POZNAN UNIVERSITY OF TECHNOLOGY

DW Loading and Refreshing Techniques: ETL

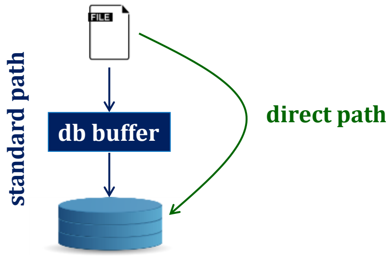
Robert Wrembel
Poznan University of Technology
Institute of Computing Science
Poznań, Poland
Robert.Wrembel@cs.put.poznan.pl
www.cs.put.poznan.pl/rwrembel

Outline

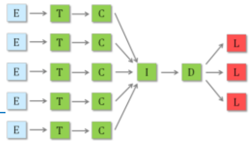
- ➲ **Loading data into DW**
- ➲ **Metadata**

Loading

- Parallel loading
- Direct path loading vs. standard path loading



- Collecting DW statistics after refreshing
- DW defragmentation



DW refreshing

- When?
  - synchronous (after a source transaction was committed) ⇨ (near) real-time DW
  - asynchronous ⇨ traditional DW
    - with a defined frequency
    - on demand
- How?
  - full (1st DW load)
  - incremental (all next loads)
- How data arrive?
  - batch ⇨ traditional DW
  - stream ⇨ (near) real-time DW

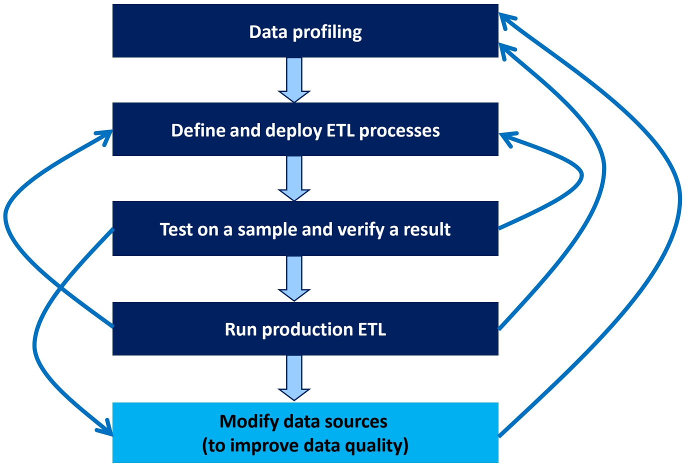
Ingesting data: tips

- Do not execute these operations in a data source
  - sorting
    - DISTINCT
    - set operators
    - GROUP BY
  - NOT and non-equijoins (typically require full scan)
  - functions in the WHERE clause

Ingesting data: tips

- Where to filter data?
  - at a data source (push down optimization), if
    - not overloaded with its proper processing
    - powerful query optimizer
    - low selectivity + good use of indexes
  - in an ETL layer, otherwise
    - sorting in a database
    - sorting in an OS (awk)
- Separate inserts from updates
  - updates → standard path
  - inserts → direct load path
- Decide how to maintain additional data structures
  - indexes
  - materialized views
- Integrity constraints in a DW?

Summary: ETL design process


Data profiling
Define and deploy ETL processes
Test on a sample and verify a result
Run production ETL
Modify data sources
(to improve data quality)

Jarke M., et. al.: Improving OLTP Data Quality Using Data Warehouse Mechanisms. SIGMOD Record, (28):2, 1999

Metadata

- On data sources
- On ETL processes
- On data warehouse

- **On data sources**
  - location (IP address)
  - hardware + operating system
  - type (RBD, OBD, XML, spreadsheet, ...)
  - schema
  - access methods (SQL, XQuery, dump file, ...)
  - connection credentials
  - results of data profiling
  - volume
  - performance characteristics

Metadata

- On ETL
  - data storage architecture of ODS and DW (e.g., disk capacities, row-store / column-store)
  - metadata on a dataset to be uploaded into DW (e.g., size, avg. record lengths)
  - definitions of ETL tasks/steps
  - available dictionaries (e.g., cities, zip codes, names)
  - workflow execution schedules
  - execution statistics (e.g., elapsed time, CPU time, #I/O, RAM usage, throughput, disc access conflicts, #records uploaded, #records rejected)
  - dependencies between workflows
  - dependencies between tasks for impact analysis
  - mappings between DS and DW structures
  - data lineage
  - execution logs

Requirements for ETL

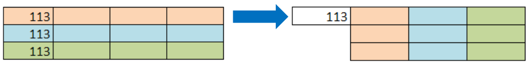
- Efficiency
  - finishing in a predefined time window
  - estimating execution termination
- Optimizable
- Fault-tolerance
  - restart after removing errors from a break point
  - restart from the beginning
  - recovery after crash
- Manageability
  - scheduling executions
    - time-based
    - token-based
  - stopping and restarting tasks
  - impact analysis
  - easy modifiable workflows

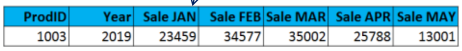
Requirements for ETL

- ➲ Producing data of high quality
- ➲ Security: access control
- ➲ Automatic code generation
- ➲ Support for user defined functions
- ➲ Automatic reporting on termination, errors, exceptions, and progress
- ➲ On line monitoring of work
- ➲ Parallel processing
- ➲ Direct path loading

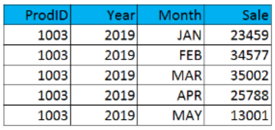
Requirements for ETL

- ➲ GUI for designing and managing processes
- ➲ A palette of predefined tasks
- ➲ Typical predefined tasks (ordered by usage frequency)
  1. Filter
  2. Aggregator
  3. Lookup
  4. Join
  5. Sort
  6. Combine record → combines records whose keys are identical into vectors of sub-records
  7. Modify → alters record structure
  8. Pivot





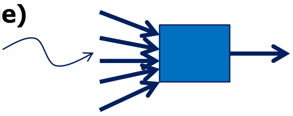
ProdID	Year	Sale JAN	Sale FEB	Sale MAR	Sale APR	Sale MAY
1003	2019	23459	34577	35002	25788	13001



ProdID	Year	Month	Sale
1003	2019	JAN	23459
1003	2019	FEB	34577
1003	2019	MAR	35002
1003	2019	APR	25788
1003	2019	MAY	13001

Requirements for ETL

- Typical predefined tasks (cont.)
  9. Merge → merging records (like SQL merge)
  10. Funnel → merging n input flows into one
  11. Transformer → transformation of data
  12. Remove duplicates
  13. Tail (usually in combination with sort)
  14. Head (usually in combination with sort)
  15. Compare → column-by-column comparison of records in two presorted input data sets
  16. Switch → dataflow split on a condition on column(s)
  17. Checksum → generates a checksum for a record
  18. Compress → dataset compression
  19. Expand → decompression



Off-the-shelf vs. in-house

- Off-the-shelf
  - faster design and deployment
  - integrated data repository
  - metadata management
  - workflow execution scheduling
  - built-in drivers to multiple DSs
  - impact analysis
  - incremental data loading
  - parallel processing
  - price
  - often require more advanced architectures → cost

- In-house-developed
  - longer design and development
  - thorough testing
  - dedicated to a given scenario
  - not customizable
  - may be tuned to a given scenario
  - may be less expensive