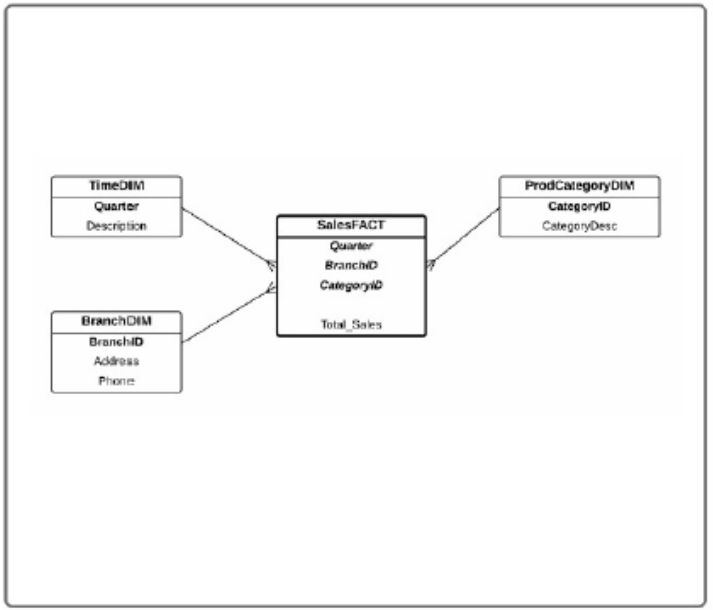
ETL Process

Operational Database (E/R Diagram)



Transformation (ETL)

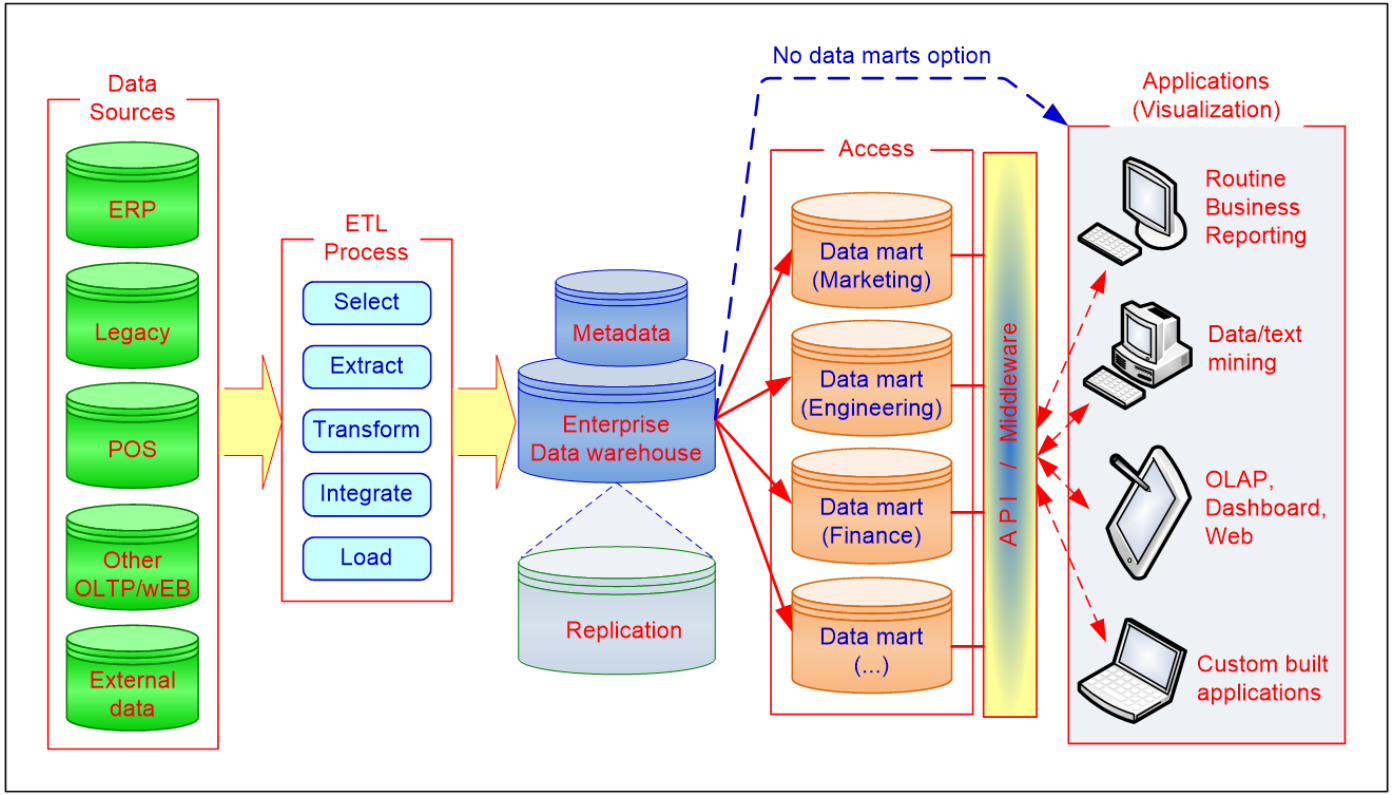
Data Warehouse (Star Schema)



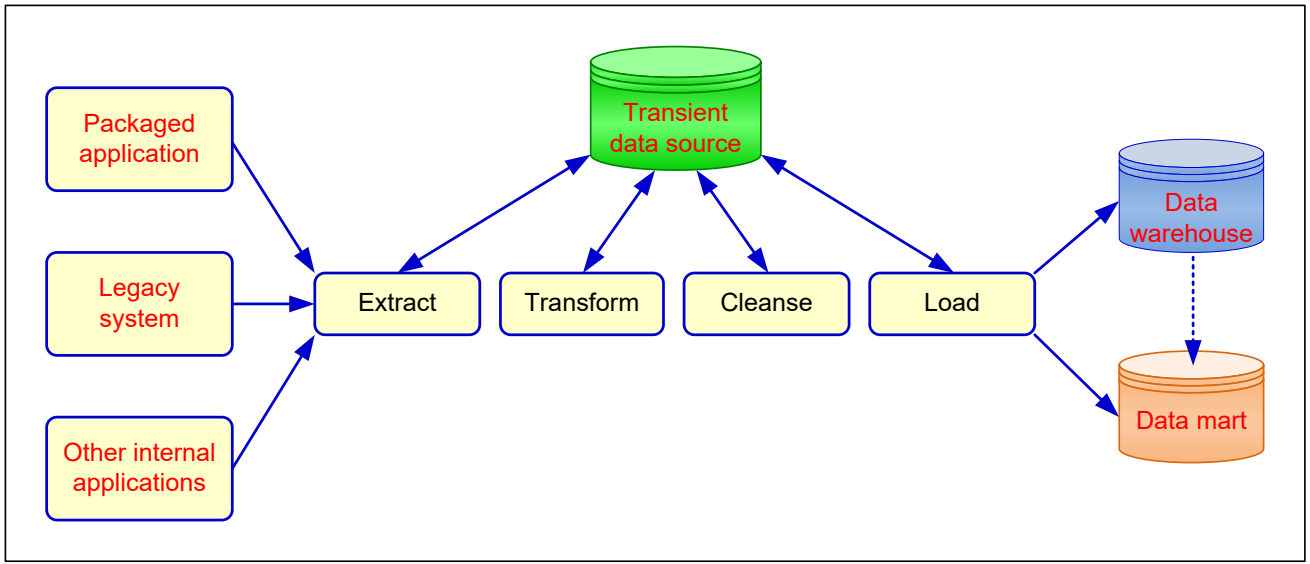
The ETL Process– Extraction, Transformation, Loading

- ETL stands for Extraction, Transformation, and Loading
  - Extraction: Collect the data from heterogenous data sources
  - Transformation: transform, clean, and standardize the data such that it can be integrated in the same data warehouse
  - Loading: consists of loading the data to the data warehouse
- Data staging area: the part of the data warehouse where transformations happens
- The staging area should not be available for querying

General Data Warehouse Architecture



Data Integration and the Extraction, Transformation, and Load Process



ETL (Extract, Transform, Load)

- Issues affecting the purchase of an ETL tool
  - Data transformation tools are expensive
  - Data transformation tools may have a long learning curve
- Important criteria in selecting an ETL tool
  - Ability to read from and write to an unlimited number of data sources/architectures
  - Automatic capturing and delivery of metadata
  - A history of conforming to open standards
  - An easy-to-use interface for the developer and the functional user

Staging Area

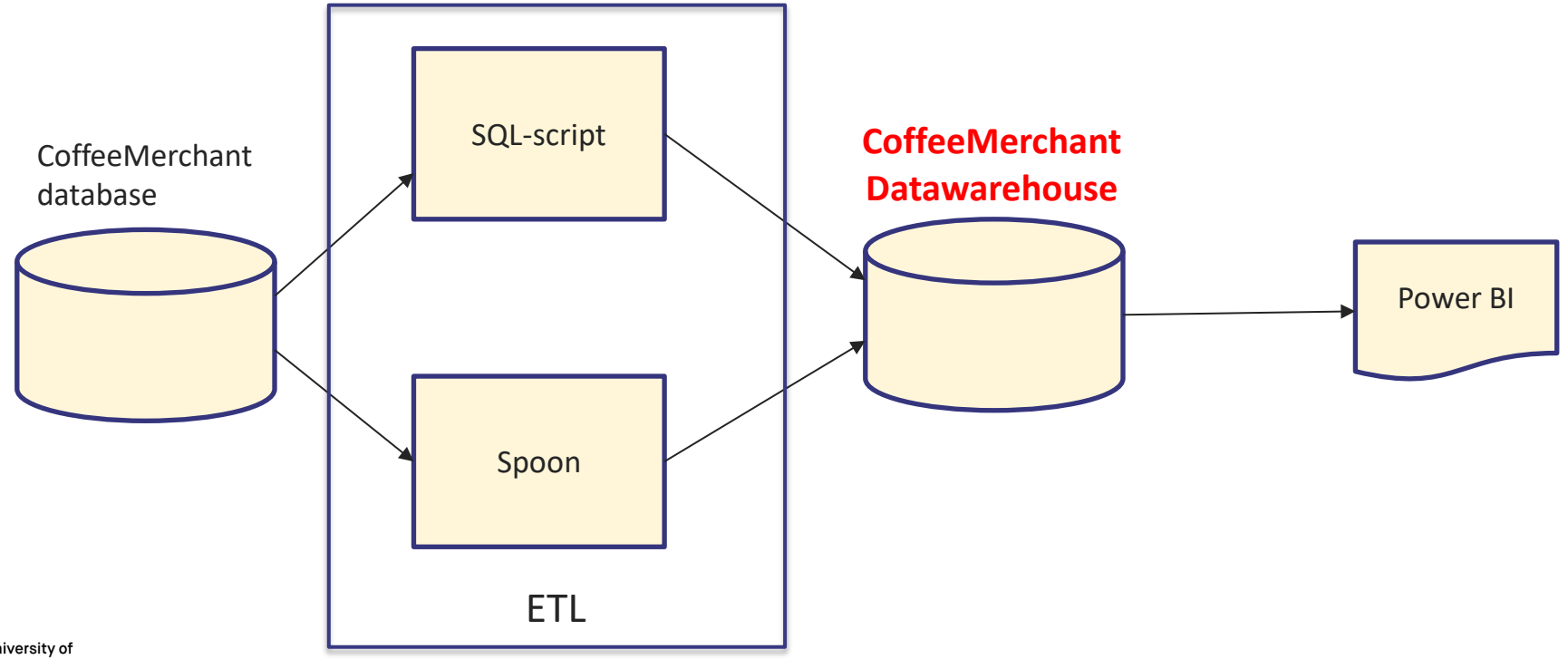
- Staging Area: Where extracted data is stored and possibly transformed before loading the data into the central warehouse.
  - Source system load times should be kept to an absolute minimum
  - Using a separate staging area enables you to work on a specific subset of the data
  - A dedicated schema allows for specific sorting or indexing to further optimise and support the ETL process
  - It’s a safety net: a process can fail before completing.

Best ETL Tools In 2024

- 1) Integrate.io
- 2) Skyvia
- 3) IRI Voracity
- 4) Dataddo
- 5) Dextrus
- 6) DBConvert Studio By SLOTIX s.r.o.
- 7) Informatica – PowerCenter
- 8) IBM–Infosphere Information Server
- 9) Oracle Data Integrator
- 10) Microsoft – SQL Server Integrated Services (SSIS)
- 11) Ab Initio
- 12) Talend – Talend Open Studio for Data Integration
- 13) CloverDX Data Integration Software
- 14) **Pentaho Data Integration**
- 15) Apache Nifi
- 16) SAS – Data Integration Studio
- 17) SAP–BusinessObjects Data Integrator
- 18) Oracle Warehouse Builder
- 19) Sybase ETL
- 20) DBSoftlab
- 21) Jasper

Tutorial on Spoon

Objective: Building coffeemerchant Data Warehouse Using Spoon PDI



Assignment

Creating the Data Warehouse Using Spoon