POZNAN UNIVERSITY OF TECHNOLOGY

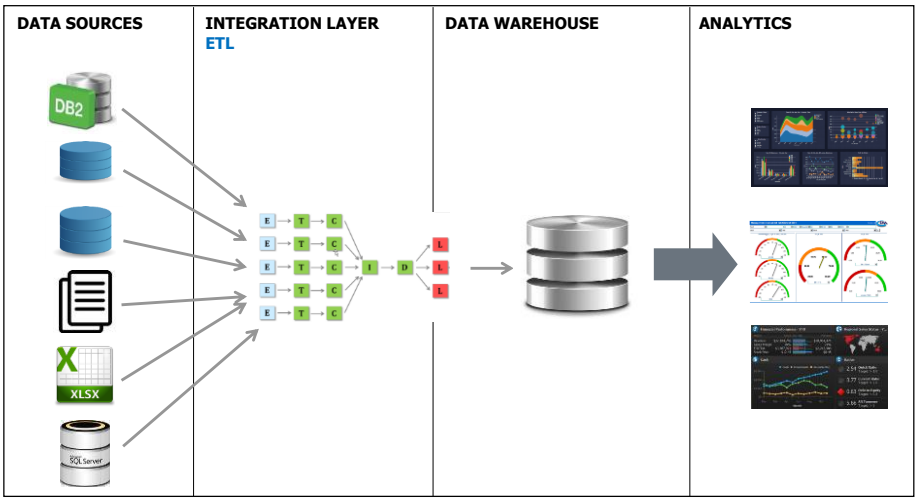
DW Loading and Refreshing Techniques: ETL
part 1

Robert Wrembel
Poznan University of Technology
Institute of Computing Science
Poznań, Poland
Robert.Wrembel@cs.put.poznan.pl
www.cs.put.poznan.pl/rwrembel

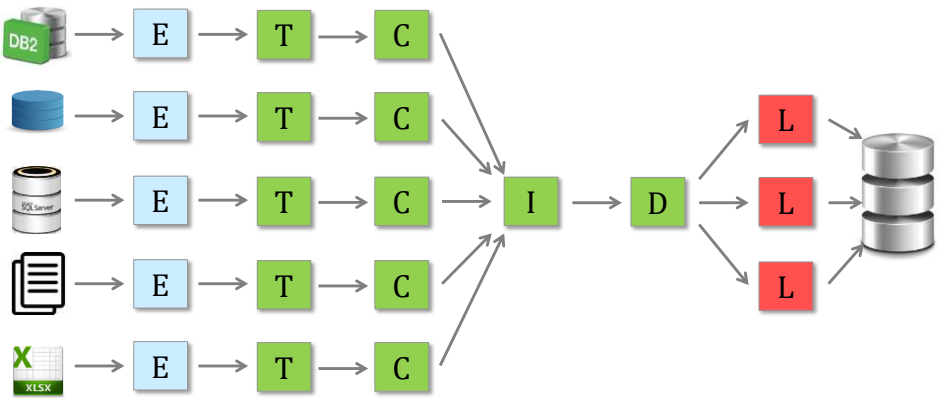
Outline

- On complexity of developing ETL layers
- In details
  - Extraction
  - Data profiling
  - Transformation
  - Cleaning
  - Integration

ETL in DWS architecture

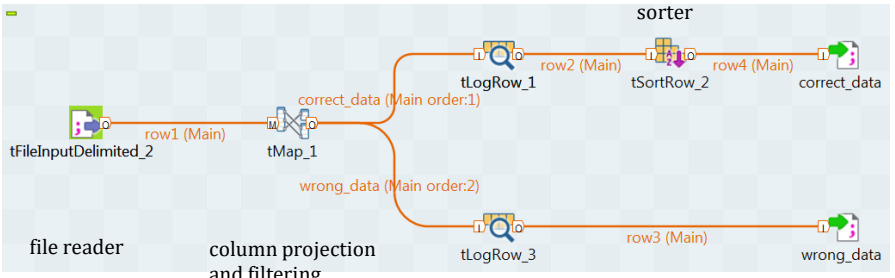


ETL in DWS architecture



Example ETL process

- Talend Open Studio



Developing ETL

- Designing and developing ETL processes
  - critical for DW functioning
  - challenges
    - data quality
    - data freshness
    - performance of ETL execution (time window for a DW refreshing)
    - source evolution
    - ETL optimization
  - costly
    - up to 70% project resources
      - staff
      - hardware
      - software

Developing ETL

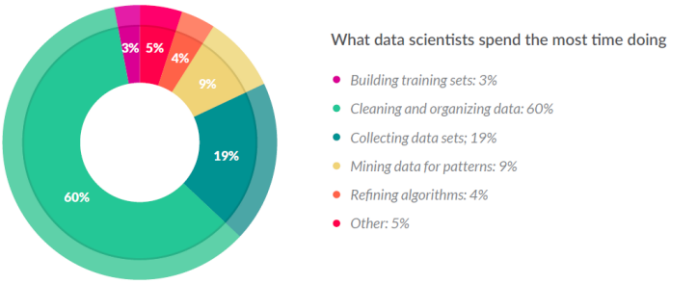
- Gartner Report on DW projects in financial institutions from the Fortune 500 list
  - 100 of staff in a DW project
  - 55 ETL
  - 17 system admins (DB, hardware)
  - 4 system architects
  - 9 BI consultants
  - 5 programmers
  - 9 managers
  - hardware (multiproc. servers, TB disks, 5mln USD)
  - ETL software (1mln USD)
  - # data sources: 10 to 50

Developing ETL

- # data sources to integrate
  - large banks: over 100
- Types of data sources to integrate
  - databases (all possible)
  - text files
  - spreadsheets
  - streaming data (more and more frequently)

Developing ETL

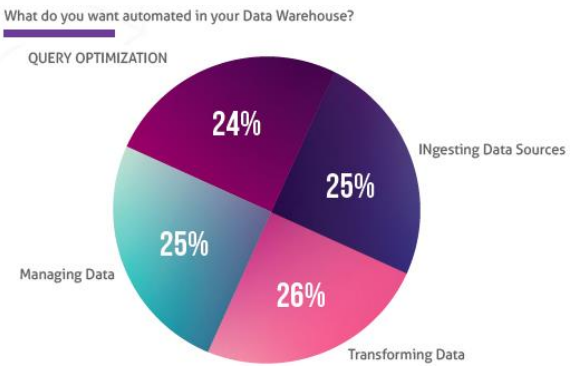
- **Data Science Report. 2016, CrowdFlower**



- **45% of time on data preparation tasks**
  - **Practical Data Preparation: Solutions to the Top 5 Most Common Mistakes. 2021, DataIku**
  - **2020 State of Data Science. 2020, Anaconda**
  - https://know.anaconda.com/rs/387-XNW-688/images/Anaconda-SODS-Report-2020-Final.pdf

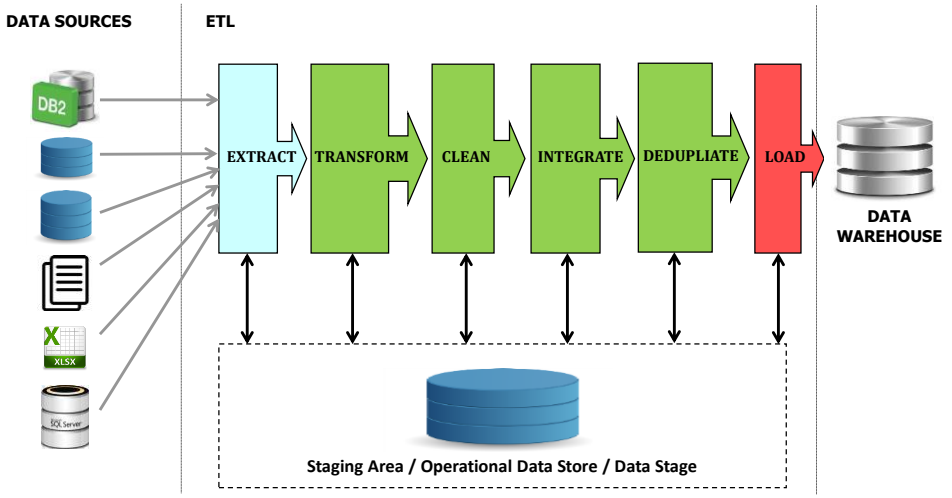
Developing ETL

- **Panoply Data Warehouse Trends Report 2018**



- **Data preparation and engineering tasks represent over 80% of the whole project (Data Engineering, Preparation, and Labeling for AI 2019. Cognilytica Research)**

ETL architecture

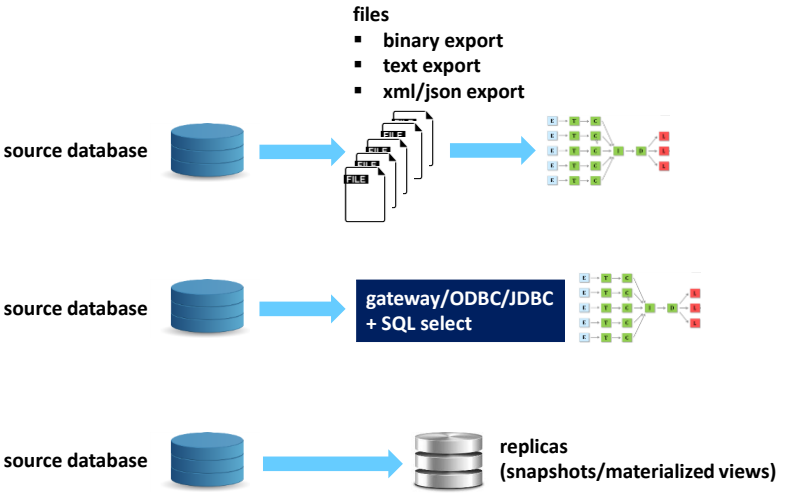


Extract

- Typical predefined connectors from an ETL tool to data sources

- IBM DB2
- SQL Server
- Oracle
- Sybase ASE, IQ
- Netezza
- Vertica
- Teradata
- SAS
- SAP Hana
- Greenplum
- PostgreSQL
- MySQL
- SQLite
- FireBird
- ODBC data source
- JDBC data source
- Excell
- Access
- Text, XML, JSON files
- Hive
- Impala
- MongoDB
- Cassandra
- ...

Ingesting data by ETL



Ingesting data by ETL

- How much to read from DS?
  - each time the whole content
  - deltas (increments) → how to detect changes?

Detecting data changes

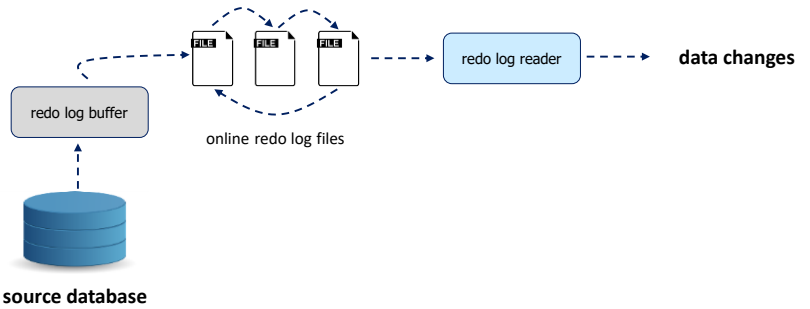
- Requirements
  - minimum or none source system changes
  - minimum interference with a data source
- Solutions
  - audit columns (implemented by a programmer)
  - log of changes on a table (implemented by a programmer)
  - snapshot comparison
  - triggers → synchronous data transfer
  - snapshots/materialized views
  - analysis of a redo log

Snapshot

- Snapshot = replica = materialized view
- Copy of a table or a subset of its columns and rows
- Refreshing
  - automatic with a defined interval
  - on demand
- SQL Server
- IBM DB2
- Oracle
- and others

Redo log

- Redo log = transaction log
  - periodical reading (log scraping)
  - continuous reading (log sniffing)



Designing ETL layer

- Analysis of available data sources
- Deciding on DS access technologies (see Topic 1)
- Data profiling
- Data ingest
  - full
  - incremental
- Transforming
- Cleaning and homogenizing
- Merging
- Duplicate elimination
- Aggregation (optional)
- Uploading into a DW

Data sources

- Identify relevant DSs
- DS description
  - business area (e.g., HR, payroll, sales, loans, marketing, ...)
  - business user
  - business owner
  - technical/infrastructure owner
  - hardware + OS
  - DBMS
  - schema
  - # transactions/day (workload)
  - data volume increase/day
  - total DB size

Data profiling

- Analyzing data sources
- Main categories of tasks
  - structure discovery (schema, relationships)
  - content analysis (data values, data quality)
  - relationships discovery between data sets
- Application areas of data profiling
  - ETL for DW
  - data conversion and migration
  - data quality analysis in production (transactional) databases
- Tools
  - statistics
  - data mining

Structure discovery

- Discovering schema
  - identify domains of attributes
  - identifying NULL / NOT NULL columns
  - UNIQUE attributes
  - PK candidates
  - FK candidates
  - functional dependencies
  - FK→PK integrity
- Discovering relationships between data sets (e.g., 1:1, 1:M, M:N)
  - typically for non-relational data
  - assessing costs of potential joins
- Discovering embedded value dependencies (in a denormalized schema)

Content analysis

- Basic column-level analysis
  - # or % of distinct values
    - discovering natural or artificial keys
  - # or % of NULL values
  - identifying valid allowed values for attributes
  - identifying default values
    - # or % of rows
  - identifying values other than expected / outliers
    - # or % or rows
  - identifying wrong values
    - # or % of rows
  - Min, Max, Avg of character string length
    - estimating target datatypes

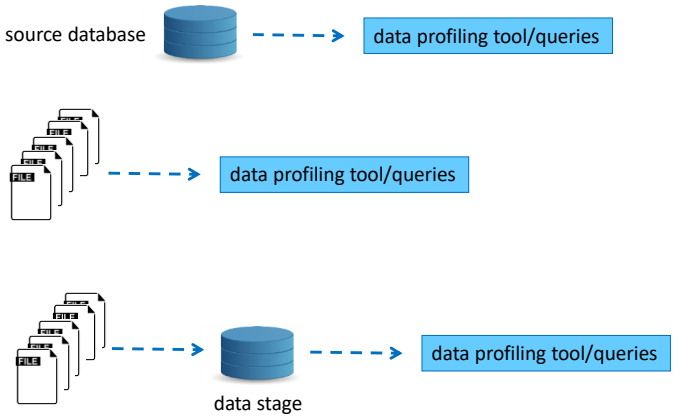
Content analysis

- Basic column-level analysis
  - Min and Max of numerical values
    - detecting outliers
    - estimating target datatypes, e.g., int or real, with or without decimal precision
  - Avg, Stdev, Var of numerical values
    - detecting outliers
  - Min and Max date values
  - Value formats
    - dates, phone no, addresses, money, email
  - computing data distributions (histograms)

Data profiling tools

- Open source
  - Quadient DataCleaner
  - Aggregate Profiler
  - Talend Open Studio for Data Quality
  - Melissa Data Profiler
  - ...
- Commercial
  - AbInitio
  - Precisely Trillium
  - IBM InfoSphere Information Analyzer (DataStage)
  - Informatica Data Profiling Solution – Data Explorer
  - SAP Business Objects Data Services for Data Profiling
  - Oracle Enterprise Data Quality
  - SAS DataFlux
  - ...

Data profiling architectures



Parsing

- **To understand and standardize values**
- **Build-in parsers (e.g., Trillium, Informatica, AbInitio)**
- **Parser features (examples)**
  - **personal names recognition**
    - **(Robert, Wrembel) = (Wrembel, Robert)**
  - **business names recognition**
  - **address recognition**

'Poznan Unviersity of Technology, Institute of Computing Science, Piotrowo 2, 60-965, Berdychowo, Poznań'

'Skłodowskiej-Curie 5, 60-965 Poznań, Wilda Poznan University of Technology'

'Laboratoire ERIC; Université Lumière Lyon 2 5 avenue Pierre Mendès, France, 69676 Bron Cedex'

parsing

organization	dept	city	district	street	no	zip
Poznan University of Technology	Institute of Computing Science	Poznań	Berdychowo	Piotrowo	2	60-965
Poznan University of Technology		Poznań	Wilda	Skłodowskiej-Curie	5	60-965
Université Lumière Lyon 2	Laboratoire ERIC	Lyon	Bron	Pierre Mendès	5	69676

Data transformation

- Transform to a common data model
  - relational
  - object-relational
  - semi-structured
  - NoSQL
  - graph
  - ...
- Transform semantically identical data to a common representation (structure)

prodID	name	category	group	net	gorss	pres
113	orange juice	drink	food	2	2,5	E103, E250, E321, E400, E502, E605, E944

barCode	pName	group_categ	netPrice	tax	E100	E200	E300	E400	E500	E600	E900
AABB01D	orange j.	food-drink	3	0.08	106	201	345	407	579	654	901

- Remove unnecessary columns

Data cleaning

- Missing (incomplete) data
  - remove/replace null values
- Invalid data
  - correct typos
    - dictionaries (spelling, names, cities, countries)
- Inconsistent data
  - zip inconsistent with city
  - multiple abbreviations for the same text value
  - gross = net + vat
- Standardize values
  - date format
  - currency
  - capital/small letters
  - abbreviations
  - synonyms (Word Net)

Data transformation & cleaning

- Requirements
  - iterative and interactive process
    - define transformation
    - run process
    - verify results

  - extendible and easy to modify
  - as much data as possible should be transformed automatically
  - as much steps as possible should be automatic

Data cleaning: problems

- Place of birth
  - how to verify if a name is correct? no zip code given as a reference point
- Cleaning addresses
  - Polish Mail provides dictionaries of cities and zip codes
  - TERYT provides dictionaries of cities and streets
- Cleaning city names
  - zip code may be useful
  - finding the most similar name in a cities dictionary
  - CZARNKOW vs. CZARNKÓW vs. CZARNKOWO
  - similarity measures: Jaro-Winkler, Sorensen, StrCmp95, Overlap
  - avg(sim(CZARNKOW, CZARNKÓW))=0.9125
  - avg(sim(CZARNKOW, CZARNKOWO))=0.9742

Data cleaning: problems

- Cleaning first names
  - finding the most similar name in a names dictionary
  - avg(sim(Alaksandra, Aliaksandra))=0.9760
  - sim(Alaksandra, Aleksandra)=0.8829
  - sim(Aliaksandra, Aleksandra)=0.8611

Data cleaning: problems

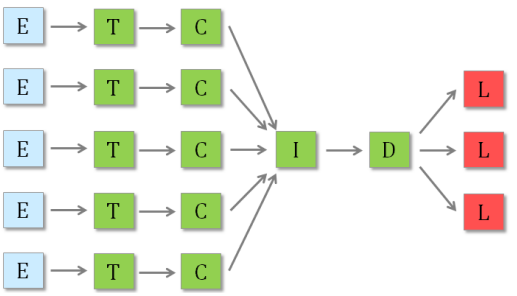
- Street names change in time
- International addresses
  - street and city names
- Fathers name interchanged with mothers name
  - problem in correcting foreigners
- First name interchanged with last name
  - in rare cases valid last names can take a value of a first name
- TERYT
  - M. Skłodowskiej-Curie in Poznań
  - M Skłodowskiej-Curie in Warsaw

Data cleaning: problems

- Other examples
  - ŁAŻNIKI - ŁAŹNIKI
  - WIERZBOWICE - WIERZBOCICE
  - KRETA - KRĘTA
  - DOLNA - DOLINA
  - PRZEMYSŁAWA - PRZEMYSŁOWA
  - STUDZIENICE - STUDZIENIEC
  - KAROLEWO - KORALEWO
  - WOŁOWA - WAŁOWA
  - GARBACKA - GARBATKA
  - WŁOCINA - SŁOCINA
  - SOKOŁOWICE - SOKOŁOWIEC

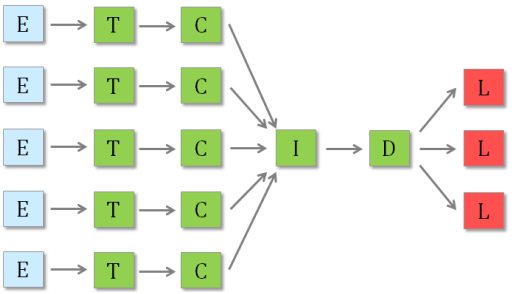
Integration

- Parallel flows are merged into a single flow for deduplication



Deduplication

➲ **Removing duplicate data from an integrated dataset**



Gartner: ETL tools

- **Open-source**
  - **Talend Open Studio**
  - **Pentaho Data Integration**
  - **CloverETL**
  - **Apache NiFi**
- **Commercial**
  - **IBM Data Stage**
  - **Informatica**
  - **Microsoft Integration Services**
  - **AbInitio**

Source: Gartner (March 2019)