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 three levels of data profiling
6. Enumerate six mechanisms to extract data
7. Enumerate five kinds of transformation tasks
8. Enumerate three criteria to select the sources
9. Explain the two integration problems
10. Explain the three kinds of data cleaning activities
11. Justify the reduction of the data size
12. Explain two possibilities to reduce the size of data
13. Enumerate six kinds of preparation activities
14. Discuss the three abstraction levels of ETL process design
15. Discuss the implementation alternatives of ETL flows
16. Enumerate four criteria of ETL process quality
17. Discuss the differences between Data and Control flow in ETL
18. Justify the existence of a staging area
19. Discuss five kinds of ETL metadata
20. Enumerate the different ETL operation
21. Distinguish between blocking and non-blocking operations

Application Objectives

1. Given 2-3 sources, a target and some requirements create the corresponding ETL flow

MOTIVATION AND DEFINITION

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

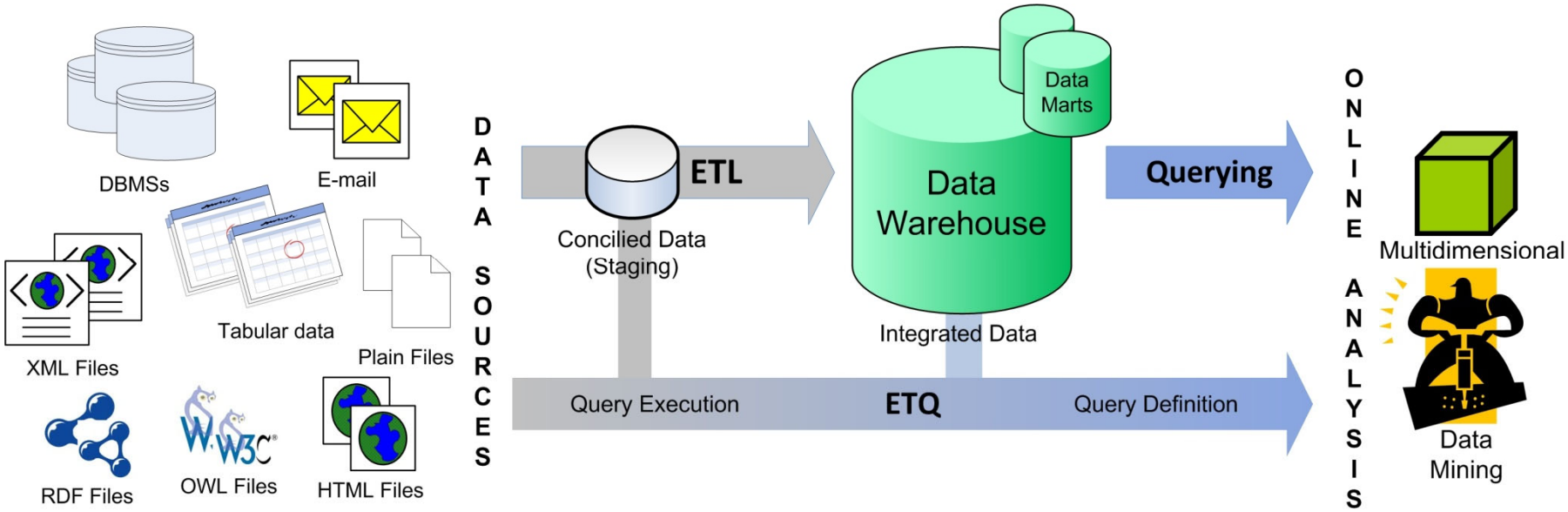
Legal and ethical aspects

- 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

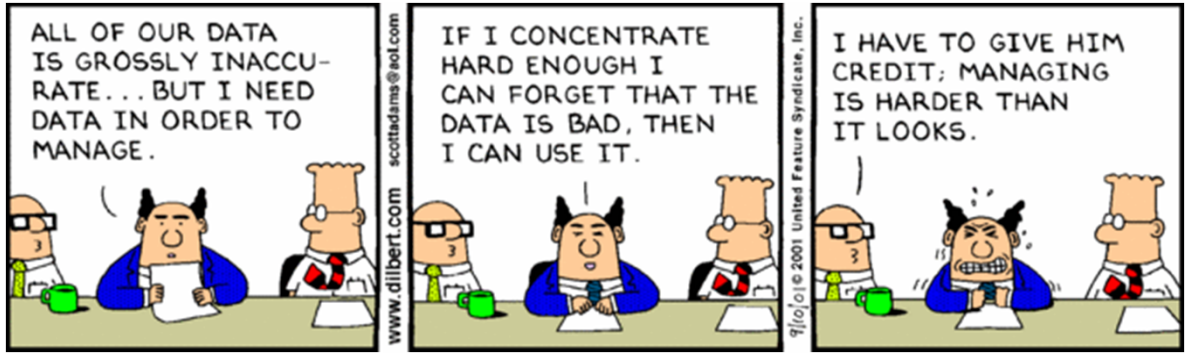
ETL flows


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

ETL STEPS

Source selection

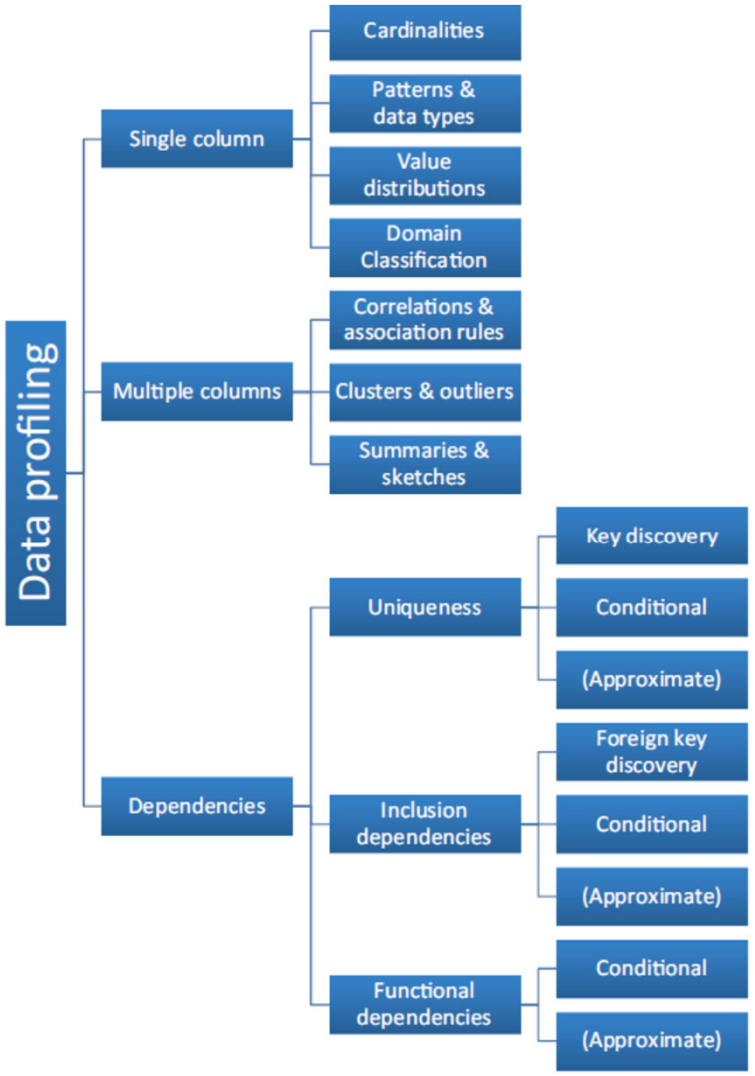
- Starting from the high-level business objectives and target MD schema
  - Relevant
  - Non-redundant
  - High quality
    - Complete
    - Accurate
    - Consistent
    - Timely
- Data profiling



https://www.wearetheliving.com/myth-garbage-garbage/

Source selection: data profiling

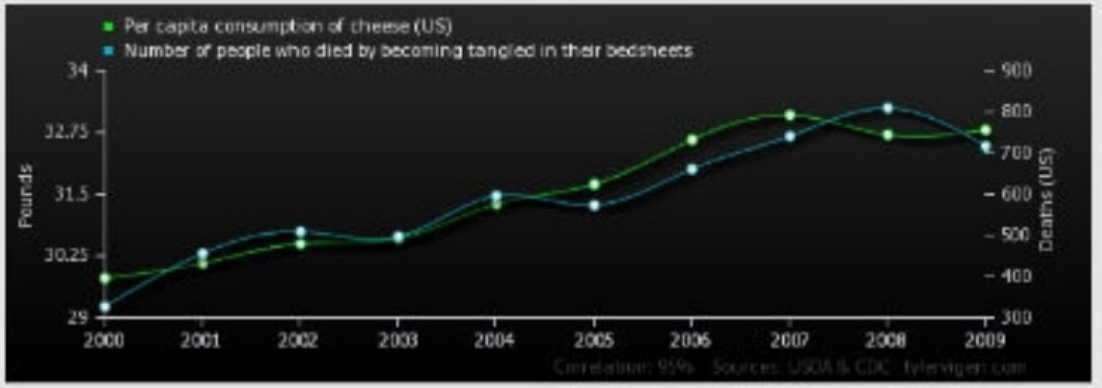
- Selected source data profiled for
  - Data quality
  - Data characteristics
  - Fitness for the purpose



Z. Abedjan et al. VLDB 2015

Source selection: data relevance example

Per capita consumption of cheese (US)
correlates with
Number of people who died by becoming tangled in their bedsheets



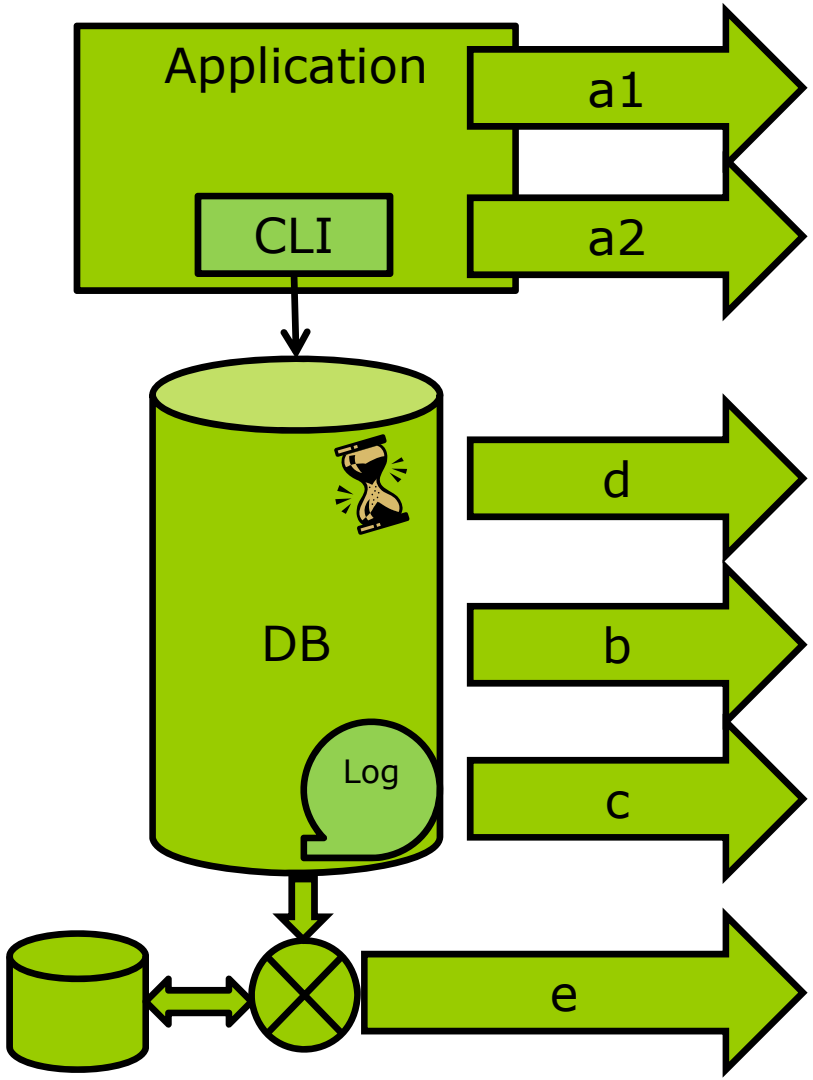
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) Pounds (USDA)	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in their bedsheets Deaths (US) (CDC)	327	456	509	497	596	573	661	741	809	717

Correlation: 0.947091

https://dangerousminds.net/comments/spurious_correlations_between_nicolas_cage_movies_and_swimming_pool

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
- Cleaning
- Reduction
- Preparation
- Integration

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 (use a lookup table)
  - Correct
    - Match dictionary (or lookup table)
    - 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)

ETL PROCESS DESIGN

Design abstraction levels

- Conceptual design
  - Understanding, sharing, and documentation
  - Existing formalizations
    - BPMN, UML, Ontology/Semantic web technologies
  - Domain-specific languages
- Logical design
  - Detailed description of the workflow, dependencies, schedule and recovery plans
  - Directed Acyclic Graph (DAG)
  - Automated translation from conceptual model
  - Drag&drop manual logical design
- Physical design
  - Automatic code generation
  - Template-based
  - ETL tools (GUI)
  - Hand-coded ETL

ETL tools (GUI)

"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

Data flow vs. control flow

Staging area

Metadata management

ARCHITECTURAL SETTING

Data flow vs. control flow

- Data Flow
  - Transformation over data
  - Pipeline execution
  - No strict order of execution
- Control flow
  - Orchestrating the execution of data flows
  - Does not work with data directly
  - Execution in strictly defined order (sequential or parallel)
  - Important for dependent data processing
    - Example: dimension tables -> fact tables

Staging area

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

Metadata management

- Source system metadata
- Data-Staging metadata
- Target DW metadata
- Business rules
  - Business metadata
  - Technical metadata
- Process execution metadata

ETL OPERATIONS

ETL operations (I)

Operation Level	Operation Type	Pentaho PDI	Talend Data Integration	SSIS	Oracle Warehouse Builder
Field	Field Value Alteration	Add constant Formula Number ranges Add sequence Calculator Add a checksum	tMap tConvertType tReplaceList	Character Map Derived Column Copy Column Data Conversion	Constant Operator Expression Operator Data Generator Transformation Mapping Sequence
Dataset	Duplicate Removal	Unique Rows Unique Rows (HashSet)	tUniqRow	Fuzzy Grouping	Deduplicator
	Sort	Sort Rows	tSortRow	Sort	Sorter
	Sampling	Reservoir Sampling Sample Rows	tSampleRow	Percentage Sampling Row Sampling	
	Aggregation	Group by Memory Group by	tAggregateRow tAggregateSortedRow	Aggregate	Aggregator
	Dataset Copy		tReplicate	Multicast	
Row	Duplicate Row	Clone Row	tRowGenerator		
	Filter	Filter Rows Data Validator	tFilterRow tMap tSchemaComplianceCheck	Conditional Split	Filter
	Join	Merge Join Stream Lookup Database lookup Merge Rows Multiway Merge Join Fuzzy Match	tJoin tFuzzyMatch	Merge Join Fuzzy Lookup	Joiner Key Lookup Operator
	Router	Switch/Case	tMap	Conditional Split	Splitter
	Set Operation - Intersect	Merge Rows (diff)	tMap	Merge Join	Set Operation
	Set Operation - Difference	Merge Rows (diff)	tMap		Set Operation
	Set Operation - Union	Sorted MergeAppend streams	tUnite	Merge Union All	Set Operation

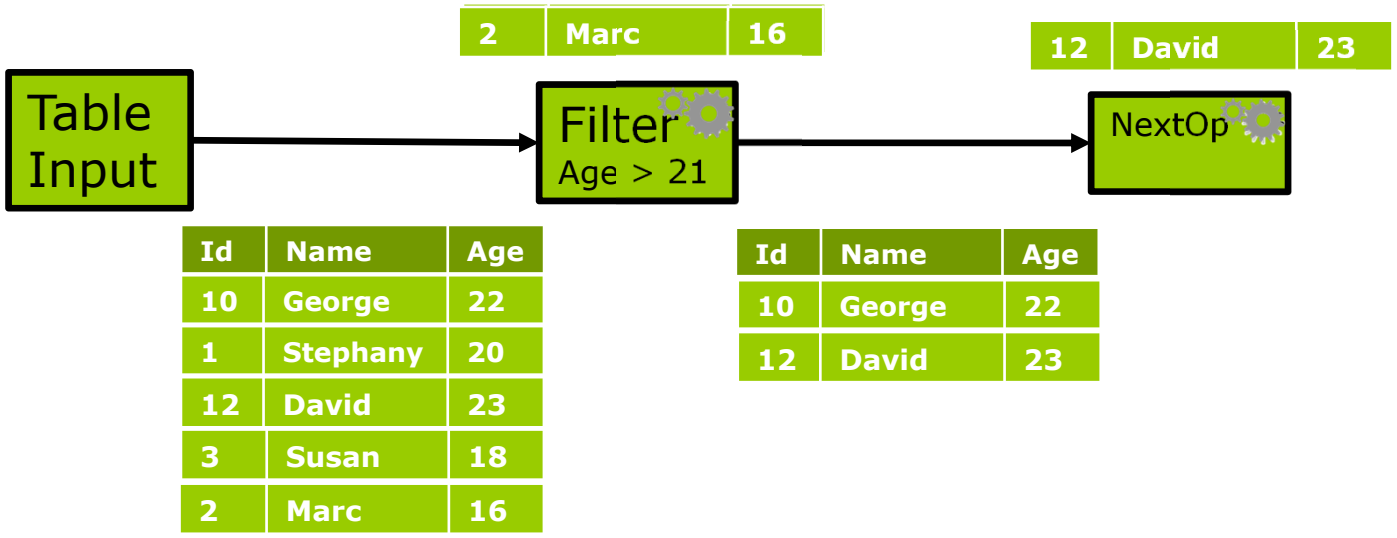
Vasileios Theodorou

ETL operations (II)

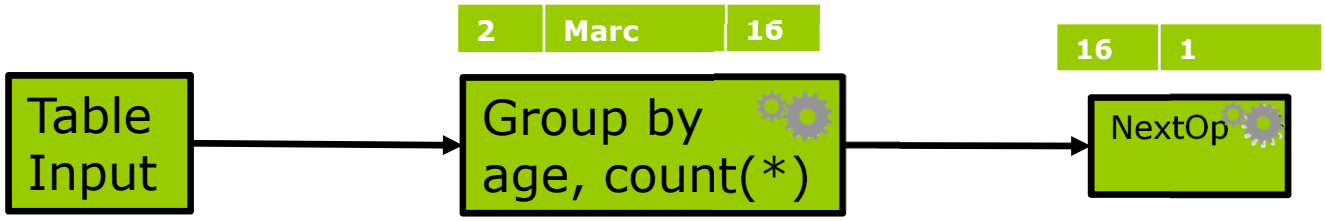
Operation Level	Operation Type	Pentaho PDI	Talend Data Integration	SSIS	Oracle Warehouse Builder
Schema	Field Addition	Set field value Set field value to a constant String operations Strings cut Replace in string Formula Split Fields Concat Fields Add value fields changing sequence Sample rows	tMap tExtractRegexFields tAddCRCRow	Derived Column Character Map Row Count Audit Transformation	Constant Operator Expression Operator Data Generator Mapping Input/Output parameter
	Datatype Conversion	Select Values	tConvertType	Data Conversion	Anydata Cast Operator
	Field Renaming	Select Values	tMap	Derived Column	
	Projection	Select Values	tFilterColumns		
Table	Pivoting	Row Denormalizer	tDenormalize tDenormalizeSortedRow	Pivot	Unpivot
	Unpivoting	Row Normalizer Split field to rows	tNormalize tSplitRow	Unpivot	Pivot
Value	Single Value Alteration	If field value is null Null if Modified Java Script Value SQL Execute	tMap tReplace	Derived Column	Constant Operator Expression Operator Match-Merge Operator Mapping Input/Output parameter
Source Operation	Extraction	CSV file input Microsoft Excel Input Table input Text file input XML Input	tFileInputDelimited tDBInput tFileInputExcel	ADO .NET / DataReader Source Excel Source Flat File Source OLE DB Source XML Source	Table Operator Flat File Operator Dimension Operator Cube Operator
Target Operation	Loading	Text file output Microsoft Excel Output Table output Text file output XML Output	tFileOutpu tDelimited tDBOutput tFileOutputExcel	Dimension Processing Excel Destination Flat File Destination OLE DB Destination SQL Server Destination	Table Operator Flat File Operator Dimension Operator Cube Operator

Vasileios Theodorou

Non-blocking operation (example)



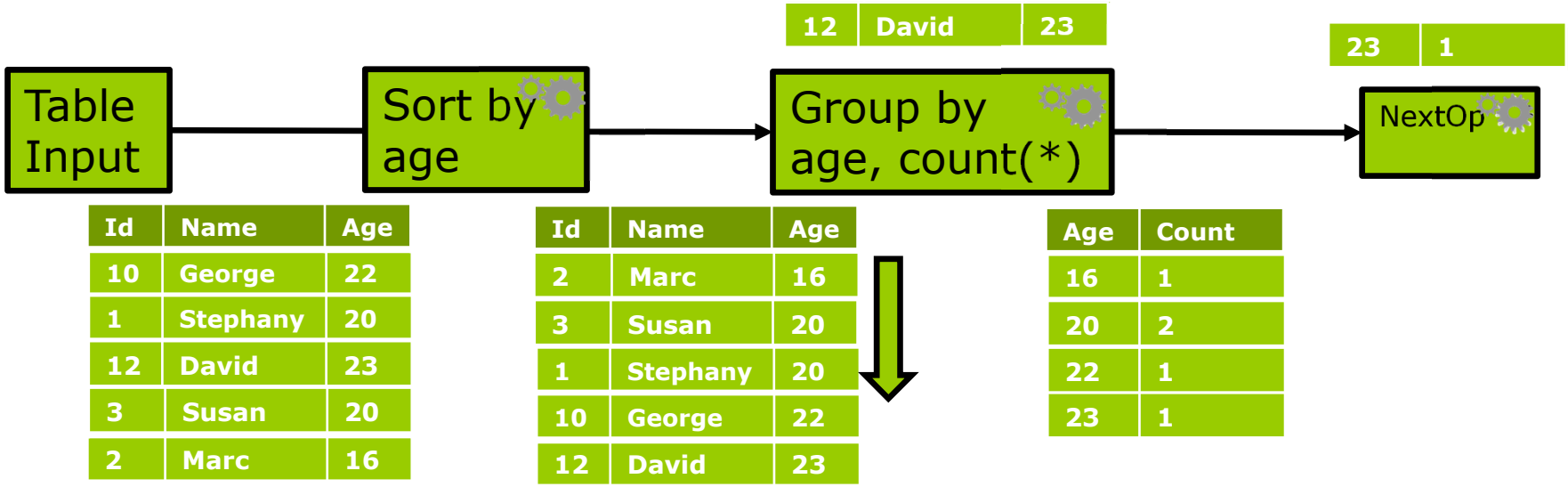
Blocking operation (example)



Id	Name	Age
10	George	22
1	Stephany	20
12	David	23
3	Susan	20
2	Marc	16

Age	Count
16	1
20	2
22	1
23	1

Blocking operation - optimized (example)



Id	Name	Age
10	George	22
1	Stephany	20
12	David	23
3	Susan	20
2	Marc	16

Id	Name	Age
2	Marc	16
3	Susan	20
1	Stephany	20
10	George	22
12	David	23

Age	Count
16	1
20	2
22	1
23	1

Kinds of operations

- Blocking
  - Duplicate Removal
  - Sort
  - Aggregation
  - Join
  - Intersect
  - Difference
  - Unpivot
- Non-blocking
  - Field Value Alteration
  - Single Value Alteration
  - Sampling
  - Dataset Copy
  - Duplicate Row
  - Router
  - Union
  - Field Addition
  - Datatype Conversion
  - Field Renaming
  - Projection
  - Pivot
  - Extraction
  - Loading

CLOSING

Summary

- Extraction, Transformation and Load
- Kinds of transformation tasks
  - Selection
  - Integration
  - Cleaning
  - Reduction
  - Preparation
- ETL process design
- ETL process quality
- Architectural setting
  - Data flow vs. Control flow
  - Staging area
  - Metadata management
- Operations
  - Blocking vs. non-blocking

Bibliography

- M. Golfarelli and S. Rizzi. *Data Warehouse Design*. McGrau-Hill, 2009
- R. Kimball, J. Caserta. The Data Warehouse ETL Toolkit. Wiley Publishing, 2004
- A. Vaisman and E. Zimányi. Data Warehouse Systems - Design and Implementation. Data-Centric Systems and Applications, Springer, 2014
- 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
- Z. Abedjan, L. Golab, and F. Naumann. Profiling relational data: a survey. *The VLDB Journal* 24, no. 4 (2015): 557-581.