Extraction, Transformation & Load

Knowledge objectives

1. Enumerate six reasons to have an ETL process
2. Explain some legal and ethical concerns
3. Define ETL
4. Compare ETL, ELT and ETQ
5. Enumerate six mechanisms to extract data
6. Enumerate five kinds of transformation tasks
7. Enumerate three criteria to select the sources
8. Explain the two integration problems
9. Explain the three kinds of data cleaning activities
10. Justify the reduction of the data size
11. Explain two possibilities to reduce the size of data
12. Enumerate six kinds of preparation activities
13. Justify the existence of a staging area
14. Explain how to measure the performance of an ETL flow
15. Discuss the implementation alternatives of ETL flows

Motivation

- Use multiple sources together
- Provide measures of confidence in data
- Remove mistakes
- Correct missing data
- Transactional data safekeeping
- Restructure data to be used by other tools

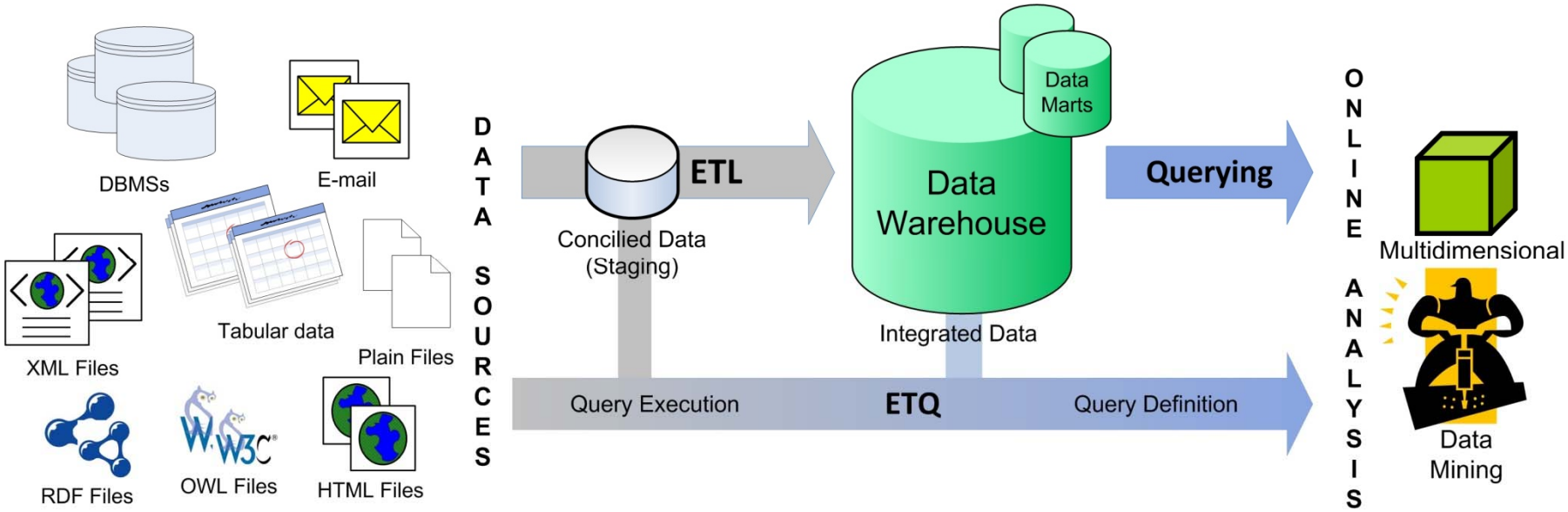
Ethical and legal aspects

- Who is the owner?
- Do we have permission …
- … To use them with our aim?
- Will we keep the confidentiality?
- How will the results be used?

Definition

- Extract
  - Multiple and heterogeneous sources
  - Different temporal characteristics of sources:
    - Transient
    - Semi-periodic
    - Temporal
- Transform
  - Change schema
  - Convert characters set
  - Cleaning
- Load
  - On-line/Off-line (i.e., update window)
  - Update/Rebuild indexes

ETL flows


DBMSs
E-mail
Tabular data
XML Files
Plain Files
RDF Files
OWL Files
HTML Files
DATA SOURCES
ETL
Concilied Data
(Staging)
Data
Warehouse
Data
Marts
Integrated Data
Querying
Query Execution
ETQ
Query Definition
ONLINE ANALYSIS
Multidimensional
Data
Mining

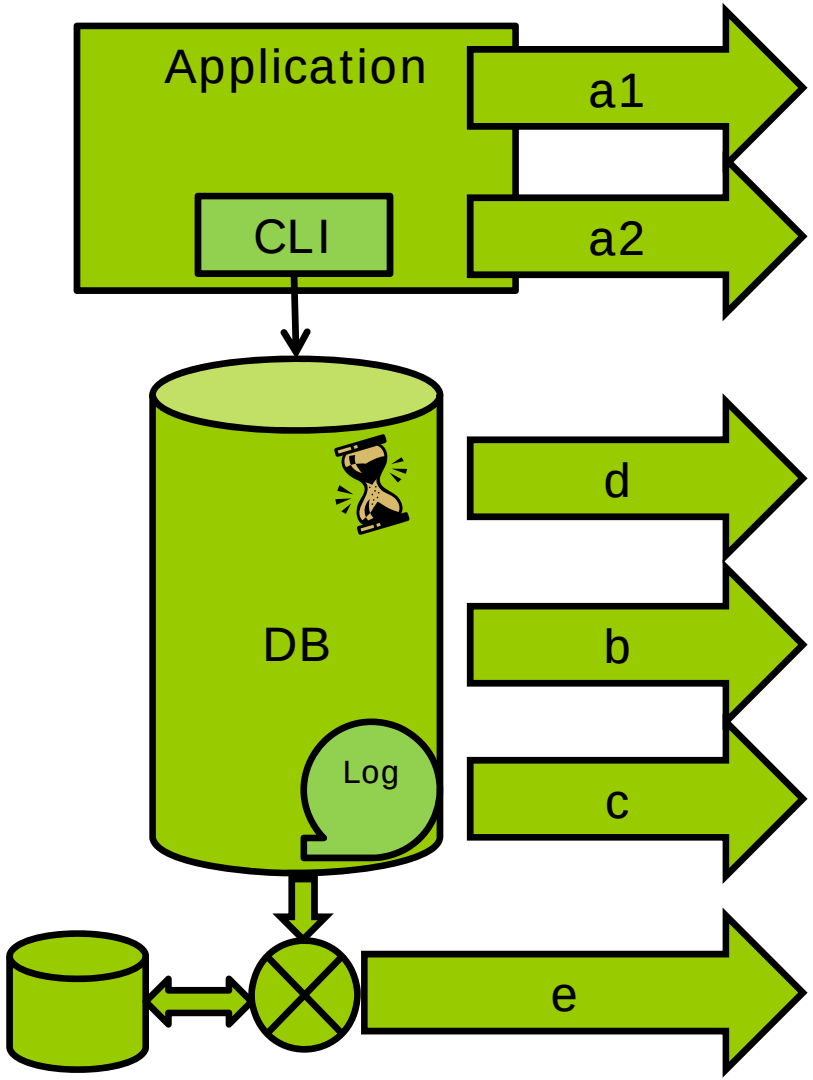
Source selection

"Garbage in, garbage out!"

- Relevant
- Non-redundant
- High quality
  - Complete
  - Accurate
  - Consistent
  - Timely

Data extraction mechanisms

a) Application-assisted
b) Trigger-based
c) Log-based
d) Timestamp-based
e) File comparison



Kinds of transforming tasks

"The process of standardizing data representation and eliminating errors in data."

- Selection
- Integration
- Cleaning
- Reduction
- Preparation

Data cleaning

- Generate data profiles
- Segment (split) columns
- Standardize
- Improve quality
  - Complete
    - Introduce values manually
    - Introduce a default value
    - Provide the average/median/mode
    - Provide the average/median/mode of the class
    - Provide the value contributing with more information
    - Cross multiple sources
  - Correct
    - Match dictionary
    - Detect outliers
      - Variance analysis
      - High distance to the regression function
      - High distance to any cluster
  - Check constraints and business rules

Size reduction

- Length (i.e., remove tuples)
  - Aggregate
  - Find a representative (i.e., clustering)
  - Sampling (keeping outliers)
- Width (i.e., remove attributes)
  - Correlation
  - Analysis of significance
  - Information gain (w.r.t. a classification)

Preparation

- Categorical to numerical
- Numerical to categorical (i.e., discretization)
  - Intervals of the same size
  - Intervals of the same provability
  - Clustering
  - Based on entropy
- Normalization
  - By the maximum (x/max)
  - By the interval size (|x-min|/|max-min|)
  - By the standard deviation $((x-\mu)/\sigma)$
  - Scaling $(x/10^j)$
- Conversion of data into metadata
- Derivation
- Enrichment (i.e., joins)

Integration

- Record matching (entity resolution)
  - Find an Object Identification Function
- Record merging
  - Schemas
  - Codification
  - Granularity
  - Units and scales
  - Data and metadata

Entity resolution

"Decide whether two tuples correspond to the same object in the real world."

- Problems to face:
  - Misspellings
  - Variant names
  - Misunderstanding of names
  - Evolution of values
  - Abbreviations
- Simplifications
  - Compare field by field
  - Normalize the strings

Architectural requirements

- Scheduler automation
- Exception handling
- Quality handling
- Recovery and restart
- Metadata management

Staging area

- Only for ETL purposes
- Structures
  - Plain files
  - XML
  - Relational tables
- Provides
  - Recoverability
  - Backup
  - Auditing

Measuring performance

- General Indicators
  - Rows read per second
  - Rows written per second
  - Throughput
- Infrastructure
  - CPU usage
  - Memory allocation
- Contention
  - Memory
  - Disk
  - Processor
  - Database

ETL tools

"The goal of a valuable tool is not to make trivial problems mundane, but to make impossible problems possible."

- Large projects
- Sophisticated processing
- Limited programming skills
- Integrated metadata repository
- Handle complex data type conversions
- Handle complex dependencies and error handling
- Automatic data lineage and dependency analysis
- Examples
  - Kettle-Pentaho Data Integration
  - cloverETL
  - JasperETL
  - Talend

Hand-coded ETL

- Not limited to vendor's abilities
- Reuse legacy routines
- Know-how already available

- MapReduce is specially appropriate
  - Schema-less data
  - "Read once" data sets
  - "Cooking" raw data ...
    - ... and loading them in a DBMS
  - Complex data flow

Summary

- Extraction, Transformation and Load
- Kinds of transformation tasks
  - Selection
  - Integration
  - Cleaning
  - Reduction
  - Preparation
- Architecture
- Tools

Bibliography

- M. Golfarelli and S. Rizzi. *Data Warehouse Design*. McGrau-Hill, 2009
- R. Kimball, J. Caserta. The Data Warehouse ETL Toolkit. Wiley Publishing, 2004
- J. Han and M. Kamber. "*Data Mining: Concepts and Techniques*". Morgan Kauffman Publishers, 2000
- M. Stonebraker et. al. "MapReduce and Parallel DBMSs: Friends or Foes?". Communications of the ACM, 53(1), 2010