The History, Present, and Future of ETL Technology

[DOLAP 2023 Test-of-Time Award – Invited Talk]

Alkis Simitsis
Athena Research Center
alkis@athenarc.gr

Spiros Skiadopoulos
University of the Peloponnese
spiros@uop.gr

Panos Vassiliadis
University of Ioannina
pvassil@cs.uoi.gr

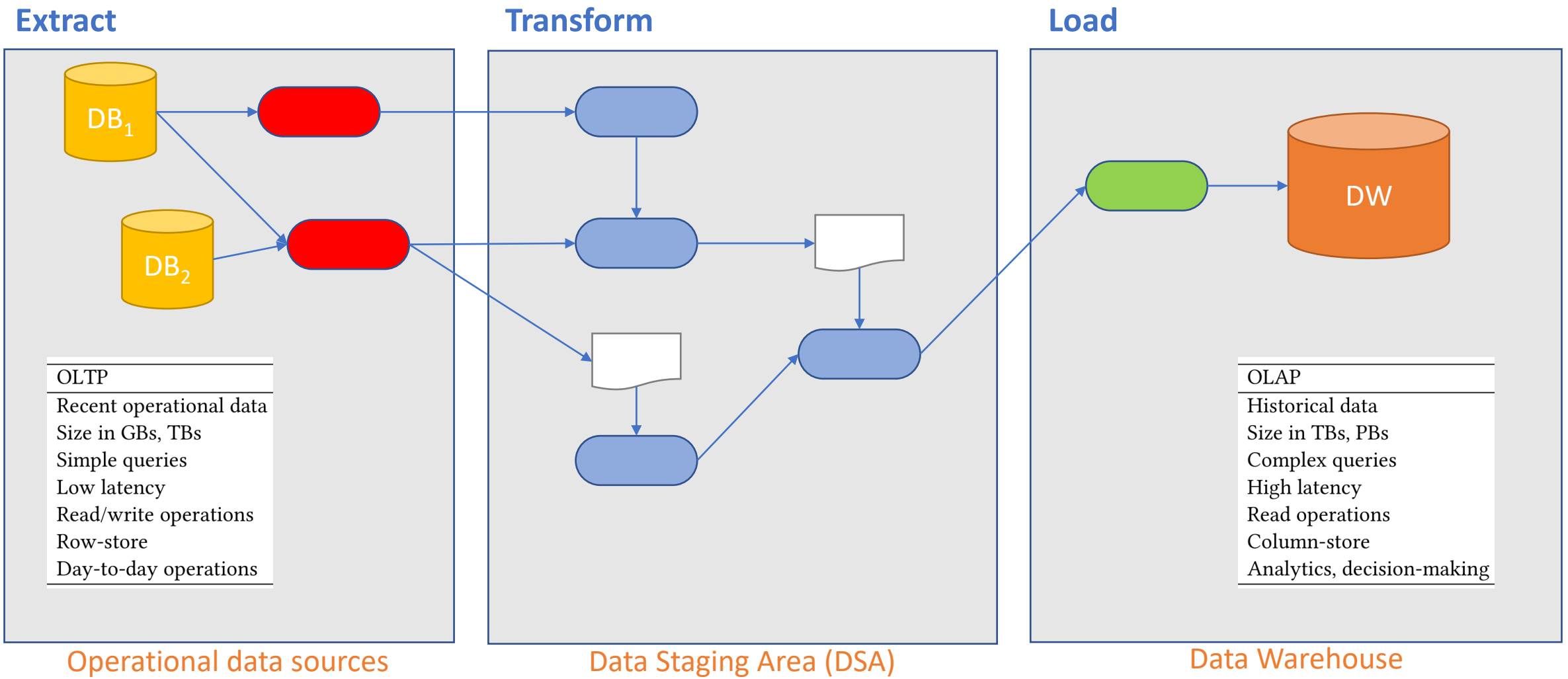
UNIVERSITY OF IOANNINA

Agenda

- What is ETL?
- A trip down history lane: a 20-year recap
- Conceptual modeling for ETL
- ETL – present times
- ETL – the future
- Conclusions

What is ETL?

What is ETL? – traditional approach



Challenges (take #1)

- Design aspects
  - Schema mappings
    - Data integration, data exchange
  - Data cleansing and data quality
    - Rules based on integrity constraints
    - Duplicate/error detection
    - ML to improve accuracy of cleansing
  - Additional complex transformations
    - Data lineage, 1-N mappings, generating new values and new fields (e.g., SK), analytics, UDFs, …
  - Hard or infeasible to express most ETL transformations w/ traditional relational ops
  - Data and control flow
  - How do we measure how 'good' an ETL flow is?

- Engineering aspects
  - Decide cadence
    - When, how often, …?
    - Batch, micro-batches, streaming?
  - Data extraction
    - Without impacting the sources significantly
    - Without losing on freshness
  - Various data types
    - Structured, semi-structured, unstructured, flexible schema
  - Many data sources, targets, engines
    - Heterogeneous, federated, distributed
  - Programming heterogeneity
    - Complex, multi-fragment flows
    - Scripts, SQL constructs, UDFs, lambda, workflows –all in the same flow
  - Conflicting objectives
    - Performance, maintenance, fault-tolerance
  - Optimization
    - End-to-end / individual transformations

A trip down history lane

ETL research: a 20-year recap

- ETL Design
  - Logical model [DMDW'02] && Conceptual design [DOLAP'02] –first work on ETL conceptual modeling
  - UML, BPMN, BPEL modeling
  - Business process models, Web services, Hypercubes
  - Semantic Web, ontologies to automate ETL design creation
  - Graph-based logical ETL design
  - Automated mapping from conceptual to logical models
- End-to-end ETL Optimization
  - Optimization as a state-space problem [ICDE'05] –first work on ETL optimization
  - Multiple optimization objectives: performance, maintainability, fault-tolerance
  - Intermediate results materialization
  - Parallelization and partition-based workload scheduling
  - Physical design and scheduling
  - Data flows with MapReduce-like UDFs
  - Multi-engine flow optimization

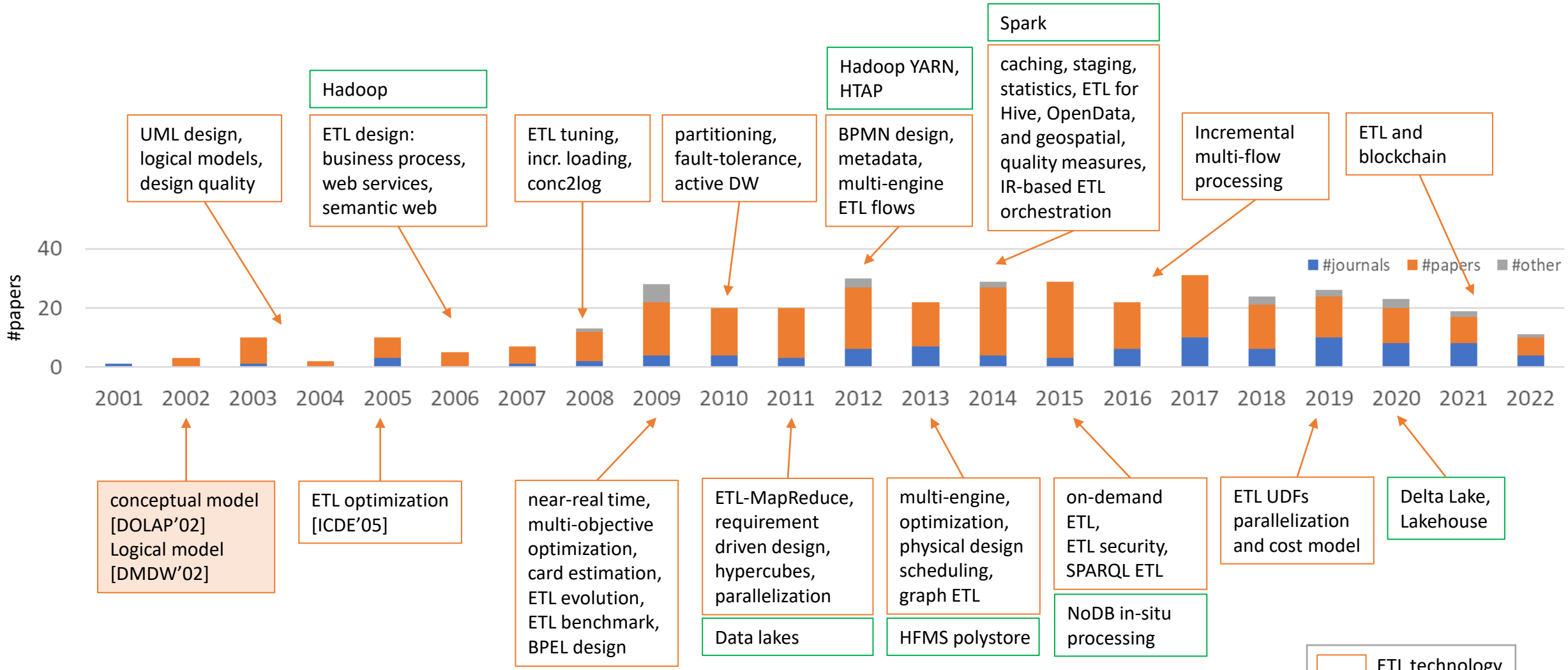
ETL research: a 20-year recap

- Optimization of ETL operations
  - Efficient extraction of delta values
  - Schema transformations
    - data mappers, pivot/unpivot
  - Data cleansing transformations
  - Lineage of data transformations
  - Efficient resumption of interrupted data flows
  - Change-table techniques for incremental view maintenance
  - Efficient data cubes
  - Cardinality estimation in ETL processes
  - ETL tasks in the context of Map-Reduce
  - Real-time processing of ETL operations
- ETL lifecycle & governance
  - Monitoring/testing through regression tests
  - Explaining ETL processes with NL descriptions
  - Managing ETL evolution
  - Cataloguing frequent ETL patterns
  - ETL benchmarks
  - 10+ research system prototypes

ETL research timeline [20-years: 2002-2022]

~400 publications
- 270 papers
- 100 articles

avg #pubs/year
01-10: 10
11-21: 26

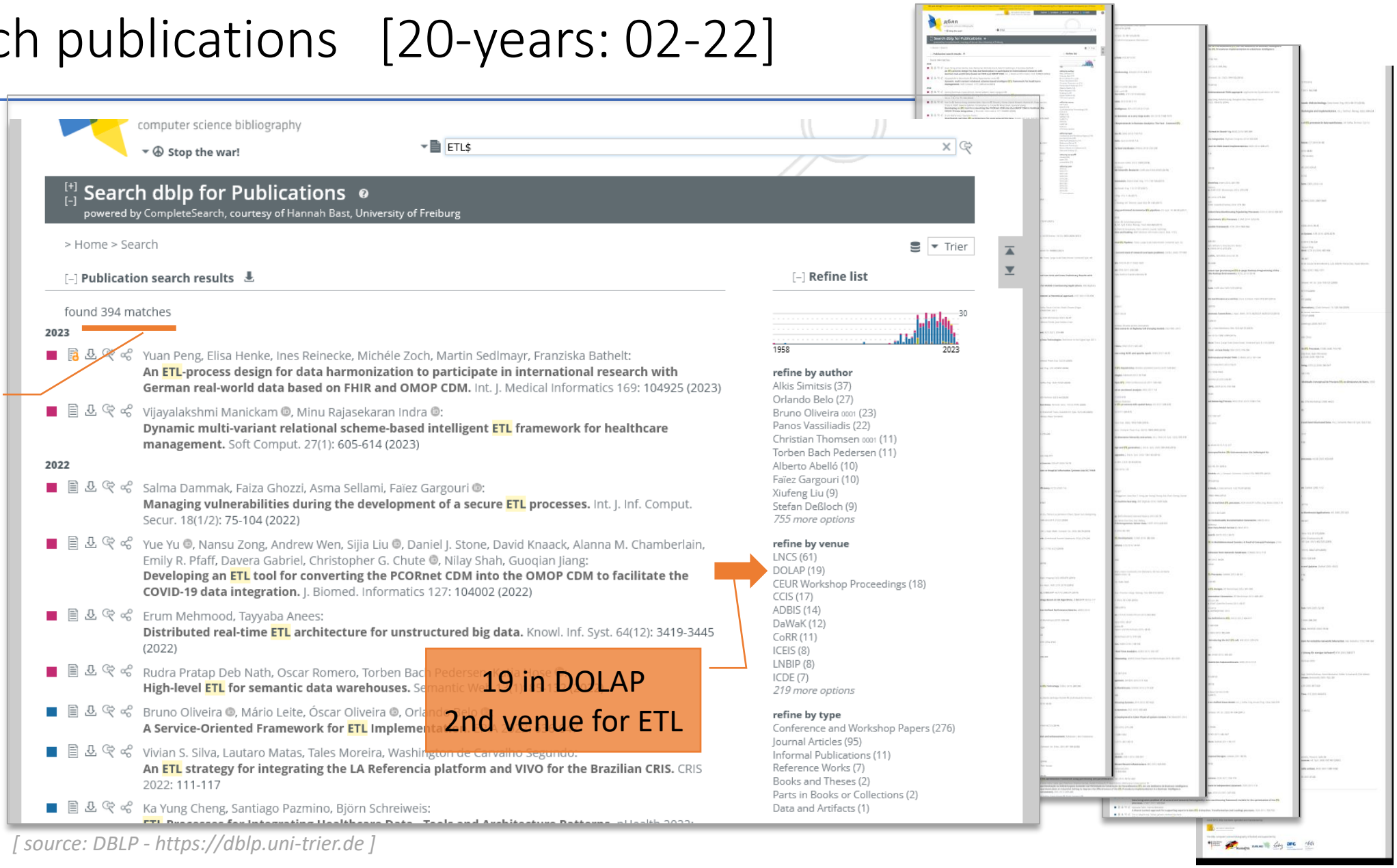


ETL research publications [20-years: 02-22]

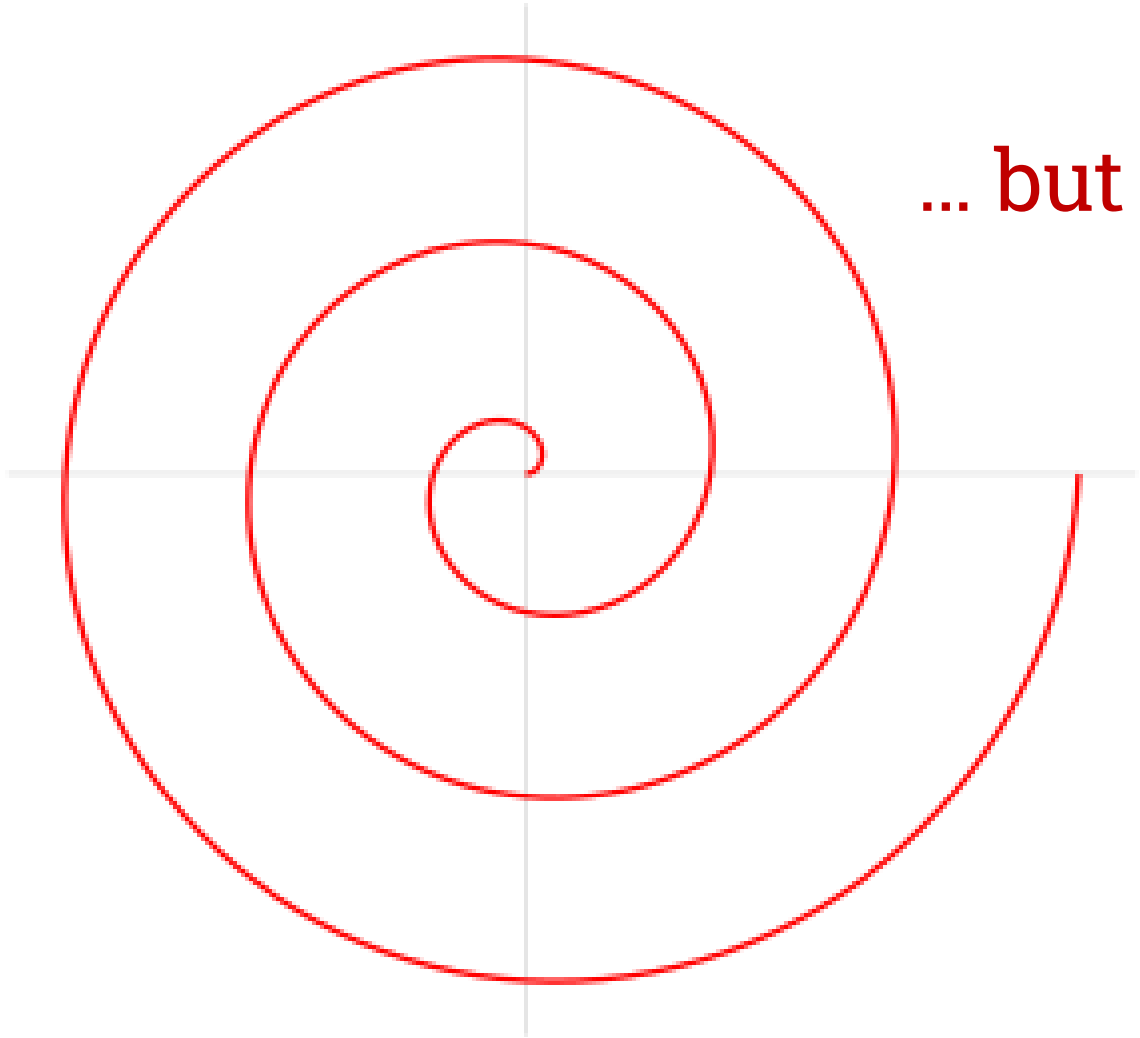
Search query: **ETL$**

[papers with ETL in the title –i.e., not all ETL related papers]

~400 publications

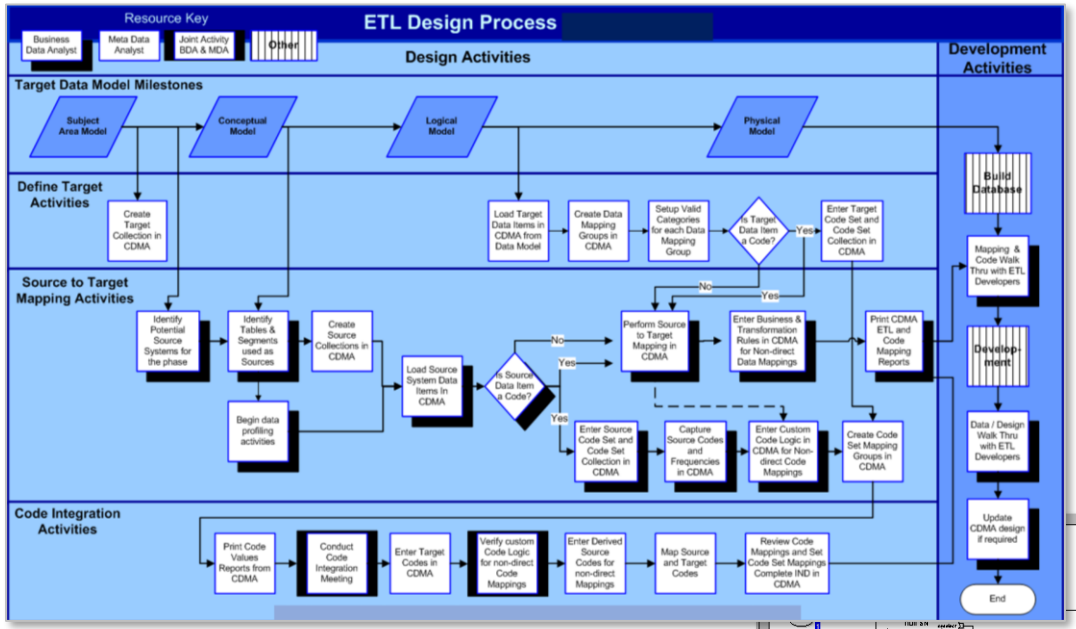


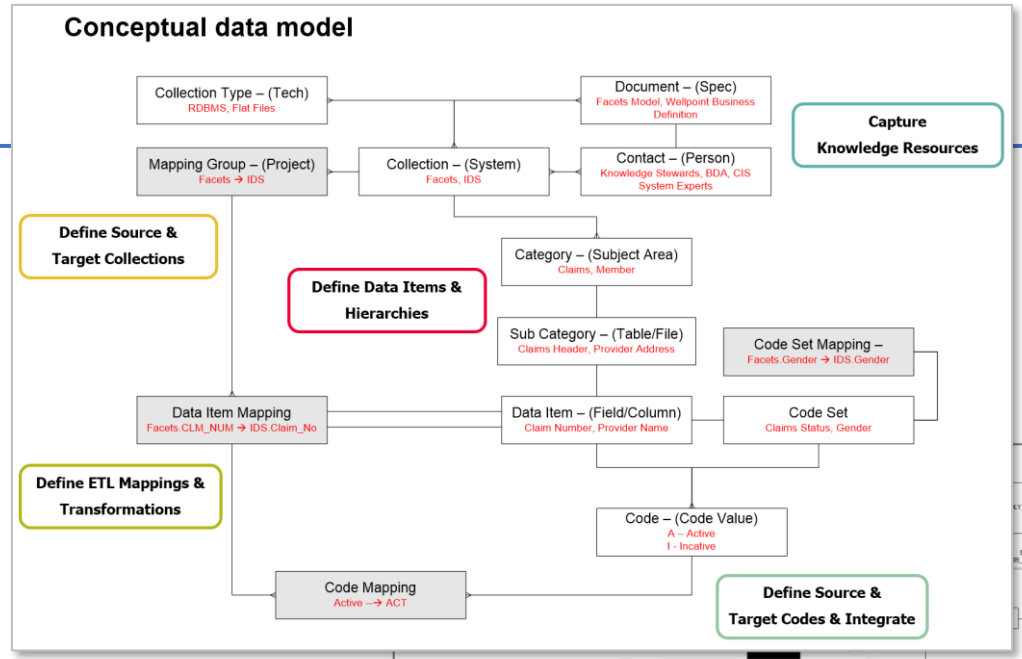
[source: DBLP - https://dblp.uni-trier.de]

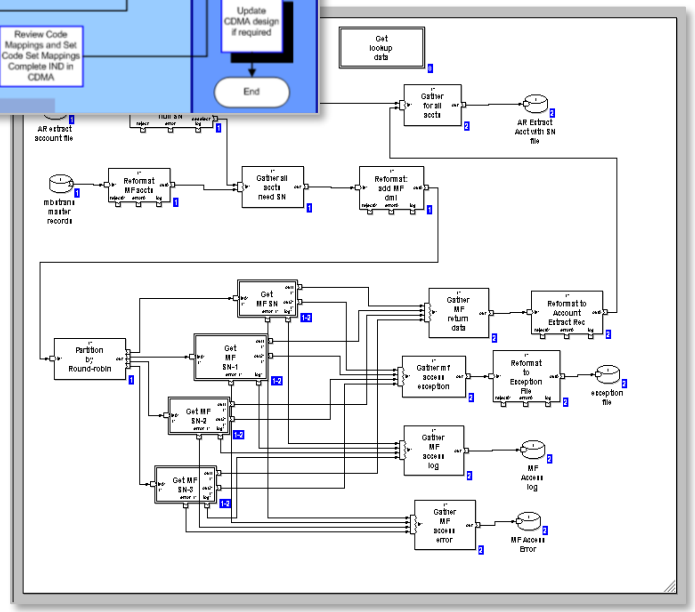


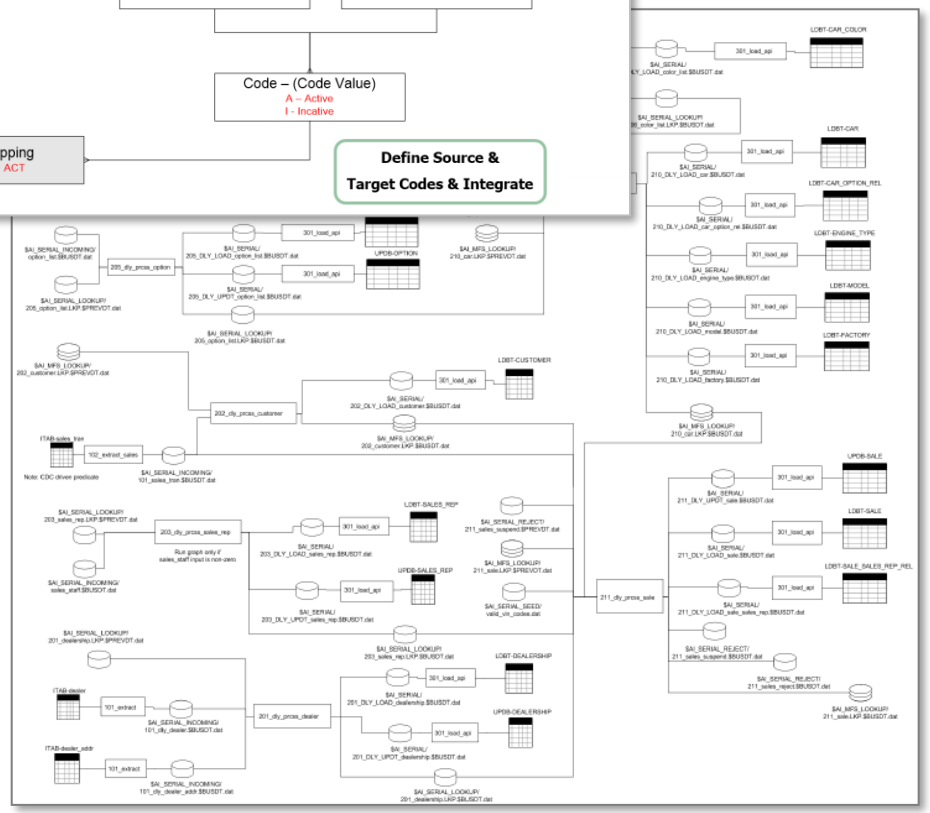
… but how did we start? …

Real-world ETL scenarios









Can you spot the commonalities?

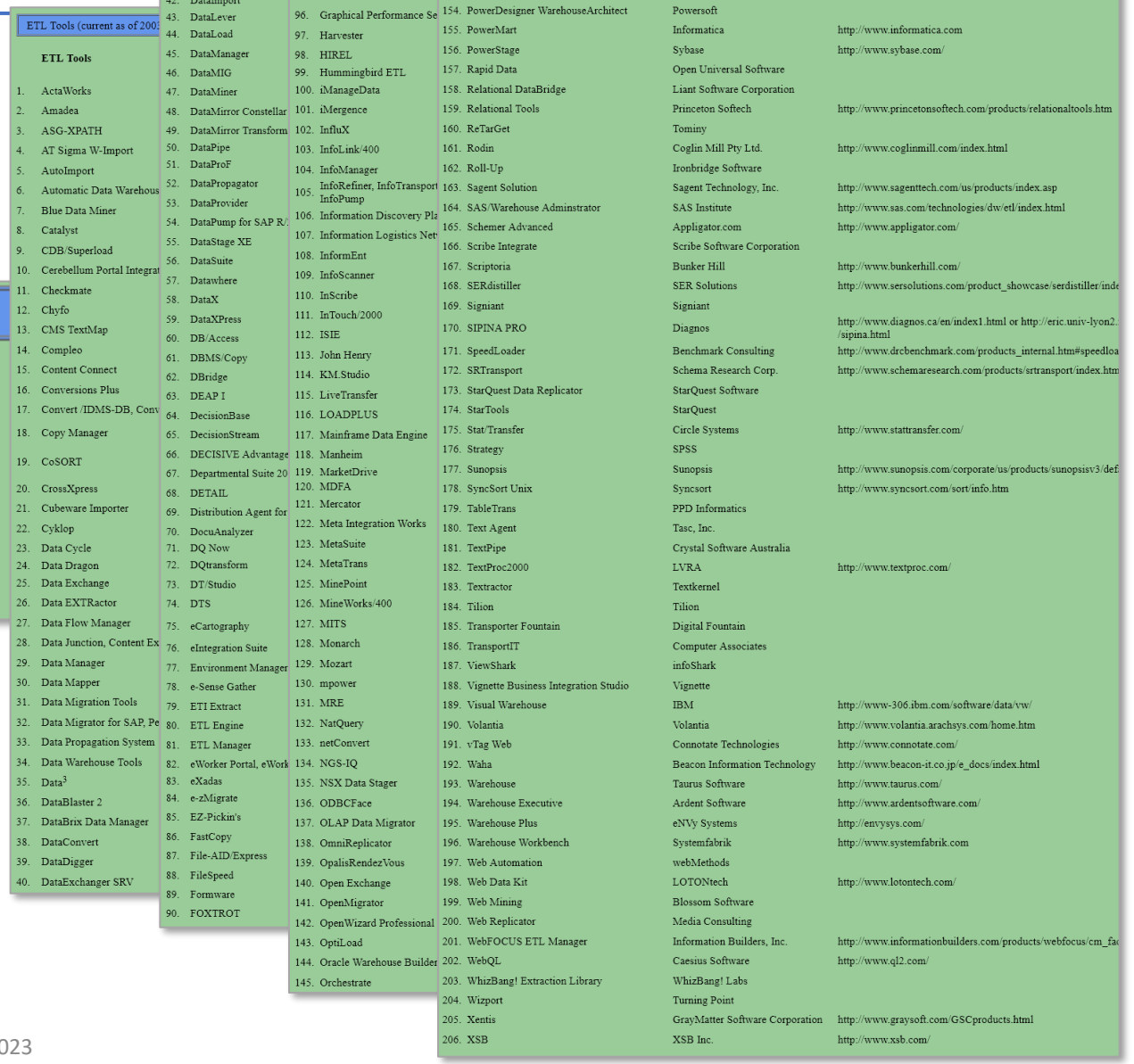
ETL technology – 20 years ago

- ETL was flourishing in the industry…
  - 200+ commercial ETL tools in 2003

ETL Tools (current as of 2003)

	ETL Tools	Vendor
1.	ActaWorks	Acta Technologies
2.	Amadea	ISoft
3.	ASG-XPATH	Allen Systems Group

[src: https://web.imsi.athenarc.gr/~alkis/publications/ETLTools.htm]



ETL technology – 20 years ago

- … but research wasn't particularly thrilled

Reviewer #3

1: Is the paper relevant to VLDB Conference?: Yes

I am unable to grasp the novelty in the idea

3: Technical depth: Weak reject

6: Presentation: Weak accept

7: Overall rating: Weak reject

8: Reviewer confidence: High

the entire effort seems identical to query optimization

11: Summary of main contribution and rationale for your recommendation (1-2 paragraphs):

The paper proposes techniques for optimization of ETL workflows in data warehousing environments.

I am unable to grasp the novelty in the idea -- the entire effort seems identical to query optimization. Maybe the differential is in

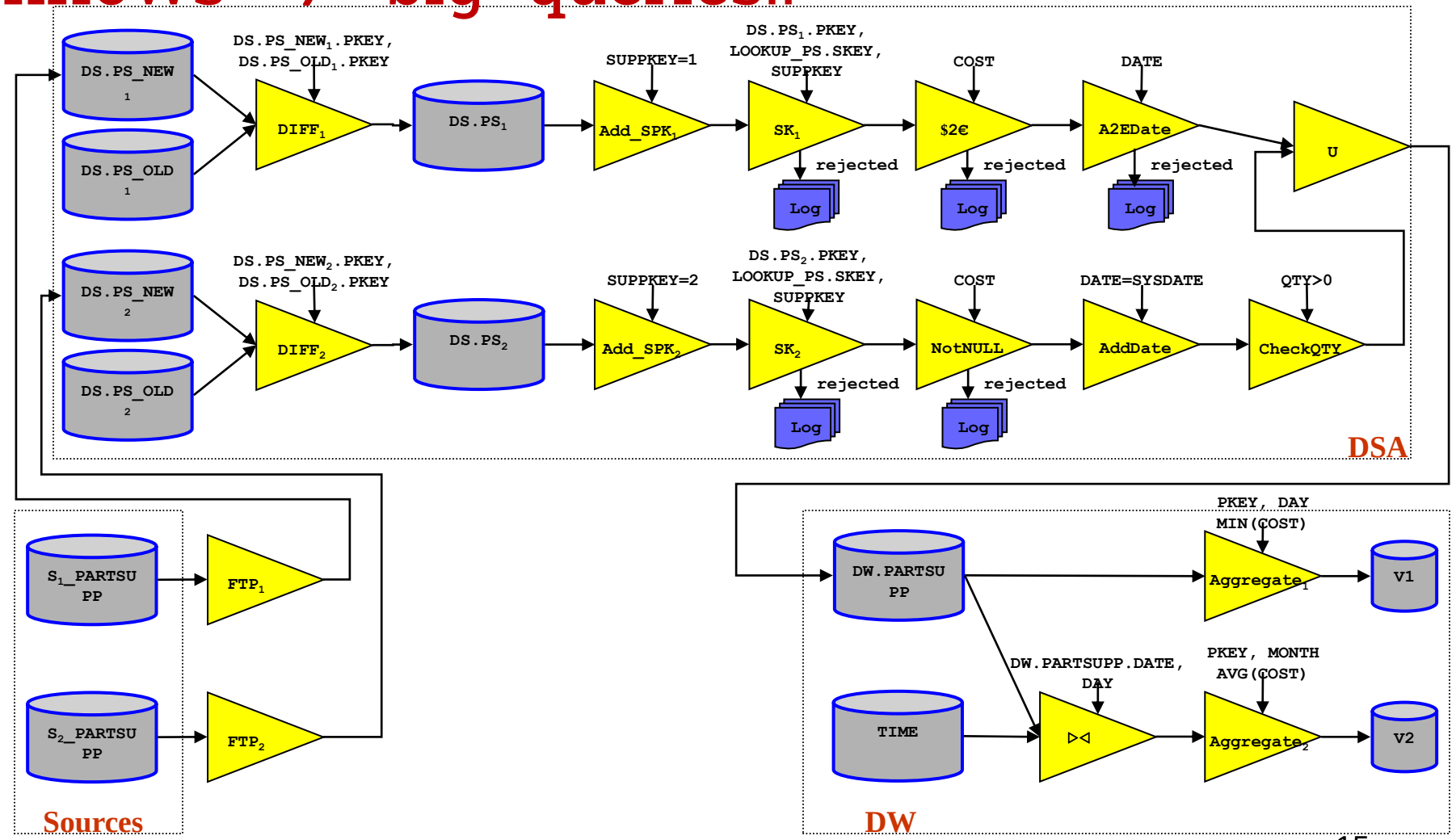
Maybe the differential is in a slightly richer set of operations

12: Detailed comments to authors:

see above

Data Warehouses ≠ collections of materialized views!!

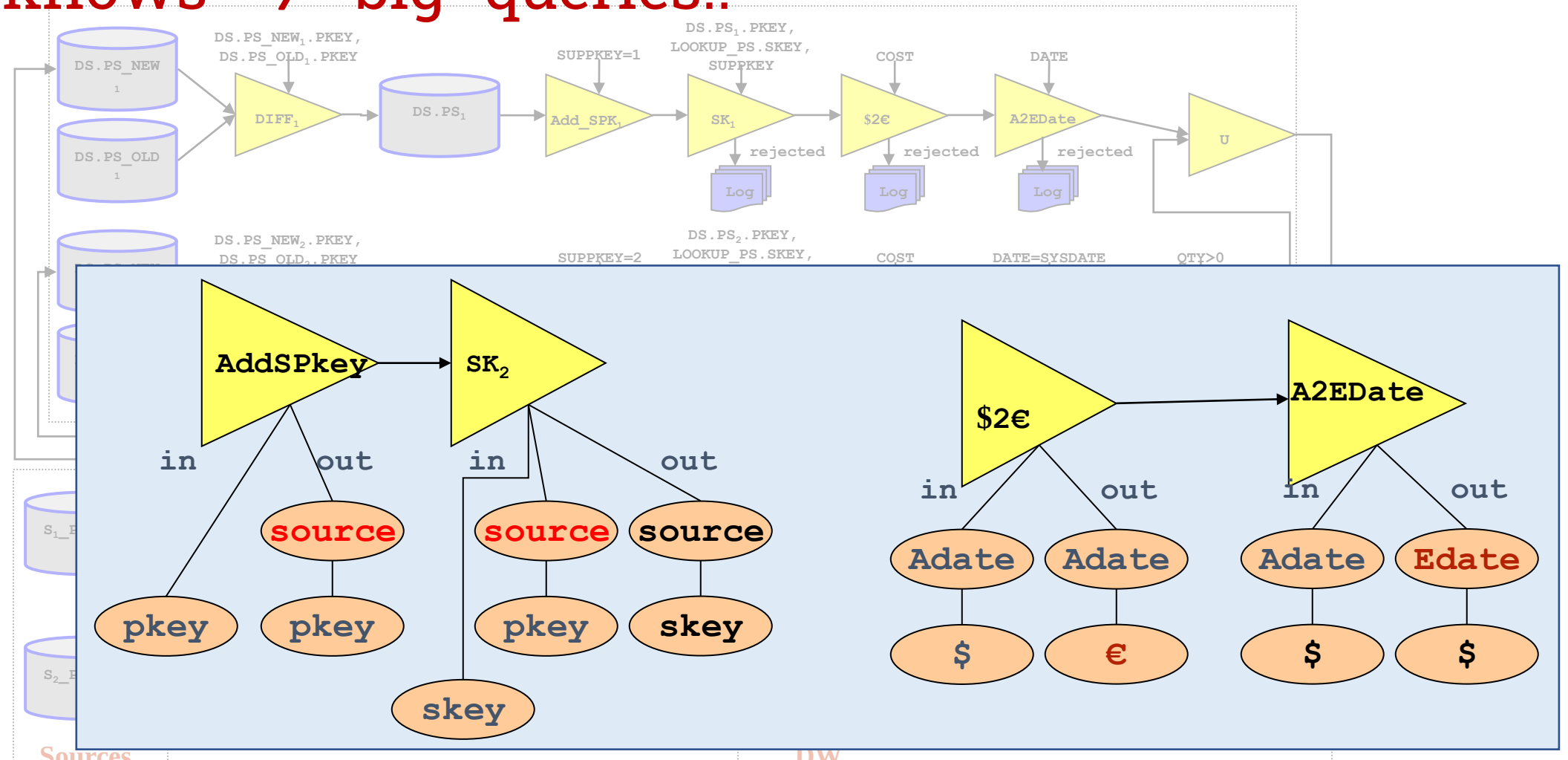
ETL workflows ≠ "big" queries!!



This has always been the vision

Data Warehouses ≠ collections of materialized views!!

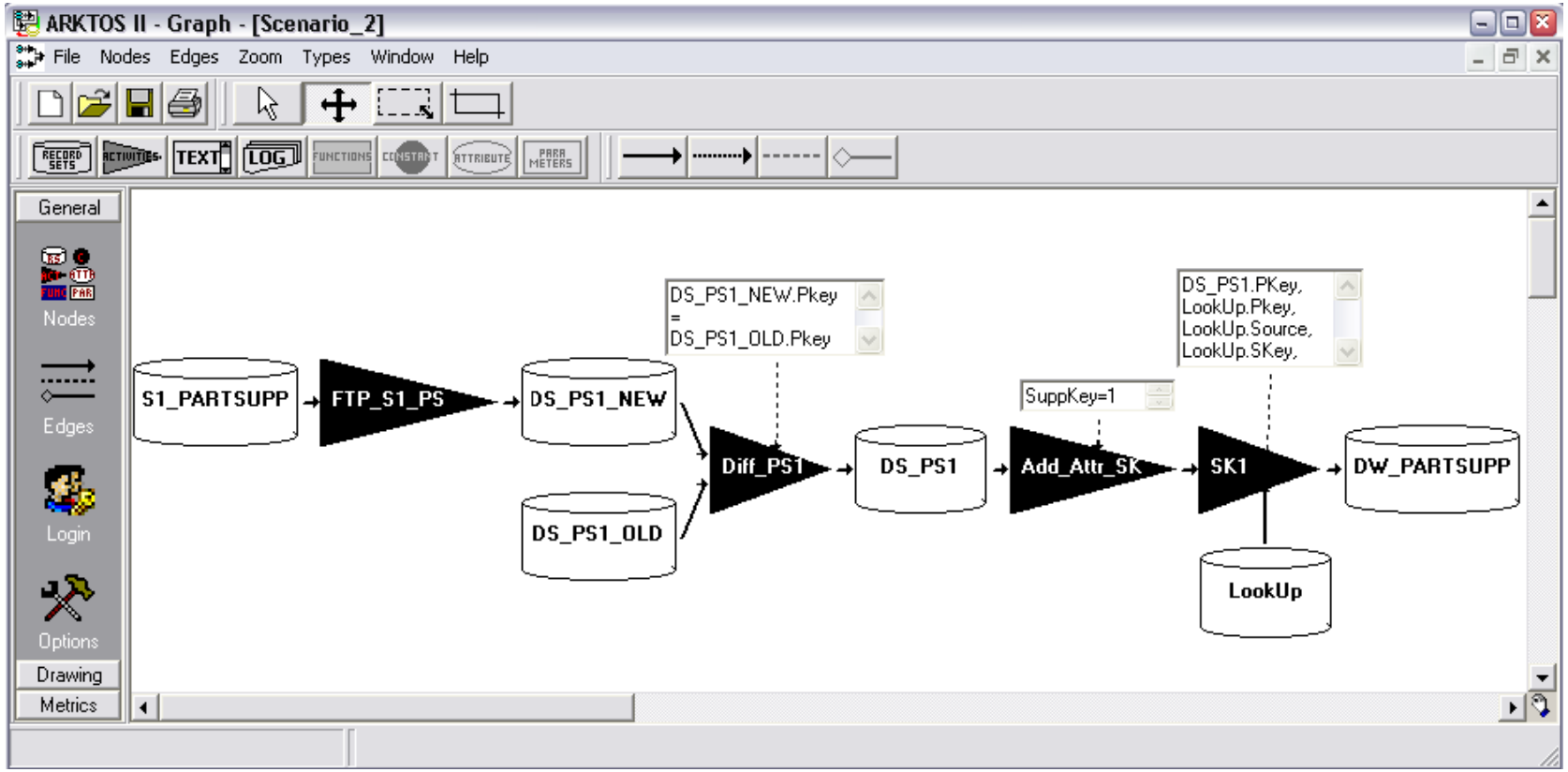
ETL workflows ≠ "big" queries!!



This has always been the vision

Arktos II – Our in-house ETL design project

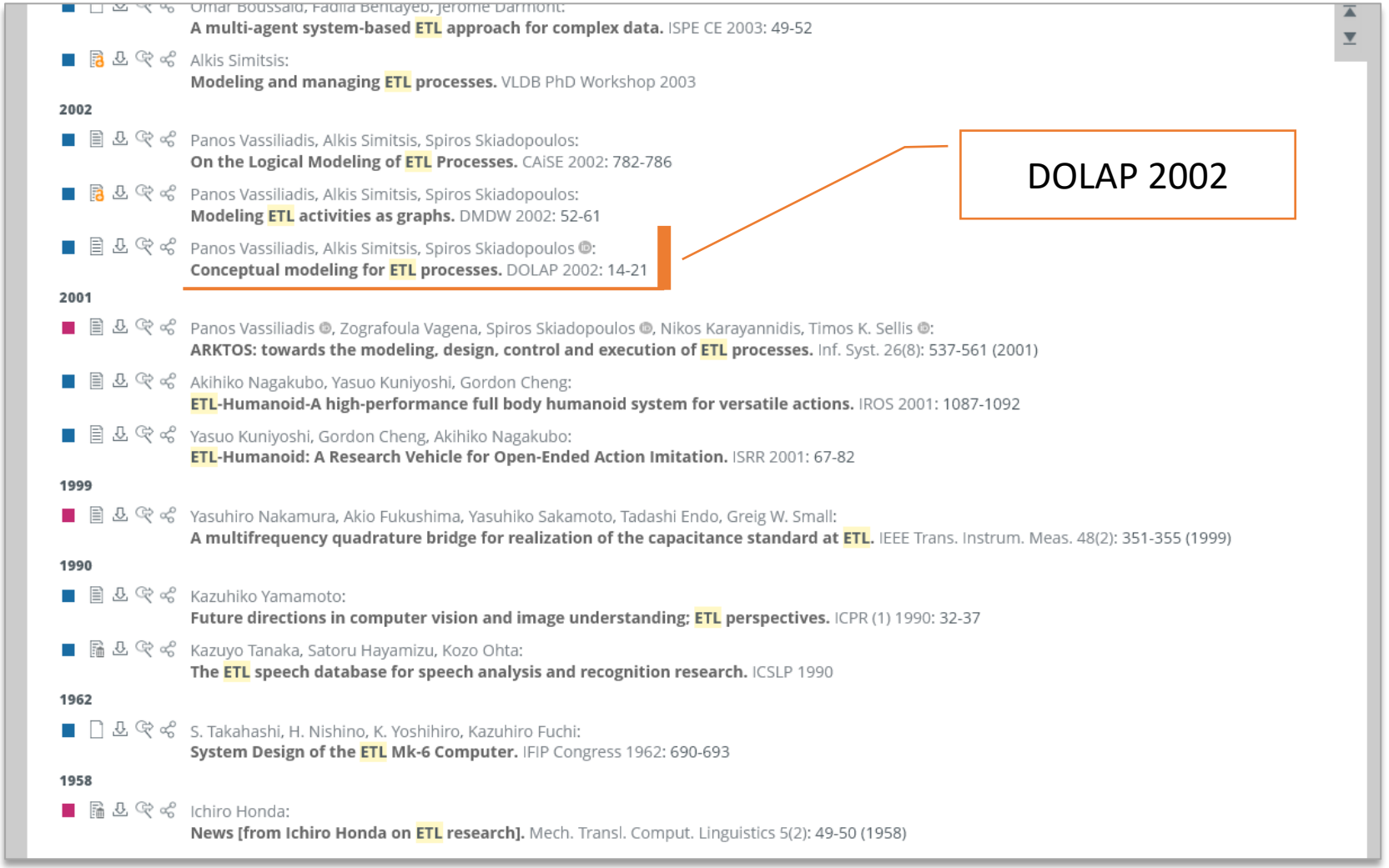
[circa 2000-2005]



Core Arktos team

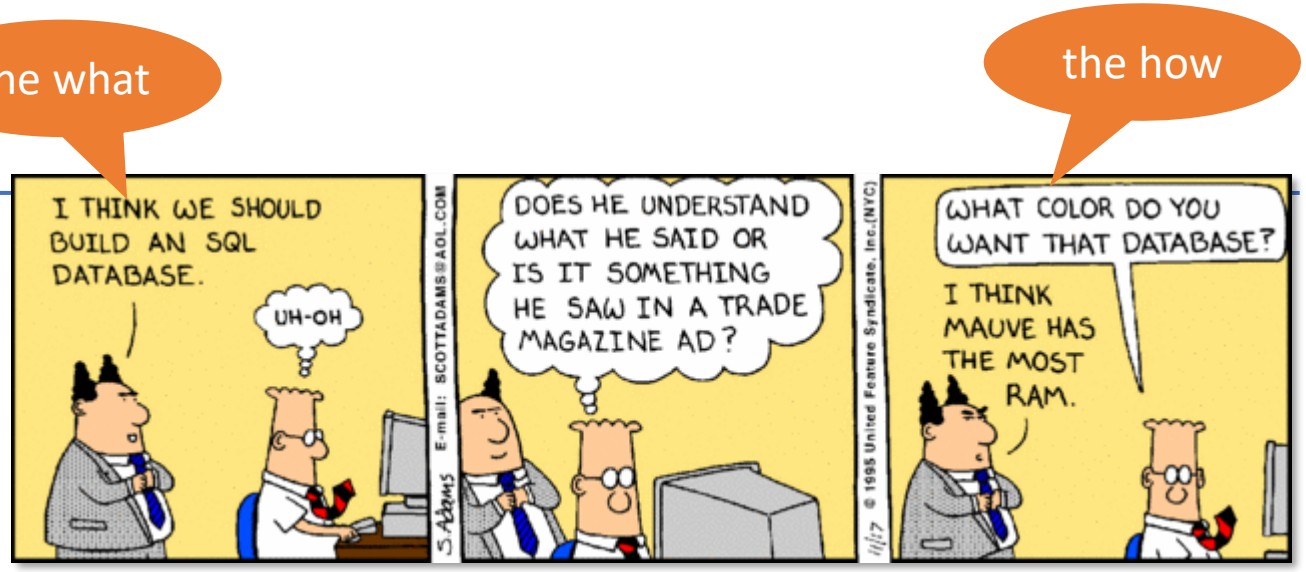
- P. Georgantas
- N. Karayannidis
- G. Papastefanatos
- A. Simitsis
- T. Sellis
- S. Skiadopoulos
- M. Terrovitis
- Z. Vagena
- P. Vassiliadis
- Y. Vassiliou

ETL research publications [circa 2002]



[source: DBLP - https://dblp.uni-trier.de]

Why a conceptual model?



[Source: dilbert.com, Nov 17, 1995]

- Task
  - Given fixed OLTP and OLAP schemas
  - Develop an efficient and scalable design to propagate data from the former to the latter
- Challenges
  - Different audiences: business users (the what) and IT professionals (the how)
  - Lack of any kind of methodology, formalism, standard, or even recorded collective experience
    - Ad hoc, in-house built solutions → hard to maintain, difficult to reuse
  - Scalable design to capture schema mappings, data/schema lineage, evolution
  - Provide a path to logical and physical models
- State-of-the-art in early 2000
  - Research: n/a
  - Industry: ad hoc, tedious, overcomplex, customized methods employing multiple documents, sheets, forms

Conceptual modeling for ETL

Conceptual modeling for ETL processes
[ACM DOLAP 2002]

Conceptual Modeling for ETL Processes

Panos Vassiliadis Alkis Simitsis Spiros Skiadopoulos
National Technical University of Athens,
Dept. of Electrical and Computer Eng.,
Iroon Polytechniou 9, 157 73, Athens, Greece,
Tel: +30-10-772-1602
pvassil@dbnet.ece.ntua.gr asimi@dbnet.ece.ntua.gr spiros@dbnet.ece.ntua.gr

ABSTRACT
Extraction-Transformation-Loading (ETL) tools are pieces of software responsible for the extraction of data from several sources, their cleansing, customization and insertion into a data warehouse. In this paper, we focus on the problem of the definition of ETL activities and provide formal foundations for their conceptual representation. The proposed conceptual model is (a) customized for the tracing of inter-attribute relationships and the respective ETL activities in the early stages of a data warehouse project; (b) enriched with a 'palette' of a set of frequently used ETL activities, like the assignment of surrogate keys, the check for null values, etc; and (c) constructed in a customizable and extensible manner, so that the designer can enrich it with his own re-occurring patterns for ETL activities.

Categories and Subject Descriptors
H.2.1 [Database Management]: Logical design - *data models, schema and subschema.*
General Terms
Design
Keywords
Data warehousing, ETL, conceptual modeling

1. INTRODUCTION
Extraction-Transformation-Loading (ETL) tools is a category of specialized tools with the task of dealing with data warehouse homogeneity, cleaning and loading problems. [29] reports that ETL and Data Cleaning tools are estimated to cost at least one third of effort and expenses in the budget of the data warehouse while [8] mentions that this number can rise up to 80% of the development time in a data warehouse project. [14] mentions that the ETL process costs 55% of the total costs of data warehouse runtime. Still, due to the complexity and the long learning curve of these tools, many organizations prefer to turn to in-house development to perform ETL and data cleaning tasks. In fact, while data warehouse expenses are expected to come up to 14 billion dollars worldwide, projected sales for ETL and data cleaning tools are expected to rise to only (!) 300 million dollars.

DOLAP'02, November 8, 2002, McLean, Virginia, USA.

Thus, it is apparent that the design, development and deployment of ETL processes, which is currently performed in an ad-hoc, in house fashion, needs modeling, design and methodological foundations. Unfortunately, as we shall show in the sequel, the research community has a lot of work to do to confront this shortcoming. In the rest of the paper, we will not discriminate between the tasks of ETL and Data Cleaning and adopt the name ETL for both these kinds of activities.

In Fig. 1, we abstractly describe the general framework for ETL processes. In the bottom layer we depict the data stores that are involved in the overall process. On the left side, we can observe the original data providers (typically, relational databases and files). The data from these sources are extracted (as shown in the upper left part of Fig. 1) by extraction routines, which provide either complete snapshots or differentials of the data sources. Then, these data are propagated to the *Data Staging Area* (DSA) where they are transformed and cleaned before being loaded to the data warehouse. The data warehouse is depicted in the right part of Fig. 1 and comprises the target data stores, i.e., fact tables and dimension tables. Eventually, the loading of the central warehouse is performed through the loading activities depicted on the upper right part of the figure.

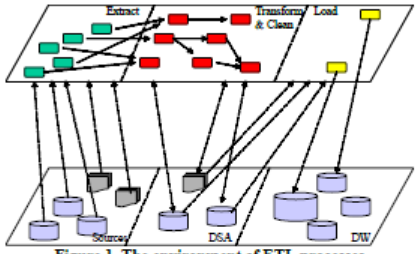


Figure 1. The environment of ETL processes

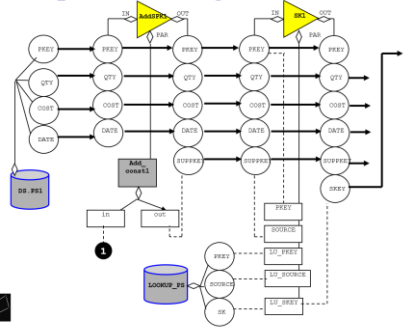
In this paper, we focus on the conceptual part of the definition of the ETL process. More specifically, we are dealing with the earliest stages of the data warehouse design. During this period, the data warehouse designer is concerned with two tasks which are practically executed in parallel. The first of these tasks involves the *collection of requirements* from the part of the users. The second task, which is of equal importance for the success of the data warehousing project, involves *the analysis of the structure and content of the existing data sources* and their *intentional mapping to the common data warehouse model.*

Conceptual Modeling for ETL processes

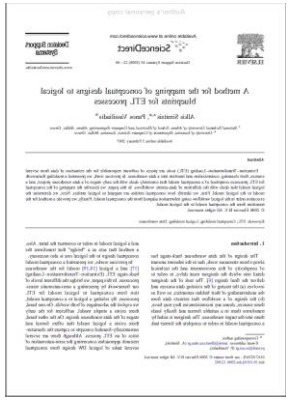
Panos Vassiliadis, Alkis Simitsis, Spiros Skiadopoulos
{pvassil,asimi,spiros}@dblab.ece.ntua.gr

National Technical University of Athens
KDBS Laboratory
http://www.dbnet.ece.ntua.gr

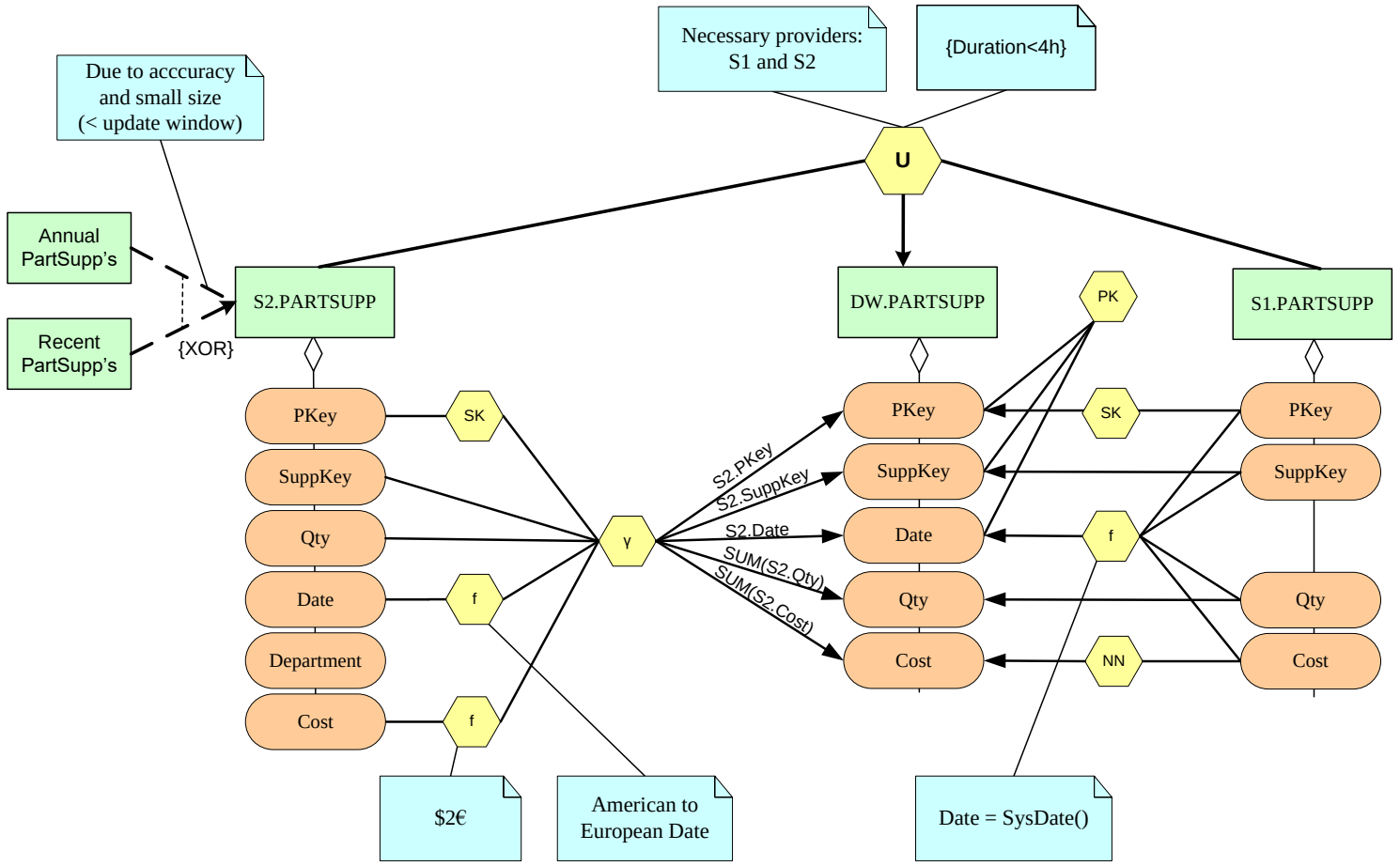
Graph Modeling [DMDW'02]



Vassiliadis,Simitsis,Skiadopoulos. On the Logical Modeling of ETL Processes. CAiSE'02 Toronto, 2002



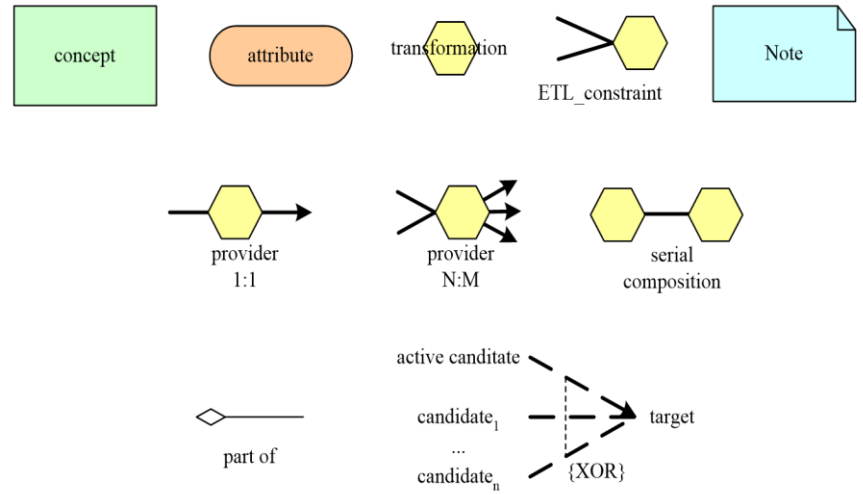
Conceptual Model for ETL



- Early stages of design
- Schema mappings: Trace the mapping of the attributes of the data sources to the attributes of the DW tables …
- … along with the necessary constraints and transformations for the ETL process…
- … as well as provision alternatives, transformation composition, and an extensible palette of transformations

- Rather surprising to have attributes and transformations as core concepts of the model

Conceptual Model

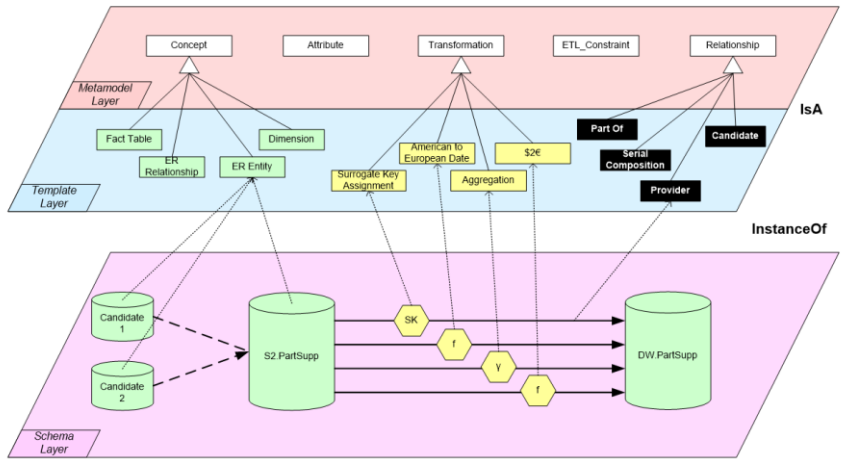


Vassiliadis,Simitsis,Skiadopoulos - DOLAP'02 37

Conceptual modeling for ETL processes

[ACM DOLAP 2002]

Instantiation & Specialization Layers



Instantiation & Specialization Layers

Filters

Selection (σ)
Not null (NN)
Primary key violation (PK)
Foreign key violation (FK)
Unique value (UN)
Domain mismatch DM)

Unary transformations

Push
Aggregation (γ)
Projection (π)
Function application (f)
Surrogate key assignment (SK)
Tuple normalization (N)
Tuple denormalization (DN)

Binary transformations

Union (U)
Join (⊳⊲)
Diff (Δ)
Update Detection (Δ_{UPD})

Composite transformations

Slowly changing dimension (Type 1,2,3) (SDC-1/2/3)
Format mismatch (FM)
Data type conversion (DTC)
Switch (σ*)
Extended union (U)

File operations

EBCDIC to ASCII conversion (EB2AS)
Sort file (Sort)

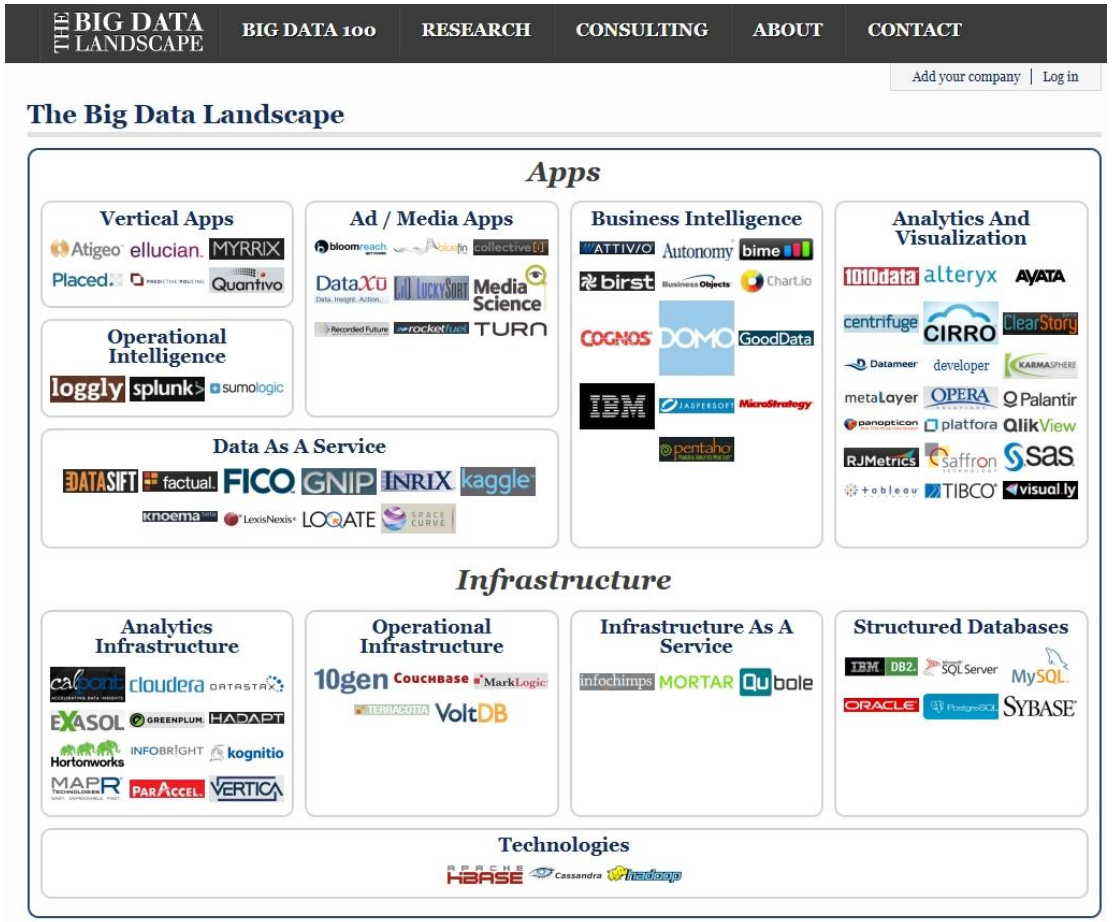
Transfer operations

Ftp (FTP)
Compress/Decompress (Z/dZ)
Encrypt/Decrypt (Cr/dCr)

ETL – present times

The analytics landscape

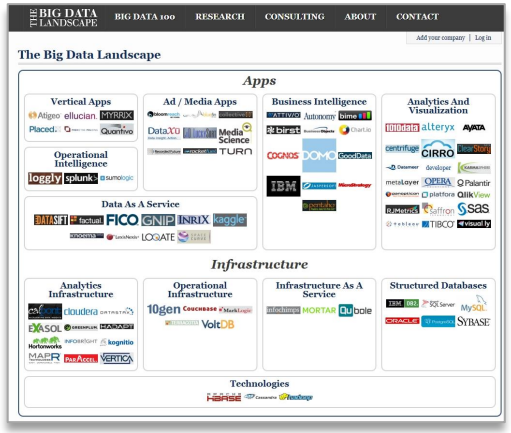
2012



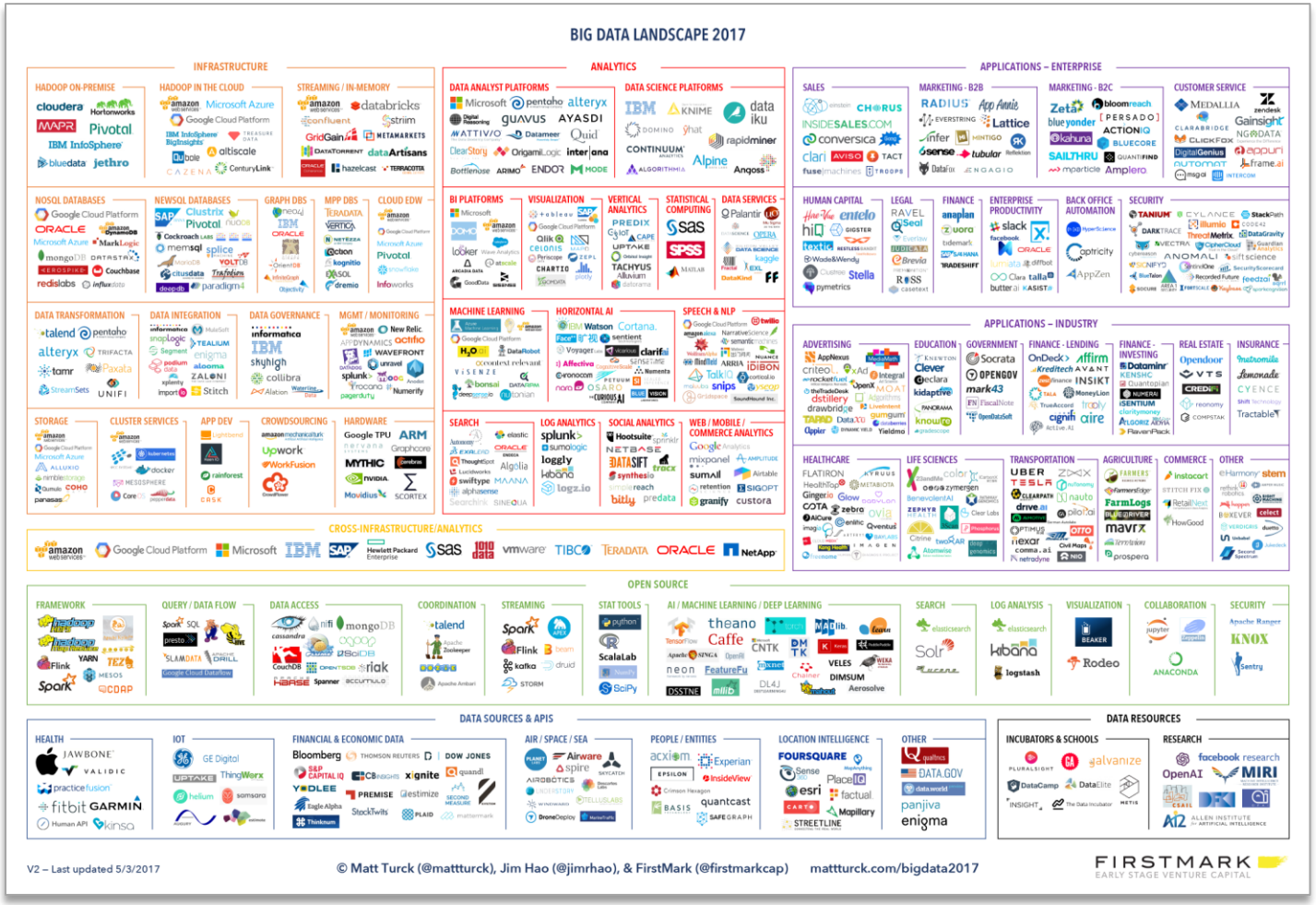
[src: https://mattturck.com]

The analytics landscape

2012 **2017**

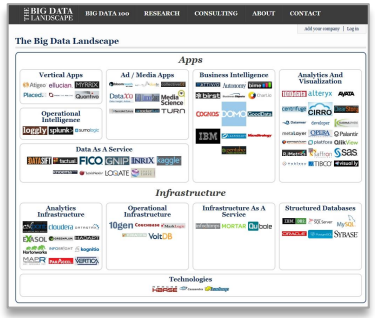


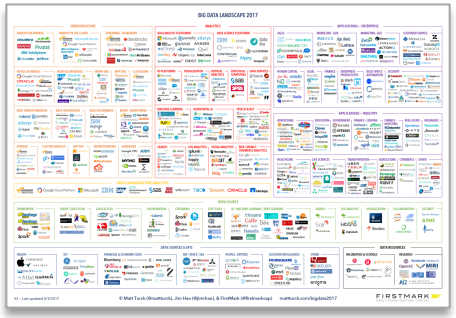
[src: https://mattturck.com]



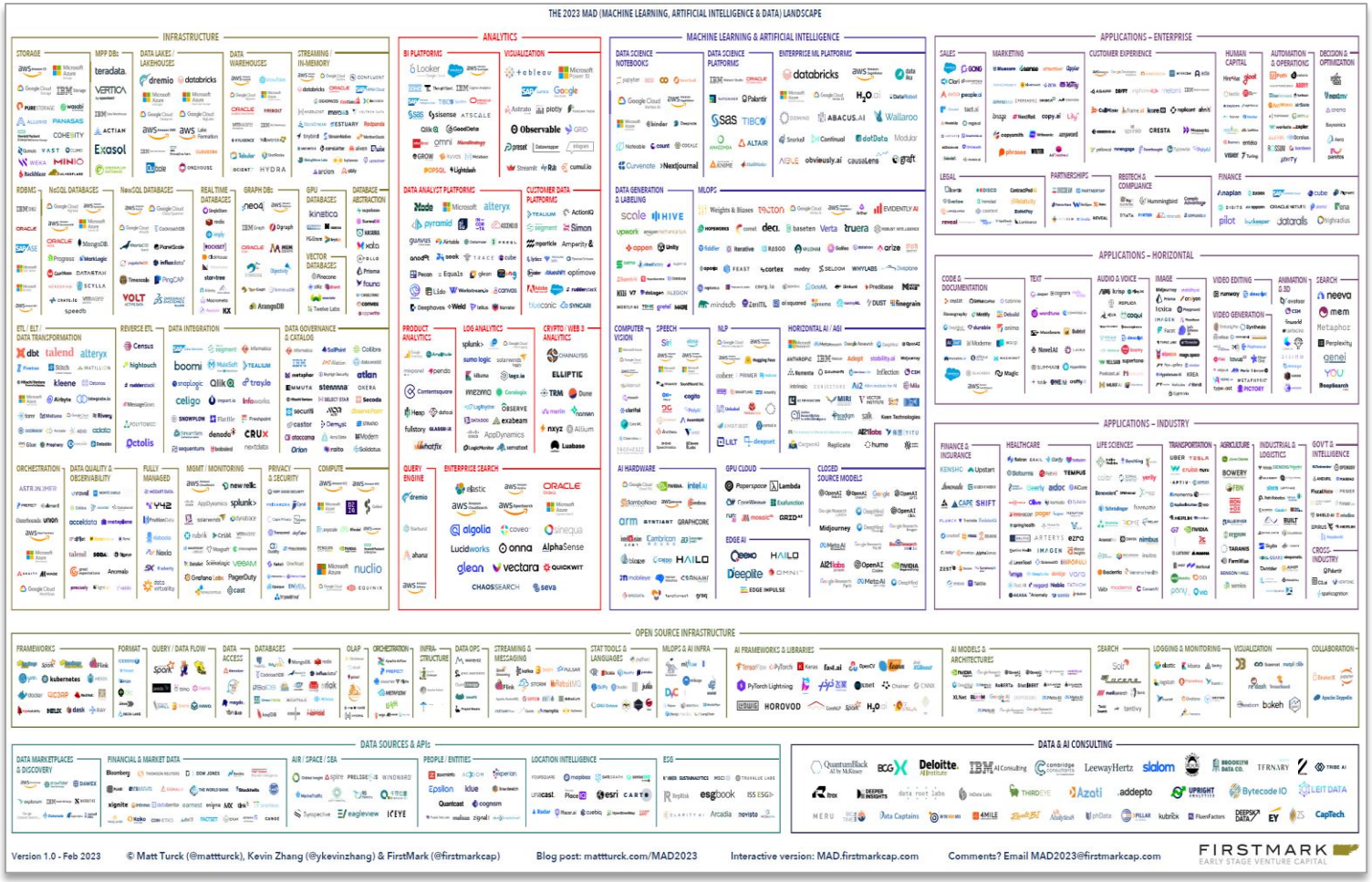
The analytics landscape

2012 **2017** **2023**

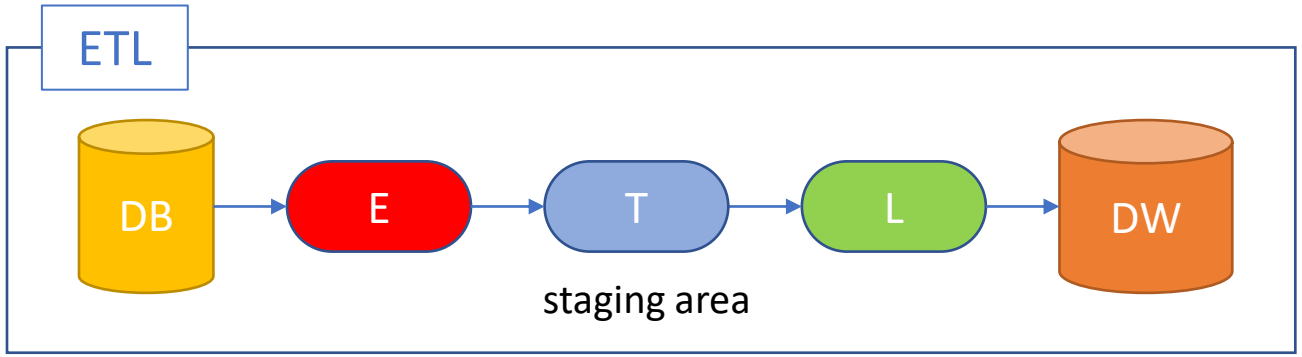


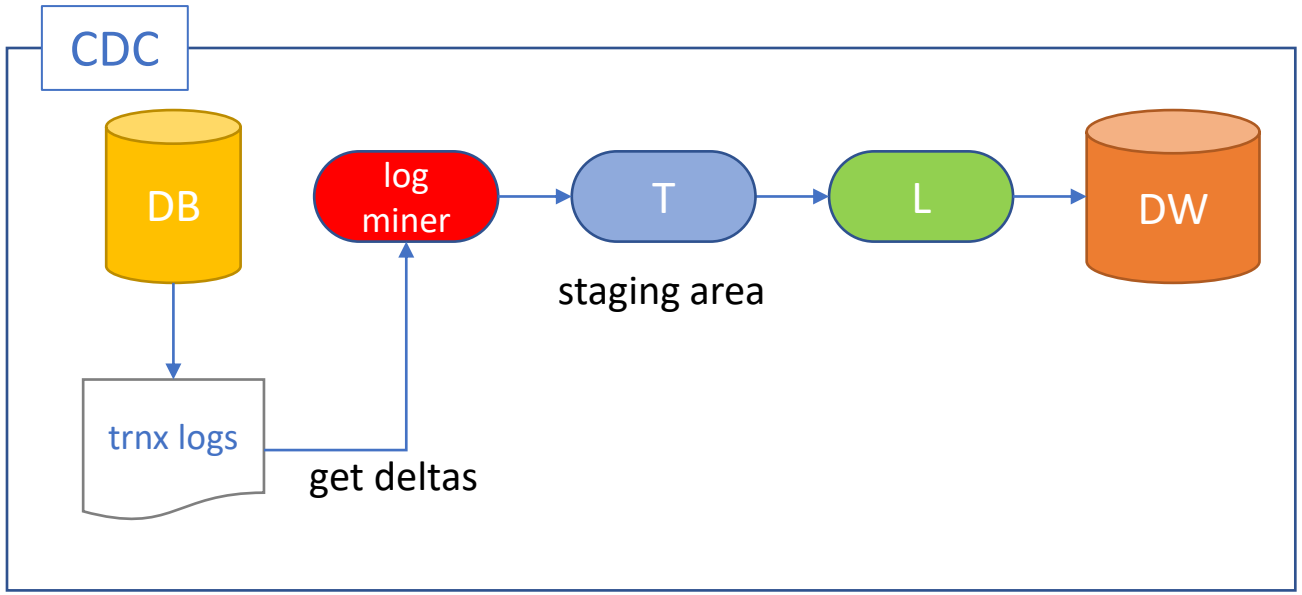


[src: https://mattturck.com]



Evolution of the ETL architecture

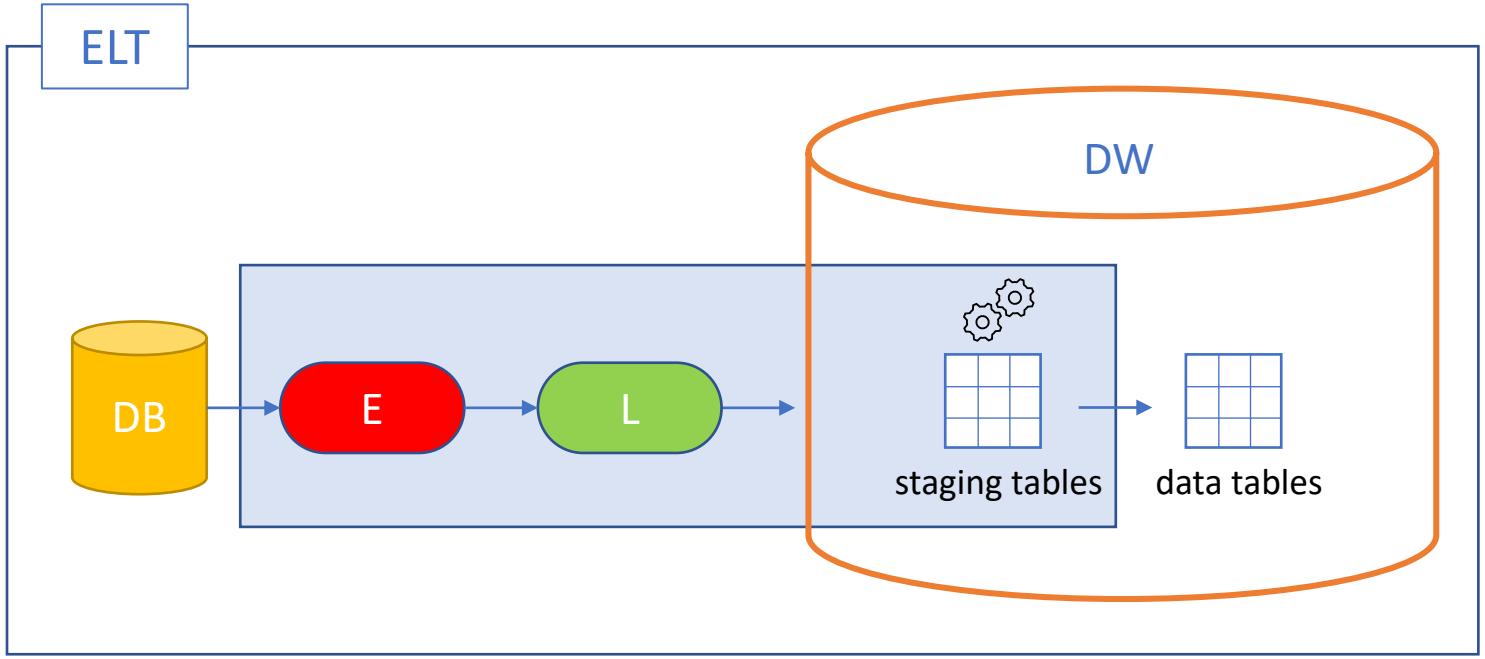




Change Data Capture optimizations

- Apply trnx in the same order
- Batch-optimized
- Load w/ native perf, use MPP
- Capture and stream data changes into msg broker (e.g., Kafka)

Evolution of the ETL architecture



ELT particularly popular in cloud deployments

- Often "EL" → data replication
- Cheaper storage on-prem / cloud
- Cheaper compute: Spark, Hadoop, Beam, cloud engines
- Streaming data, ready for analysis at target

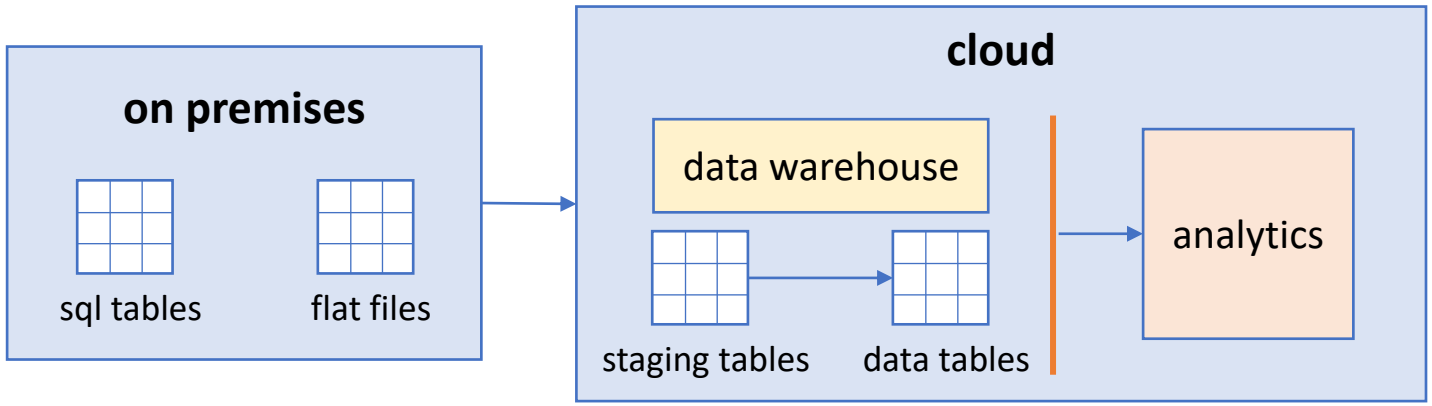
There are other flavors too

- ETLT, ELTL, …

Trends in ETL processing

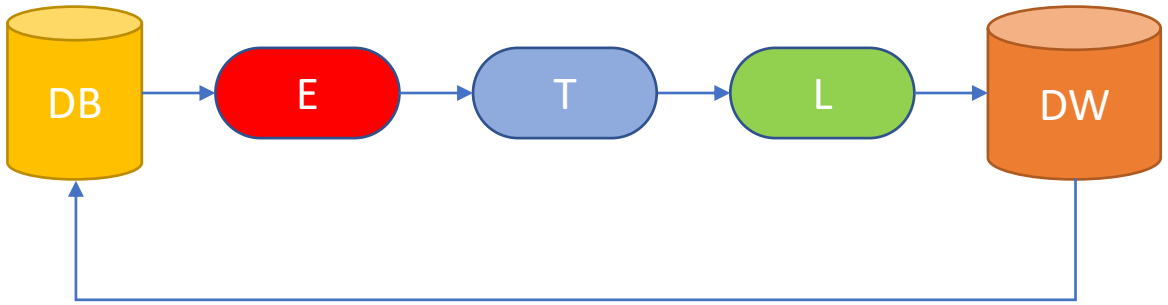
- Streaming ETL
  - Various sources, larger volumes, high speeds
  - Data sources/consumers should connect/disconnect w/o interrupting the systems (horizontal scaling)
  - Exactly-once semantics, in-memory, distributed processing

- Cloud ELT
  - elastic scalability
  - massively parallel processing jobs
  - ability of routinely start/stop jobs fast
  - horizontal/vertical autoscaling
  - run serverless pipelines
  - dynamic work rebalancing
  - flexible resource scheduling



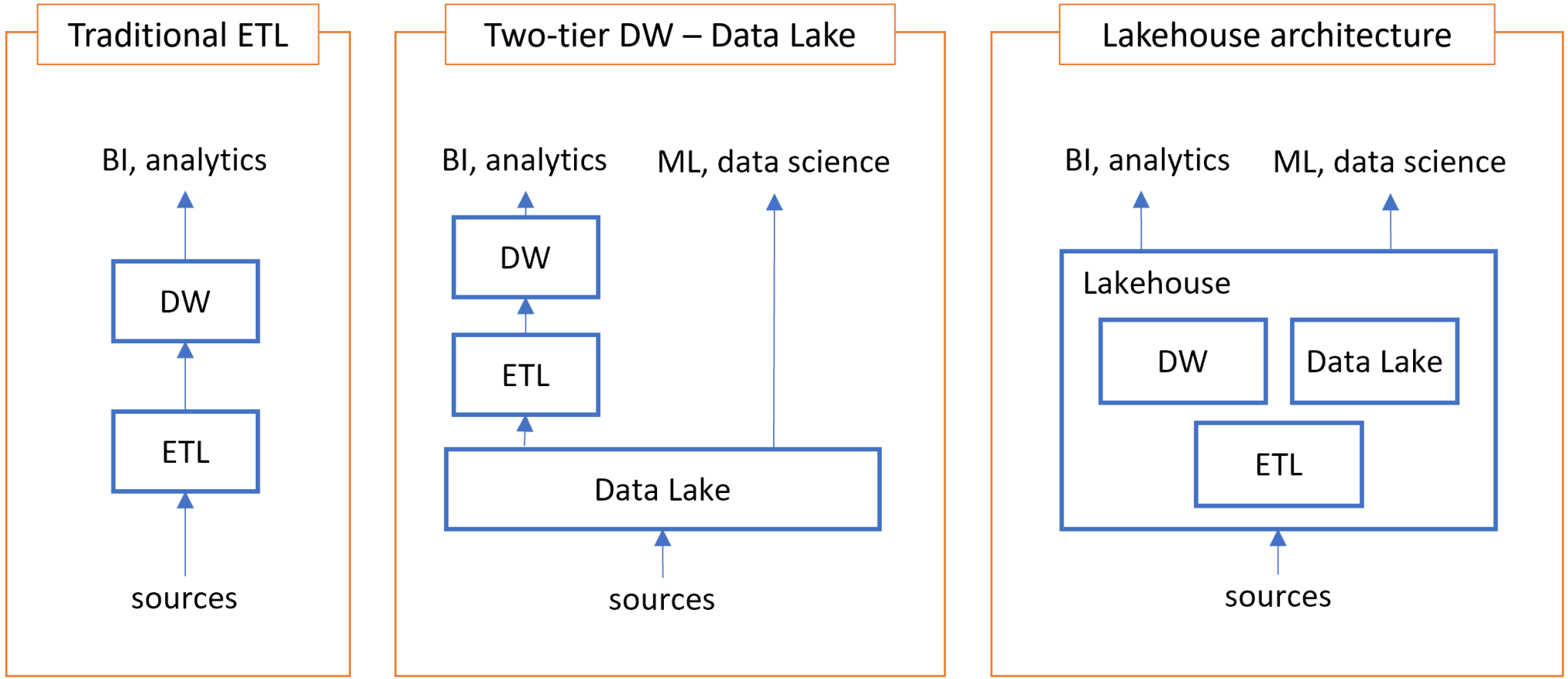
Trends in ETL processing

- Reverse ETL
  - Operational stores have isolated views of the domain
  - DW has a global view
  - Treat DW as an operational store that pushes insights back into the data sources



  - Challenges
    - schema validation,
    - efficient sync, low overhead at the sources, optimized pipeline
    - accuracy, consistency, privacy

Trends in ETL infrastructure



Alternatives to ETL

- Hybrid OLTP – OLAP or HTAP
  - Running transactional processing and scalable analytics on the same database
  - Accommodate very different workloads
    - operational (many small trnx, many updates)
    - analytical (complex, long running, resource demanding queries)
  - Hybrid column and row store setups, multi-version concurrency control

- In-situ processing
  - Avoid ETL, while still offering the features set of databases
  - Raw data files as a first-class citizen fully integrated with the query engine
  - Flexible caching and adaptive indexing to keep positional information and provide efficient access to raw files

ETL – the future

Next gen ETL – challenges (take #2)

- New ETL pipelines
  - Modern business intelligence, multimodal data processing, AI/ML ETL pipelines
- UDF-fueled in-engine ETL
  - Impedance mismatch between UDF and SQL is no more
- Learning ETL
  - Exploit learning techniques toward self-managed ETL
- Privacy preserving ETL
  - So far, focus on data protection and security (CCPA, HIPAA, GDPR, etc.)
  - Next: anonymization, differential privacy, homomorphic encryption, secure multi-party computation
- Personal ETL
  - Self-service data preparation
- ML pipelines as ETL
  - Data exploration, data discovery, feature engineering, observability, ML model auditing

Conclusions

Conclusions

- ETL technology
  - The cornerstone of business intelligence, decision making, and data analytics for over 25 years
  - Initial focus on design and optimization
  - Evolved to other forms: ELT, streaming, cloud, reverse
  - Evolving infrastructure: DW, Data lakes, Lakehouses, Multi-engine environments

- Our take
  - The ETL technology will remain relevant as long as it adapts to the modern business needs and data technology advancements

- Big THANKS to
  - The Test-of-Time award committee
  - The large and strong DOLAP community
  - Our many colleagues in this 20-year journey in the ETL-land and beyond

Big Thanks to our colleagues in this 20-year journey in the ETL-land and beyond

- A. Abelló
- E. Baikousi
- M. Castellanos
- J. Darmont
- U. Dayal
- A. Deligiannakis
- P. Georgantas
- N. Giatrakos
- M. Golfarelli
- A. Gounaris
- C. Gupta
- M. Hsu
- P. Jovanovic
- A. Karagiannis
- A. Karakasidis
- N. Karayannidis
- G. Kougka
- S. Lujan-Mora
- P. Manousis
- S. Nadal
- G. Papastefanatos
- T.B. Pedersen
- E. Pitoura
- N. Polyzotis
- O. Romero
- T. Sellis
- D. Skoutas
- M. Terrovitis
- D. Theodoratos
- J. Trujillo
- A. Tsois
- V. Tziovara
- Z. Vagena
- Y. Vassiliou
- K. Wilkinson
- A. Zarras