

# 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



# Agenda

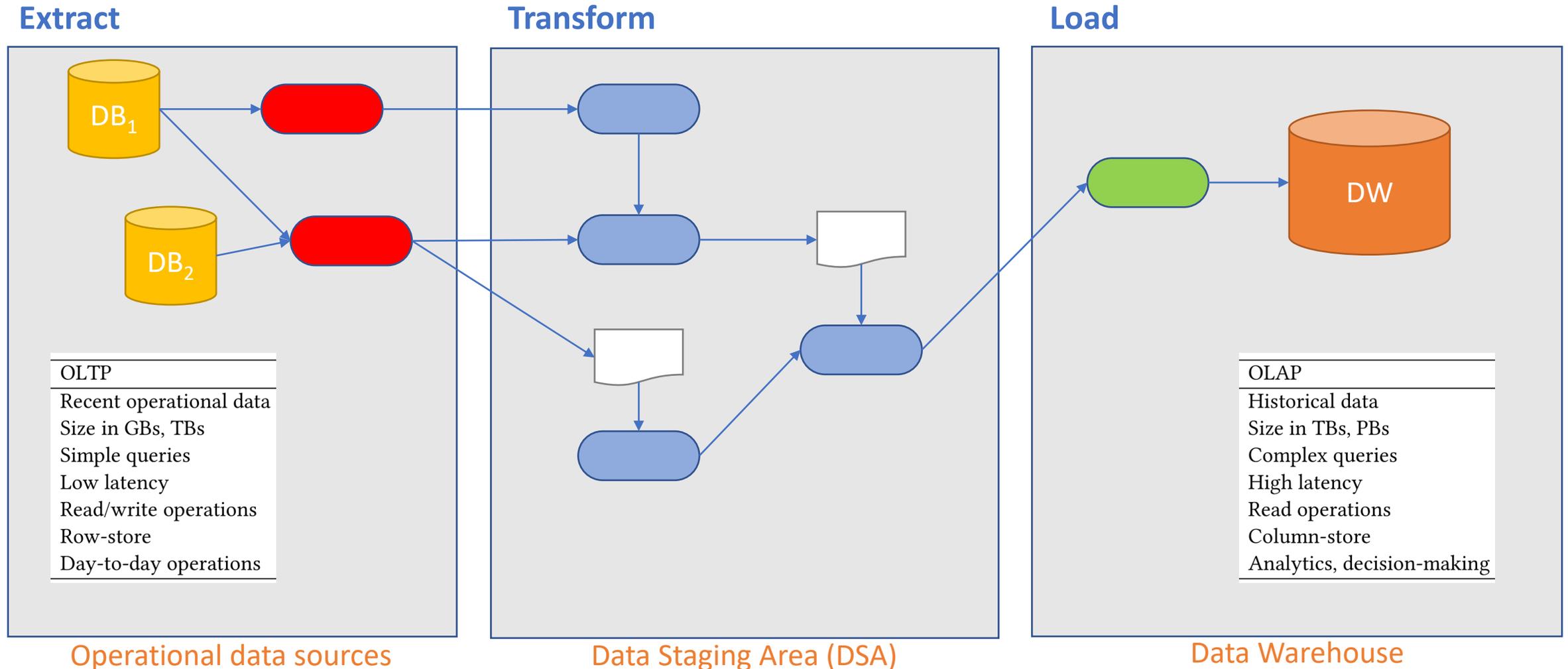
---

- 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

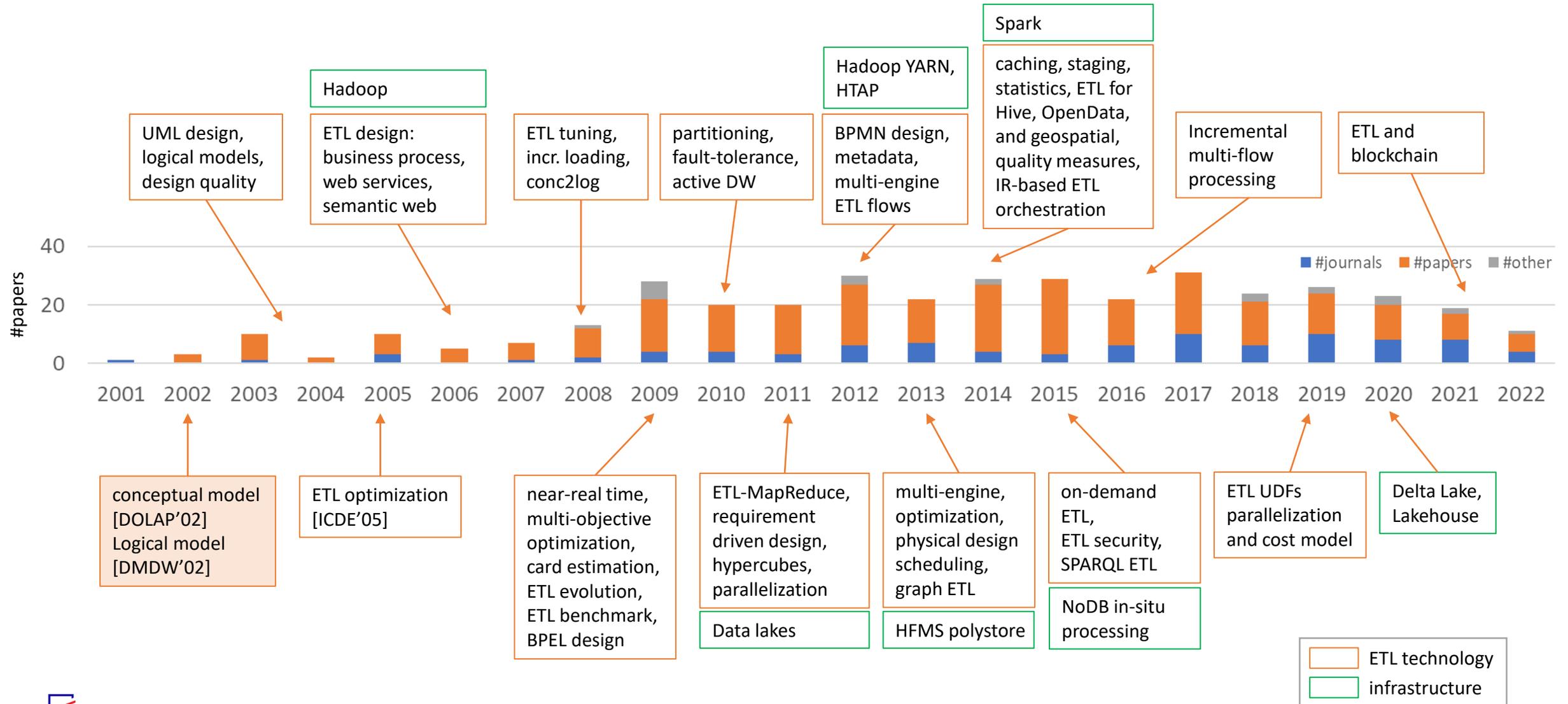
- Efficient extraction of delta values
- Schema transformations
  - data mappers, pivot/unpivot
- Data cleansing transformations
- Lineage of data transformations
- Efficient resumption of interrupted data flows
- Change-table techniques for incremental view maintenance
- Efficient data cubes
- Cardinality 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      avg #pubs/year  
 ▪ 270 papers              01-10: 10  
 ▪ 100 articles              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

The screenshot shows a search results page for 'ETL\$' on the DBLP database. The search query is 'ETL\$' and the results are filtered to show 394 matches. The results are displayed in a list format, with the first few entries highlighted. An orange box highlights the first entry: 'An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM'. Another orange box highlights the second entry: 'Dynamic multi-variant relational scheme-based intelligent ETL framework for healthcare management'. A third orange box highlights the third entry: 'Managing vulnerabilities during the development of a secure ETL processes'. A fourth orange box highlights the fourth entry: 'Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration'. A fifth orange box highlights the fifth entry: 'Distributed real-time ETL architecture for unstructured big data'. A sixth orange box highlights the sixth entry: 'High-level ETL for semantic data warehouses'. A seventh orange box highlights the seventh entry: 'A Service-Oriented Framework for ETL Implementation'. An eighth orange box highlights the eighth entry: 'An ETL strategy for integrating the LA Referencia platform and VIVO for the Brazilian CRIS'. A ninth orange box highlights the ninth entry: 'ETL Processes for Integrating Healthcare Data, Tools and Architecture Patterns'. A tenth orange box highlights the tenth entry: 'ETL Processes for Integrating Healthcare Data, Tools and Architecture Patterns'. A bar chart shows the distribution of publications over time, with a peak in 2022. A 'Refine list' on the right side of the page shows the number of publications for each author and venue. The 'refine by author' list includes: Alkis Simitis (37), Orlando Belo (27), Bruno Oliveira (23), Panos Vassiliadis (22), Christian Thomsen (11), Torben Bach Pedersen (11), Alberto Abelló (10), Faiez Gargouri (10), Xiufeng Liu (9), Stefan DeBloch (9), and 755 more options. The 'refine by venue' list includes: LNCS (61), DOLAP (19), CEUR Workshop Proceedings (18), CCIS (17), ADBIS (14), DaWaK (12), CoRR (11), ICEIS (8), LNBIP (8), ICDE (7), and 219 more options. The 'refine by type' list includes: Conference and Workshop Papers (276), Journal Articles (95), Informal Publications (11), Reference Works (7), Books and Theses (2), Parts in Books or Collections (2), and Data and Artifacts (1).

computer science bibliography  
Stop the war!  
ETL\$

Search dblp for Publications  
powered by CompleteSearch, courtesy of Hannah Bast, University of Freiburg

> Home > Search

Publication search results

found 394 matches

2023

Yuan Peng, Elisa Henke, Ines Reinecke, Michèle Zoch, Martin Sedlmayr, Franziska Bathelt:  
**An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM.** Int. J. Medical Informatics 169: 104925 (2023)

Vijayalakshmi Manickam, Minu Rajasekaran Indra:  
**Dynamic multi-variant relational scheme-based intelligent ETL framework for healthcare management.** Soft Comput. 27(1): 605-614 (2023)

2022

Salma Dammak, Faiza Ghozzi, Asma Sellami, Faïez Gargouri:  
**Managing vulnerabilities during the development of a secure ETL processes.** Int. J. Inf. Comput. Secur. 18(1/2): 75-104 (2022)

Yue Yu, Nansu Zong, Andrew Wen, Sijia Liu, Daniel J. Stone, David Knaack, Alanna M. Chamberlain, Emily R. Pfaff, Davera Gabriel, Christopher G. Chute, Nilay Shah, Guoqian Jiang:  
**Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration.** J. Biomed. Informatics 127: 104002 (2022)

Erum Mehmood, Tayyaba Anees:  
**Distributed real-time ETL architecture for unstructured big data.** Knowl. Inf. Syst. 64(12): 3419-3445 (2022)

Rudra Pratap Deb Nath, Oscar Romero, Torben Bach Pedersen, Simeon Krieger, Simeon Krieger:  
**High-level ETL for semantic data warehouses.** Semantic Web 13(1): 1-15 (2022)

Bruno Oliveira, Mário Leite, Óscar Oliveira, Orlando Belo, Panos Vassiliadis:  
**A Service-Oriented Framework for ETL Implementation.** Proceedings of the 2022 ACM Conference on Health, Information Systems and Informatics (CHIIS '22): 1-10 (2022)

Vivian S. Silva, Lautaro Matas, Tales Moreira, Washington de Carvalho Segundo:  
**An ETL strategy for integrating the LA Referencia platform and VIVO for the Brazilian CRIS.** CRIS 2022: 111-117 (2022)

Ka Yung Cheng, Santiago Pazmino, Björn Schreiweis:  
**ETL Processes for Integrating Healthcare Data, Tools and Architecture Patterns.** Health 2022: 1-10 (2022)

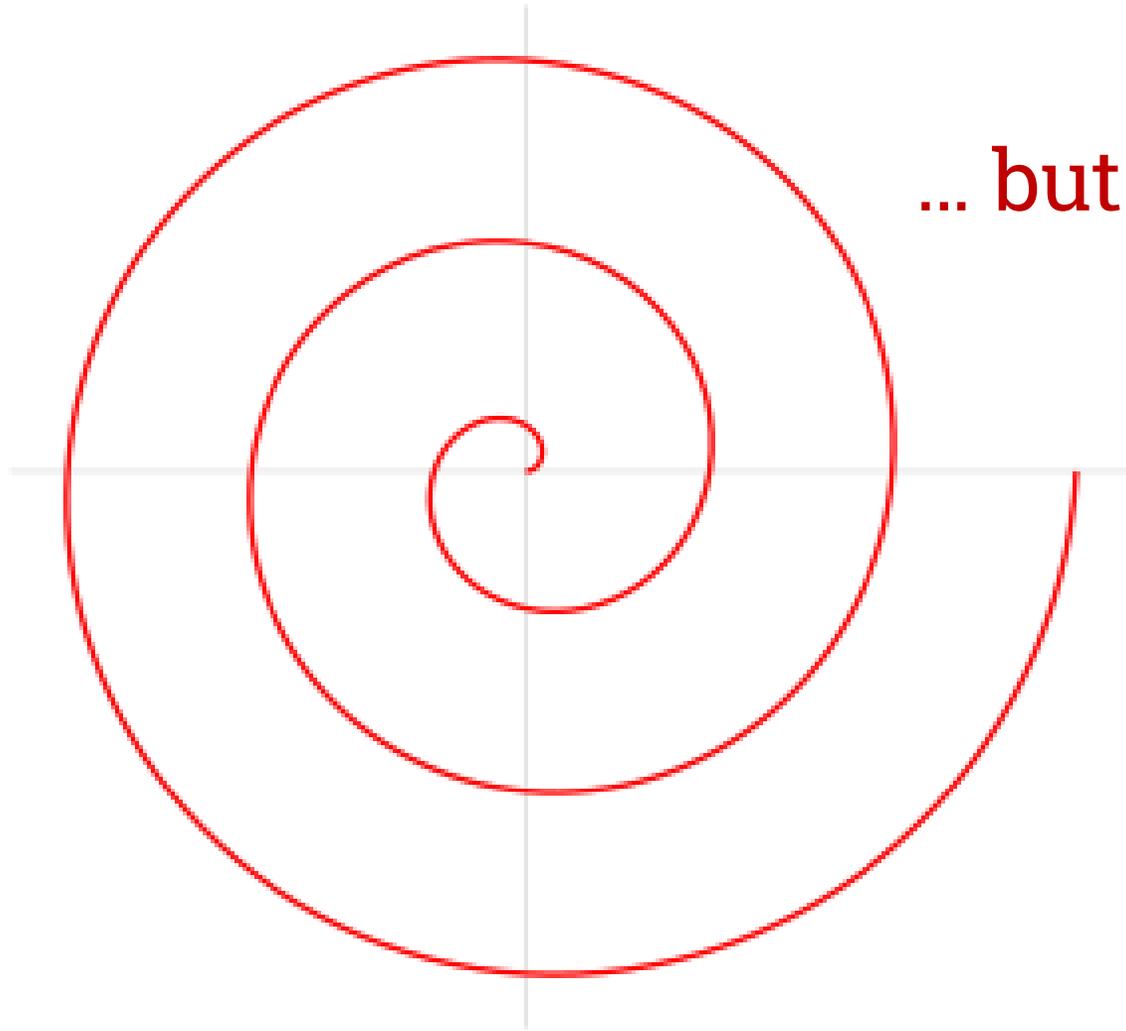
refine by author  
Alkis Simitis (37)  
Orlando Belo (27)  
Bruno Oliveira (23)  
Panos Vassiliadis (22)  
Christian Thomsen (11)  
Torben Bach Pedersen (11)  
Alberto Abelló (10)  
Faiez Gargouri (10)  
Xiufeng Liu (9)  
Stefan DeBloch (9)  
755 more options

refine by venue  
LNCS (61)  
DOLAP (19)  
CEUR Workshop Proceedings (18)  
CCIS (17)  
ADBIS (14)  
DaWaK (12)  
CoRR (11)  
ICEIS (8)  
LNBIP (8)  
ICDE (7)  
219 more options

refine by type  
Conference and Workshop Papers (276)  
Journal Articles (95)  
Informal Publications (11)  
Reference Works (7)  
Books and Theses (2)  
Parts in Books or Collections (2)  
Data and Artifacts (1)

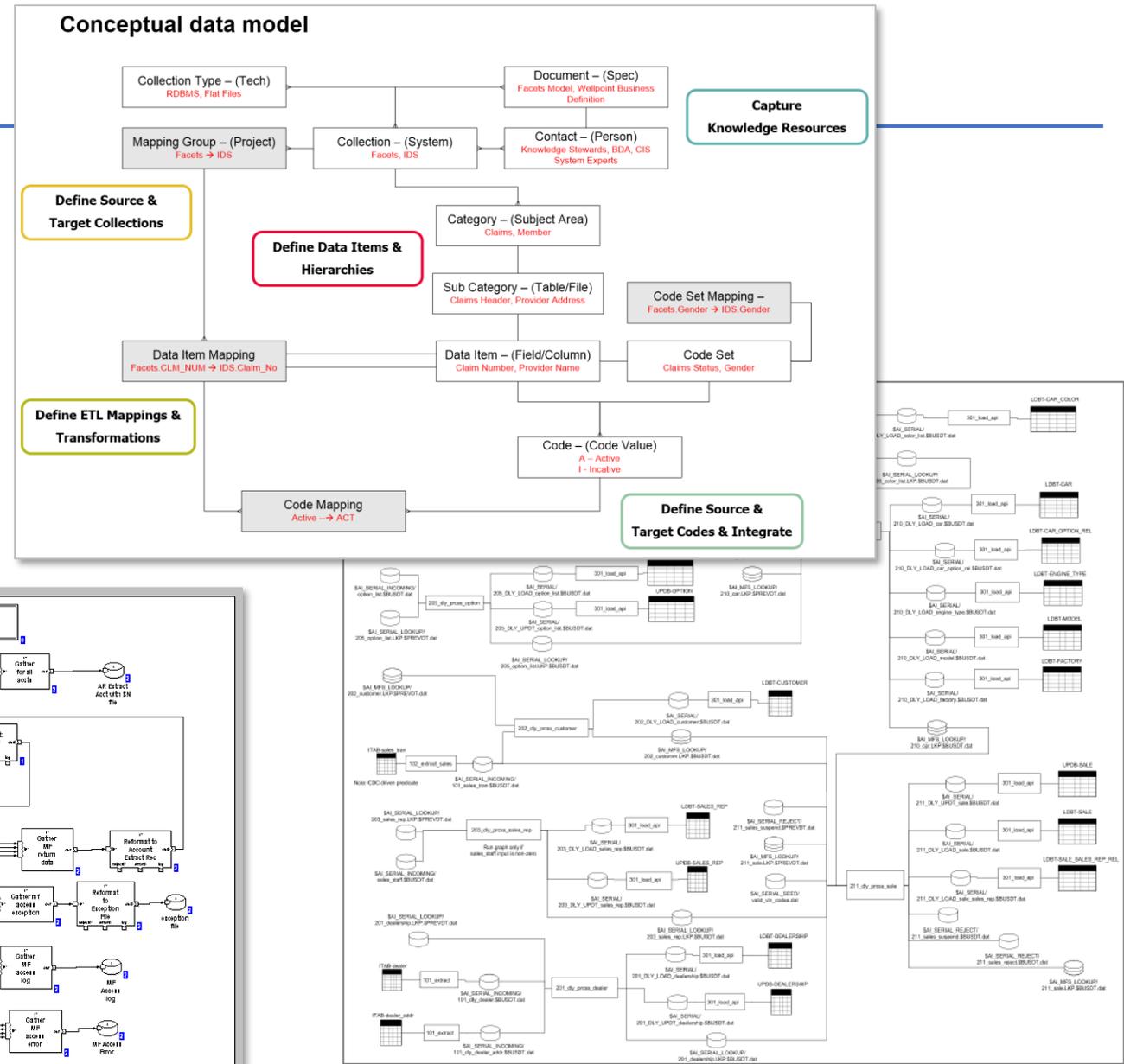
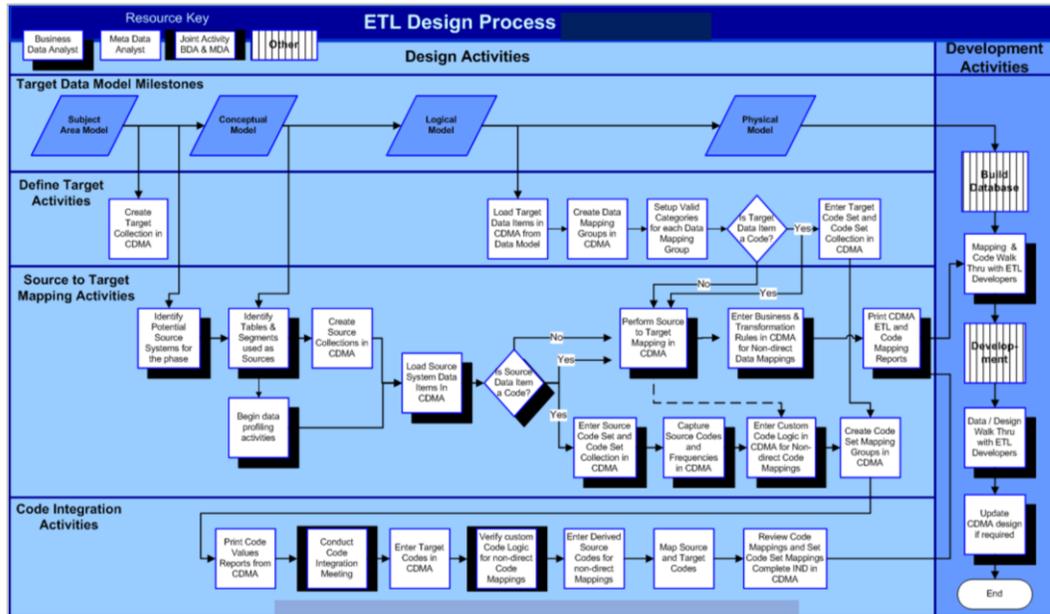
19 in DOLAP  
2nd venue for ETL

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

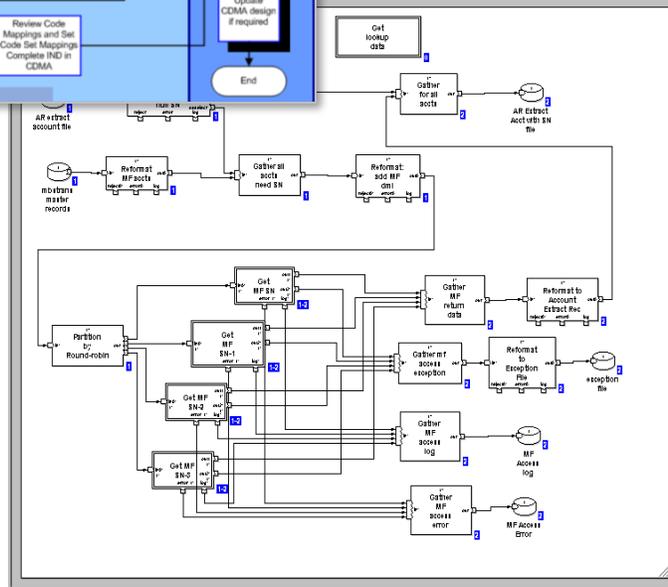


... but how did we start? ...

# 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 Tools (current as of 2003)	
ETL Tools	Vendor
1. ActaWorks	Acta Technologies
2. Amadea	ISoft
3. ASG-XPATH	Allen Systems Group
4. AT Sigma W-Import	
5. AutoImport	
6. Automatic Data Warehouse	
7. Blue Data Miner	
8. Catalyst	
9. CDB/Superload	
10. Cerebellum Portal Integrat	
11. Checkmate	
12. Chyfo	
13. CMS TextMap	
14. Compleo	
15. Content Connect	
16. Conversions Plus	
17. Convert /IDMS-DB, Conv	
18. Copy Manager	
19. CoSORT	
20. CrossXpress	
21. Cubeware Importer	
22. Cyklop	
23. Data Cycle	
24. Data Dragon	
25. Data Exchange	
26. Data EXTRACTor	
27. Data Flow Manager	
28. Data Junction, Content Ex	
29. Data Manager	
30. Data Mapper	
31. Data Migration Tools	
32. Data Migrator for SAP, Pe	
33. Data Propagation System	
34. Data Warehouse Tools	
35. Data³	
36. DataBlaster 2	
37. DataBrix Data Manager	
38. DataConvert	
39. DataDigger	
40. DataExchanger SRV	
41. Datagration	
42. DataImport	
43. DataLever	
44. DataLoad	
45. DataManager	
46. DataMIG	
47. DataMiner	
48. DataMirror Constellar	
49. DataMirror Transform	
50. DataPipe	
51. DataProF	
52. DataPropagator	
53. DataProvider	
54. DataPump for SAP R/	
55. DataStage XE	
56. DataSuite	
57. Datawhere	
58. DataX	
59. DataXPRESS	
60. DB/Access	
61. DBMS/Copy	
62. DBridge	
63. DEAP I	
64. DecisionBase	
65. DecisionStream	
66. DECISIVE Advantage	
67. Departmental Suite 20	
68. DETAIL	
69. Distribution Agent for	
70. DocuAnalyzer	
71. DQ Now	
72. DQtransform	
73. DT/Studio	
74. DTS	
75. eCartography	
76. eIntegration Suite	
77. Environment Manager	
78. e-Sense Gather	
79. ETL Extract	
80. ETL Engine	
81. ETL Manager	
82. eWorker Portal, eWork	
83. eXadas	
84. e-zMigrate	
85. EZ-Pickin's	
86. FastCopy	
87. File-AID/Express	
88. FileSpeed	
89. Formware	
90. FOXTROT	
91. Fusion FTMS	
92. Gate 1	
93. Génio	
94. Gladstone Conversion Pac	
95. GoHunter	
96. Graphical Performance Se	
97. Harvester	
98. HIREL	
99. Hummingbird ETL	
100. iManageData	
101. iMergece	
102. Influx	
103. InfoLink/400	
104. InfoManager	
105. InfoRefiner, InfoTranspor	
106. InfoPump	
107. Information Discovery Pla	
108. Information Logistics Net	
109. InformEnt	
110. InScaner	
111. InScribe	
112. InTouch/2000	
113. ISIE	
114. John Henry	
115. KM.Studio	
116. LiveTransfer	
117. LOADPLUS	
118. Mainframe Data Engine	
119. MarketDrive	
120. MDFA	
121. Mercator	
122. Meta Integration Works	
123. MetaSuite	
124. MetaTrans	
125. MinePoint	
126. MineWorks/400	
127. MITS	
128. Monarch	
129. Mozart	
130. mpower	
131. MRE	
132. NatQuery	
133. netConvert	
134. NGS-IQ	
135. NSX Data Stager	
136. ODBCFace	
137. OLAP Data Migrator	
138. OmniReplicator	
139. OpalisRendezVous	
140. Open Exchange	
141. OpenMigrator	
142. OpenWizard Professional	
143. OptiLoad	
144. Oracle Warehouse Builder	
145. Orchestrate	
146. Outbound	FireSign Computer Company
147. Parse-O-Matic	Pinnacle Software <a href="http://www.parse-o-matic.com/">http://www.parse-o-matic.com/</a>
148. ParseRat	Guy Software <a href="http://www.guysoftware.com/parserrat.htm">http://www.guysoftware.com/parserrat.htm</a>
149. pcMainframe	cfSOFTWARE <a href="http://www.cfsoftware.com/">http://www.cfsoftware.com/</a>
150. PinnPoint Plus	Pinnacle Decision Systems <a href="http://www.pinnpoint.com/products/index.html">http://www.pinnpoint.com/products/index.html</a>
151. PL/Loader	Hanlon Consulting
152. PointOut	mSE GmbH
153. Power*Loader Suite	SQL Power Group
154. PowerDesigner WarehouseArchitect	Powersoft
155. PowerMart	Informatica <a href="http://www.informatica.com">http://www.informatica.com</a>
156. PowerStage	Sybase <a href="http://www.sybase.com/">http://www.sybase.com/</a>
157. Rapid Data	Open Universal Software
158. Relational DataBridge	Liant Software Corporation
159. Relational Tools	Princeton Softech <a href="http://www.princetonsofttech.com/products/relationaltools.htm">http://www.princetonsofttech.com/products/relationaltools.htm</a>
160. ReTarGet	Tominy
161. Rodin	Coglin Mill Pty Ltd. <a href="http://www.coglinmill.com/index.html">http://www.coglinmill.com/index.html</a>
162. Roll-Up	Ironbridge Software
163. Sagent Solution	Sagent Technology, Inc. <a href="http://www.sagenttech.com/us/products/index.asp">http://www.sagenttech.com/us/products/index.asp</a>
164. SAS Warehouse Administrator	SAS Institute <a href="http://www.sas.com/technologies/dw/etl/index.html">http://www.sas.com/technologies/dw/etl/index.html</a>
165. Scheme Advanced	Appligator.com <a href="http://www.appligator.com/">http://www.appligator.com/</a>
166. Scribe Integrate	Scribe Software Corporation
167. Scriptoria	Bunker Hill <a href="http://www.bunkerhill.com/">http://www.bunkerhill.com/</a>
168. SERdistiller	SER Solutions <a href="http://www.sersolutions.com/product_showcase/serdistiller/index.html">http://www.sersolutions.com/product_showcase/serdistiller/index.html</a>
169. Signiant	Signiant
170. SIPINA PRO	Diagnos <a href="http://www.diagnos.ca/en/index1.html">http://www.diagnos.ca/en/index1.html</a> or <a href="http://eric.univ-lyon2.fr/~signia.html">http://eric.univ-lyon2.fr/~signia.html</a>
171. SpeedLoader	Benchmark Consulting <a href="http://www.drcbenchmark.com/products_internal.htm#speedloa">http://www.drcbenchmark.com/products_internal.htm#speedloa</a>
172. SRTransport	Schema Research Corp. <a href="http://www.schemaresearch.com/products/srtransport/index.html">http://www.schemaresearch.com/products/srtransport/index.html</a>
173. StarQuest Data Replicator	StarQuest Software
174. StarTools	StarQuest
175. Stat/Transfer	Circle Systems <a href="http://www.stattransfer.com/">http://www.stattransfer.com/</a>
176. Strategy	SPSS
177. Sunopsis	Sunopsis <a href="http://www.sunopsis.com/corporate/us/products/sunopsisv3/default.htm">http://www.sunopsis.com/corporate/us/products/sunopsisv3/default.htm</a>
178. SyncSort Unix	Syncsort <a href="http://www.syncsort.com/sortinfo.htm">http://www.syncsort.com/sortinfo.htm</a>
179. TableTrans	PPD Informatics
180. Text Agent	Tasc, Inc.
181. TextPipe	Crystal Software Australia
182. TextProc2000	LVRA <a href="http://www.textproc.com/">http://www.textproc.com/</a>
183. Texttractor	Textkernel
184. Tilion	Tilion
185. Transporter Fountain	Digital Fountain
186. TransportIT	Computer Associates
187. ViewShark	infoShark
188. Vignette Business Integration Studio	Vignette
189. Visual Warehouse	IBM <a href="http://www-306.ibm.com/software/data/vw/">http://www-306.ibm.com/software/data/vw/</a>
190. Volantia	Volantia <a href="http://www.volantia.arachsys.com/home.htm">http://www.volantia.arachsys.com/home.htm</a>
191. vTag Web	Connotate Technologies <a href="http://www.connotate.com/">http://www.connotate.com/</a>
192. Waha	Beacon Information Technology <a href="http://www.beacon-it.co.jp/e_docs/index.html">http://www.beacon-it.co.jp/e_docs/index.html</a>
193. Warehouse	Taurus Software <a href="http://www.taurus.com/">http://www.taurus.com/</a>
194. Warehouse Executive	Ardent Software <a href="http://www.ardentsoftware.com/">http://www.ardentsoftware.com/</a>
195. Warehouse Plus	eNvy Systems <a href="http://envysys.com/">http://envysys.com/</a>
196. Warehouse Workbench	Systemfabrik <a href="http://www.systemfabrik.com">http://www.systemfabrik.com</a>
197. Web Automation	webMethods
198. Web Data Kit	LOTONtech <a href="http://www.lotontech.com/">http://www.lotontech.com/</a>
199. Web Mining	Blossom Software
200. Web Replicator	Media Consulting
201. WebFOCUS ETL Manager	Information Builders, Inc. <a href="http://www.informationbuilders.com/products/webfocus/cm_fac">http://www.informationbuilders.com/products/webfocus/cm_fac</a>
202. WebQL	Caesius Software <a href="http://www.q12.com/">http://www.q12.com/</a>
203. WhizBang! Extraction Library	WhizBang! Labs
204. Wizport	Turning Point
205. Xentis	GrayMatter Software Corporation <a href="http://www.graysoft.com/GSCproducts.html">http://www.graysoft.com/GSCproducts.html</a>
206. XSB	XSB Inc. <a href="http://www.xsb.com/">http://www.xsb.com/</a>

# 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

5: 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

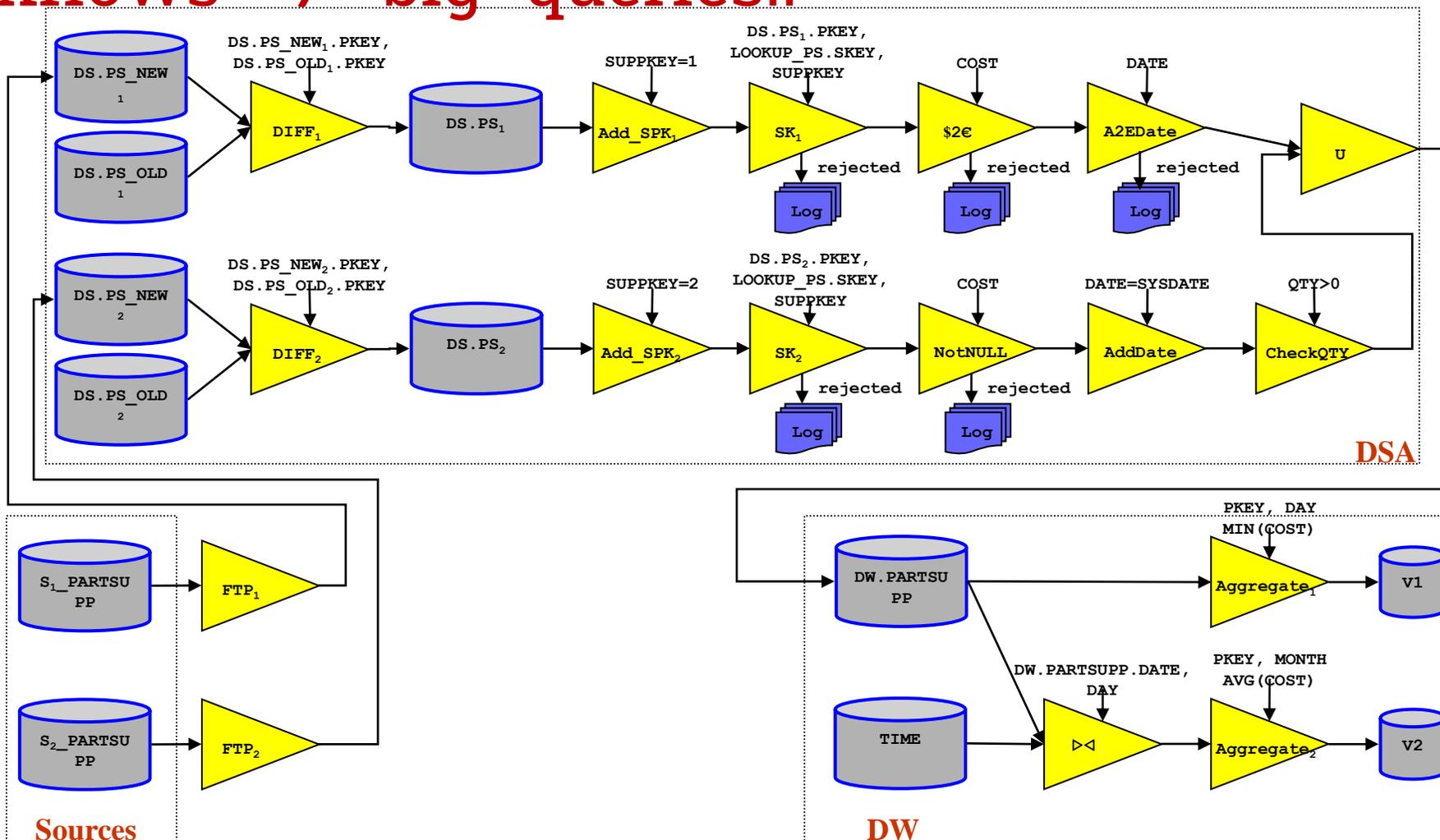
Maybe the differential is in a slightly richer set of operations

12: Detailed comments to authors:

see above

# Data Warehouses $\neq$ collections of materialized views!!

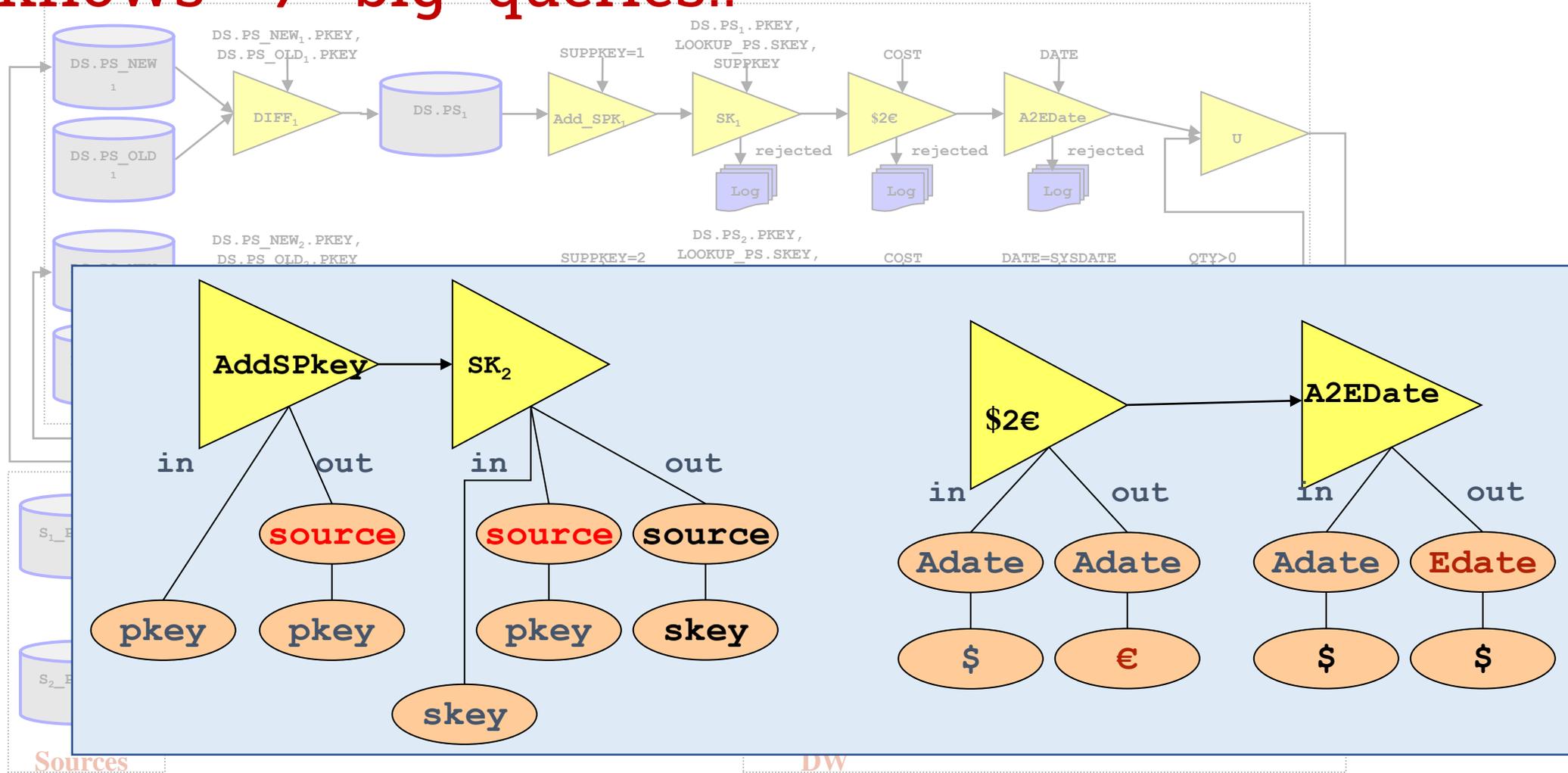
## ETL workflows $\neq$ "big" queries!!



This has always been the vision

# Data Warehouses ≠ collections of materialized views!!

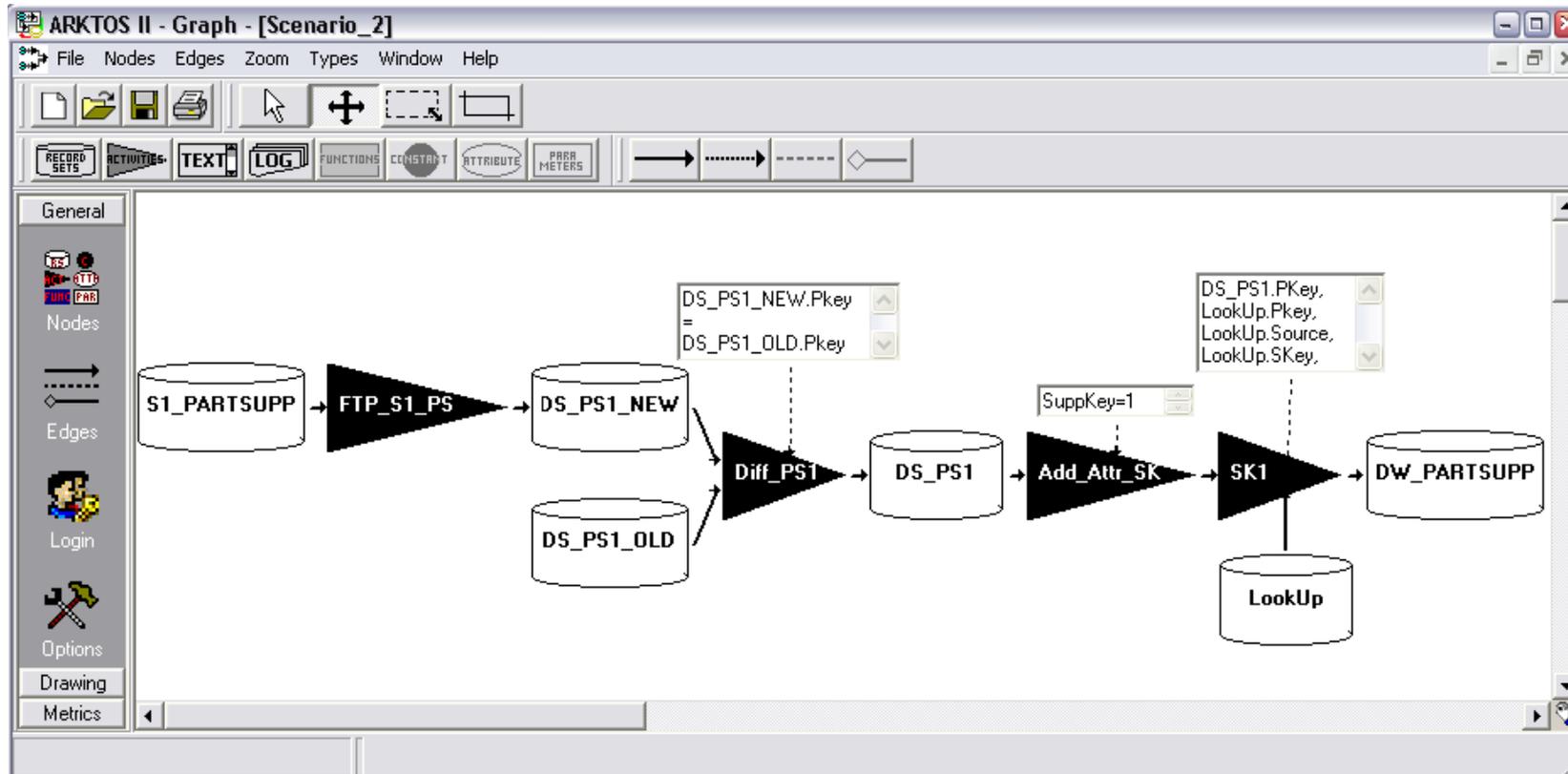
## ETL workflows ≠ "big" queries!!



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]

Omar Boussard, Fatima Bentayeb, Jerome Darmont:  
**A multi-agent system-based ETL approach for complex data.** ISPE CE 2003: 49-52

Alkis Simitsis:  
**Modeling and managing ETL processes.** VLDB PhD Workshop 2003

**2002**

Panos Vassiliadis, Alkis Simitsis, Spiros Skiadopoulos:  
**On the Logical Modeling of ETL Processes.** CAISE 2002: 782-786

Panos Vassiliadis, Alkis Simitsis, Spiros Skiadopoulos:  
**Modeling ETL activities as graphs.** DMDW 2002: 52-61

Panos Vassiliadis, Alkis Simitsis, Spiros Skiadopoulos:  
**Conceptual modeling for ETL processes.** DOLAP 2002: 14-21

**2001**

Panos Vassiliadis, Zografoula Vagena, Spiros Skiadopoulos, Nikos Karayannidis, Timos K. Sellis:  
**ARKTOS: towards the modeling, design, control and execution of ETL processes.** Inf. Syst. 26(8): 537-561 (2001)

Akihiko Nagakubo, Yasuo Kuniyoshi, Gordon Cheng:  
**ETL-Humanoid-A high-performance full body humanoid system for versatile actions.** IROS 2001: 1087-1092

Yasuo Kuniyoshi, Gordon Cheng, Akihiko Nagakubo:  
**ETL-Humanoid: A Research Vehicle for Open-Ended Action Imitation.** ISRR 2001: 67-82

**1999**

Yasuhiro Nakamura, Akio Fukushima, Yasuhiko Sakamoto, Tadashi Endo, Greig W. Small:  
**A multifrequency quadrature bridge for realization of the capacitance standard at ETL.** IEEE Trans. Instrum. Meas. 48(2): 351-355 (1999)

**1990**

Kazuhiko Yamamoto:  
**Future directions in computer vision and image understanding: ETL perspectives.** ICPR (1) 1990: 32-37

Kazuyo Tanaka, Satoru Hayamizu, Kozo Ohta:  
**The ETL speech database for speech analysis and recognition research.** ICSLP 1990

**1962**

S. Takahashi, H. Nishino, K. Yoshihiro, Kazuhiro Fuchi:  
**System Design of the ETL Mk-6 Computer.** IFIP Congress 1962: 690-693

**1958**

Ichiro Honda:  
**News [from Ichiro Honda on ETL research].** Mech. Transl. Comput. Linguistics 5(2): 49-50 (1958)

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

# Why a conceptual model?

the what

the how

- 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



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

# Conceptual modeling for ETL

---



# 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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DOI: 10.1145/590402.590401  
Copyright 2002 ACM 1-58113-590-4/02/0011...\$5.00.

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.

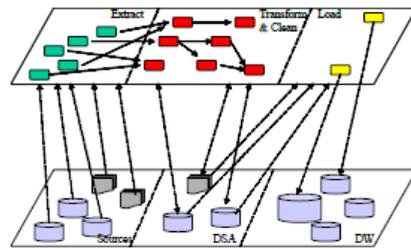


Figure 1. The environment of ETL processes

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 [ACM DOLAP 2002]

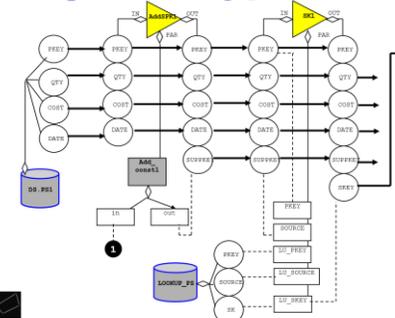
## Conceptual Modeling for ETL processes

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

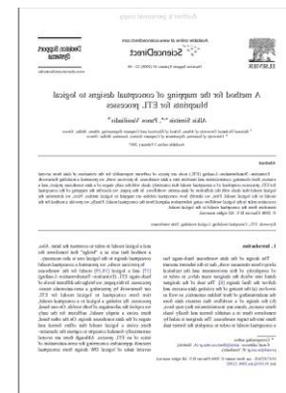


National Technical University of Athens  
KDBS Laboratory  
<http://www.dbnet.ece.ntua.gr>

## Graph Modeling [DMDW'02]

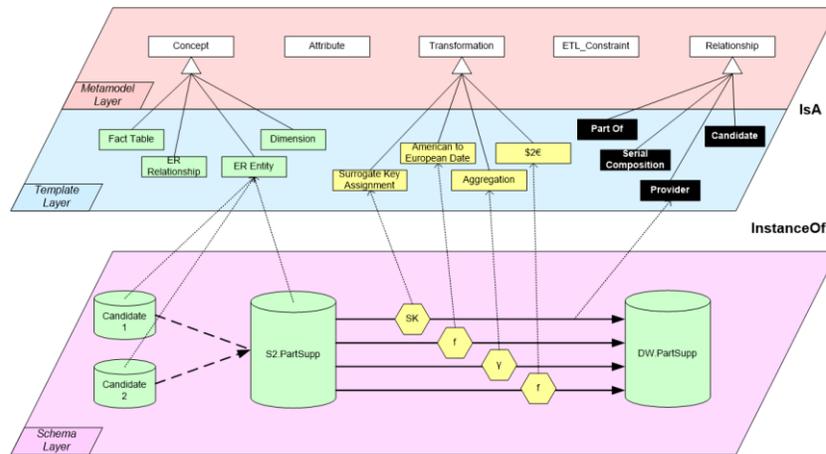


Vassiliadis, Simitsis, Skiadopoulos. On the Logical Modeling of ETL Processes. CAISE'02 Toronto, 2002





## Instantiation & Specialization Layers



Vassiliadis, Simitsis, Skiadopoulos - DOLAP'02

31

## Instantiation & Specialization Layers

### Filters

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

### Unary transformations

Push  
 Aggregation ( $\gamma$ )  
 Projection ( $\pi$ )  
 Function application ( $f$ )  
 Surrogate key assignment (SK)  
 Tuple normalization (N)  
 Tuple denormalization (DN)

### Binary transformations

Union (U)  
 Join ( $\bowtie$ )  
 Diff ( $\Delta$ )  
 Update Detection ( $\Delta_{UPD}$ )

### Composite transformations

Slowly changing dimension (Type 1,2,3) (SDC-1/2/3)  
 Format mismatch (FM)  
 Data type conversion (DTC)  
 Switch ( $\sigma^*$ )  
 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)

Vassiliadis, Simitsis, Skiadopoulos - DOLAP'02

33

# ETL – present times

---

# The analytics landscape

2012

The screenshot shows the website 'THE BIG DATA LANDSCAPE' with a navigation bar containing 'BIG DATA 100', 'RESEARCH', 'CONSULTING', 'ABOUT', and 'CONTACT'. Below the navigation bar is a search bar with 'Add your company | Log in'. The main content area is titled 'The Big Data Landscape' and is organized into three main sections: 'Apps', 'Infrastructure', and 'Technologies'. Each section contains several sub-categories with logos of companies in that space.

**Apps**

- Vertical Apps:** Atigeo, ellucian, MYRRIX, Placed, PRAXIS THE PRACTICE, Quantivo
- Ad / Media Apps:** bloomreach, Collective, DataXu, LuckyScribe, Media Science, Recorded Future, rocketHub, TURN
- Business Intelligence:** ACTIVIO, Autonomy, bime, birst, Business Objects, Chart.io, COGNOS, DOMO, GoodData, IBM, Jaspersoft, Microstrategy, pentaho
- Analytics And Visualization:** 1010data, alteryx, AYATA, centrifuge, CIRRO, ClearStory, Datameer, developer, KARMASHERE, metaLayer, OPERA, Palantir, panopticon, platforma, QlikView, RJMetrics, saffron, sas, +bleev, TIBCO, visual.ly
- Operational Intelligence:** loggly, splunk, sumologic
- Data As A Service:** DATASIFT, factual, FICO, GNIP, INRIX, kaggle, knoema, LexisNexis, LOGATE, SPACE CURVE

**Infrastructure**

- Analytics Infrastructure:** calpont, cloudera, DISTRAX, EXASOL, GREENPLUM, HADAPT, Hortonworks, INFOBRIGHT, kognitio, MAFR, PARACCEL, VERTICA
- Operational Infrastructure:** 10gen, Couchbase, MarkLogic, TERRACOTTA, VoltDB
- Infrastructure As A Service:** infochimps, MORTAR, Qu, bole
- Structured Databases:** IBM, DB2, SQL Server, MySQL, ORACLE, PostgreSQL, SYBASE

**Technologies**

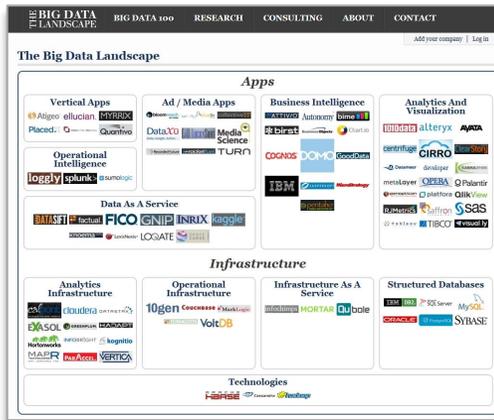
- HBASE, Cassandra, Hadoop

[src: <https://mattturck.com>]

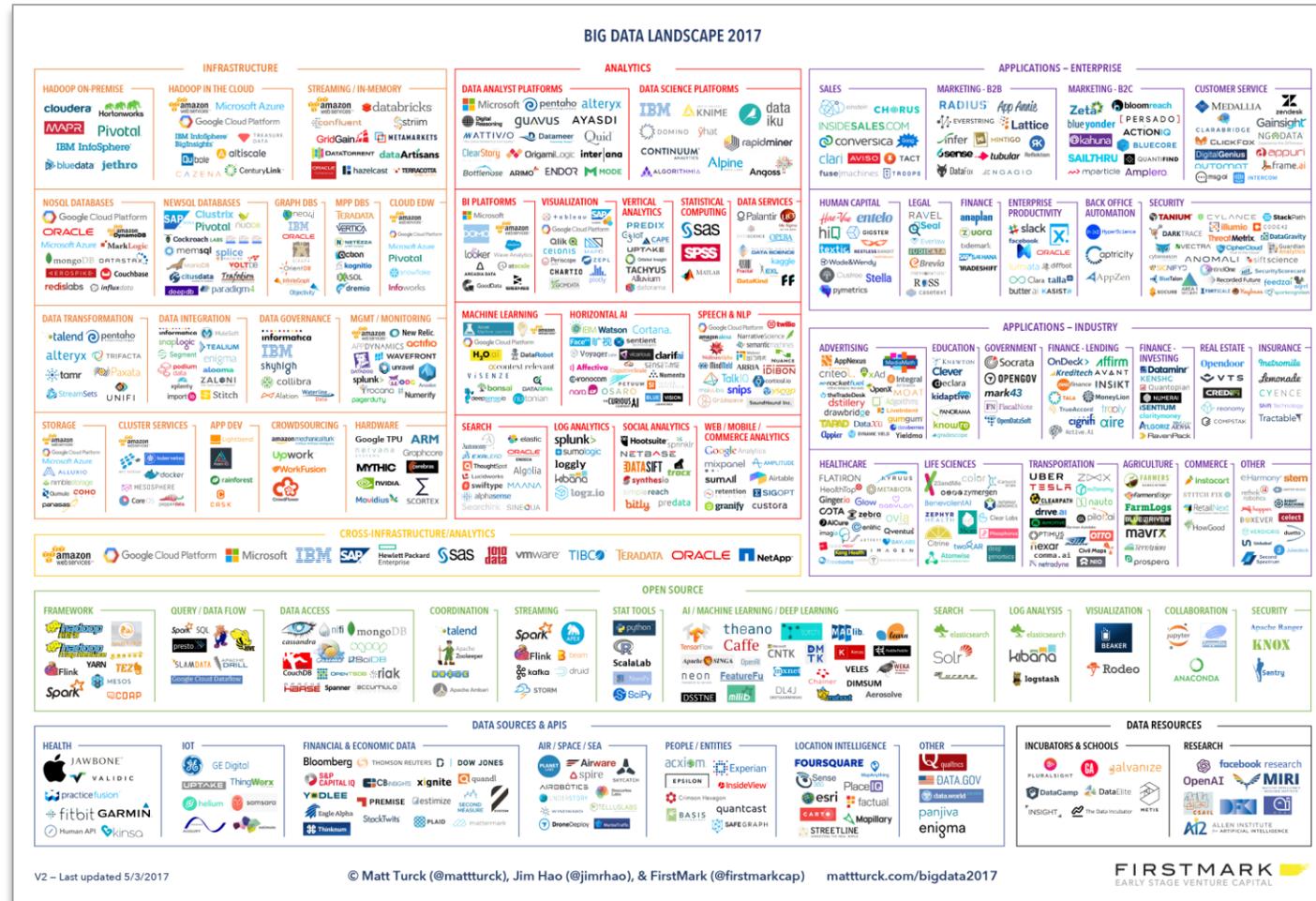
# The analytics landscape

2012

2017



[src: <https://mattturck.com>]



V2 - Last updated 5/3/2017

© Matt Turck (@mattturck), Jim Hao (@jimhao), & FirstMark (@firstmarkcap) mattturck.com/bigdata2017

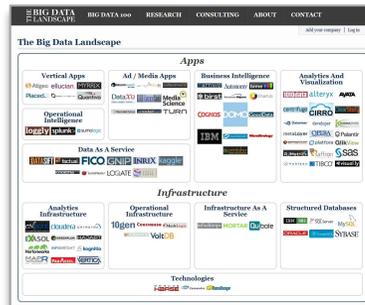
FIRSTMARK  
EARLY STAGE VENTURE CAPITAL

# The analytics landscape

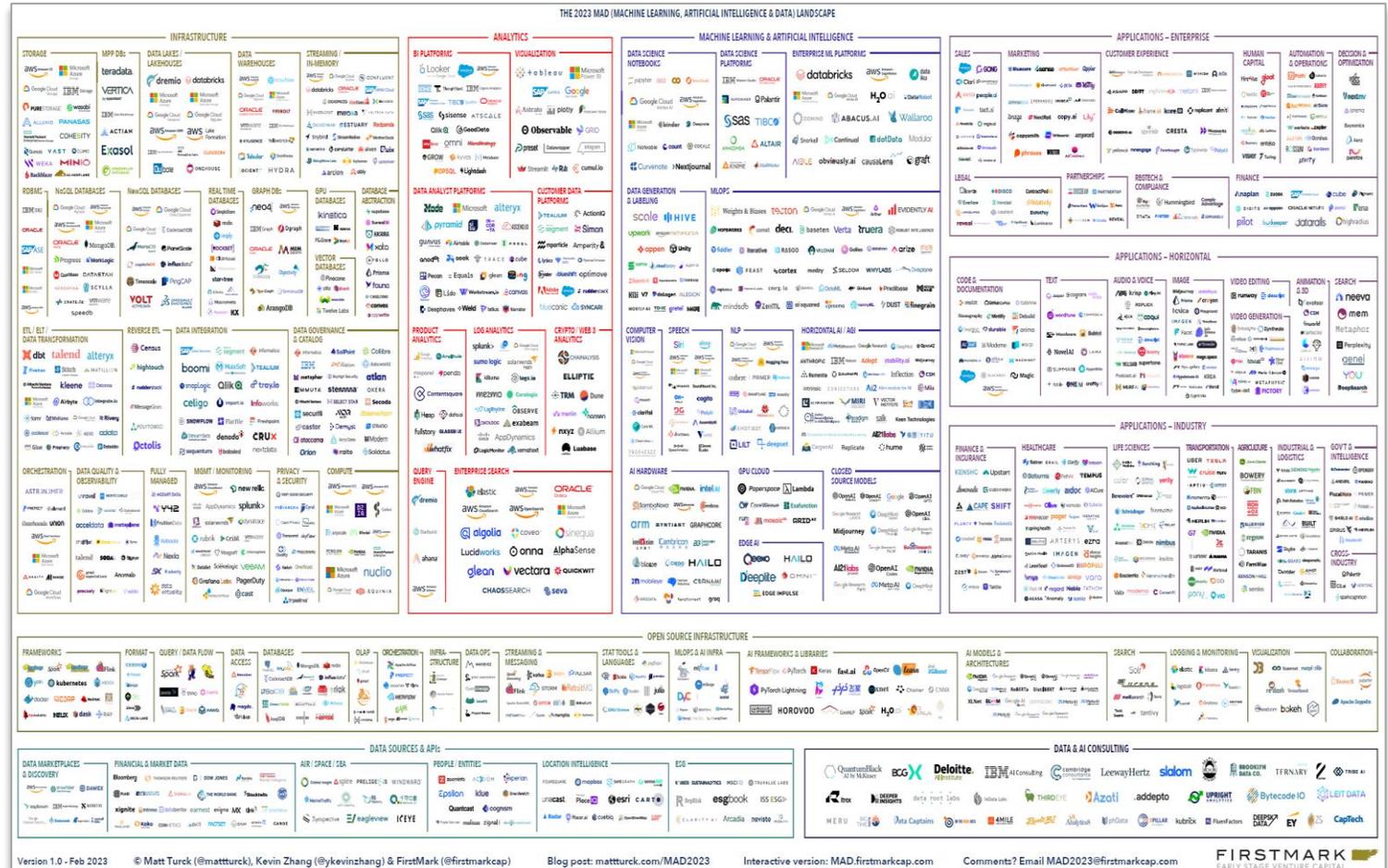
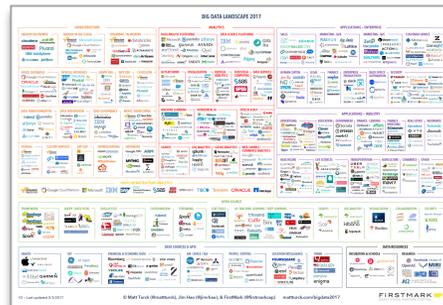
2012

2017

2023

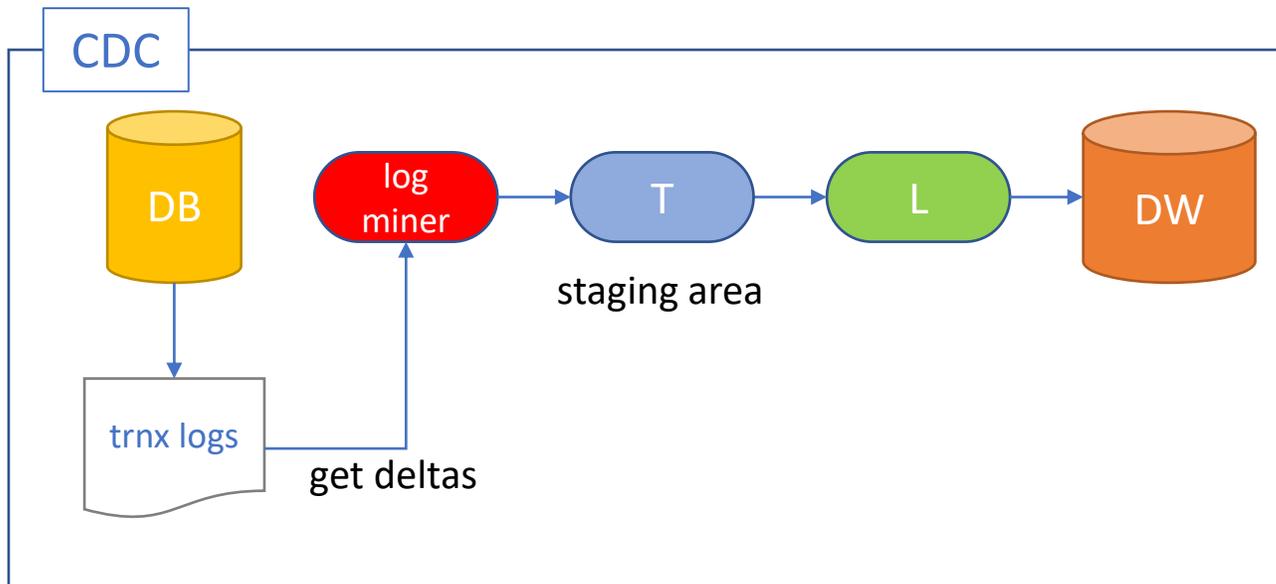
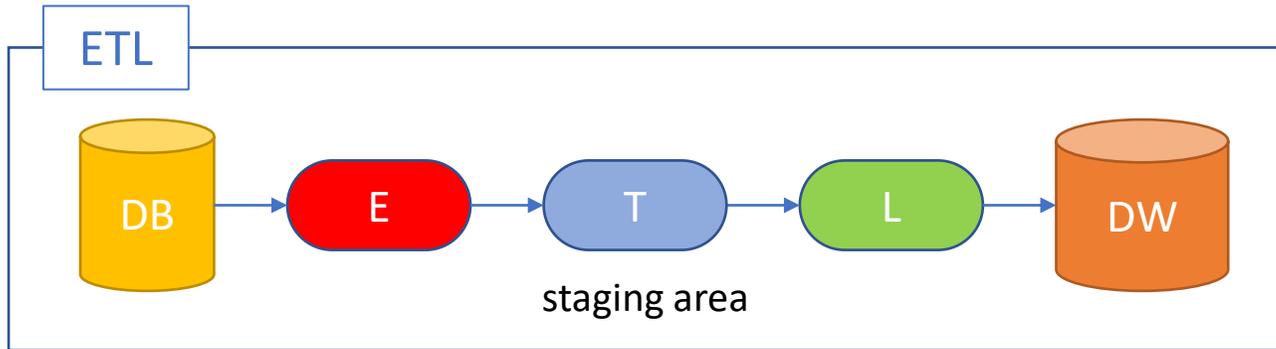


[src: <https://mattturck.com>]



Version 1.0 - Feb 2023 © Matt Turck (@mattturck), Kevin Zhang (@kyevinzhang) & FirstMark (@firstmarkcap) Blog post: [mattturck.com/MAD2023](https://mattturck.com/MAD2023) Interactive version: [MAD.firstmarkcap.com](https://MAD.firstmarkcap.com) Comments? Email [MAD2023@firstmarkcap.com](mailto:MAD2023@firstmarkcap.com)

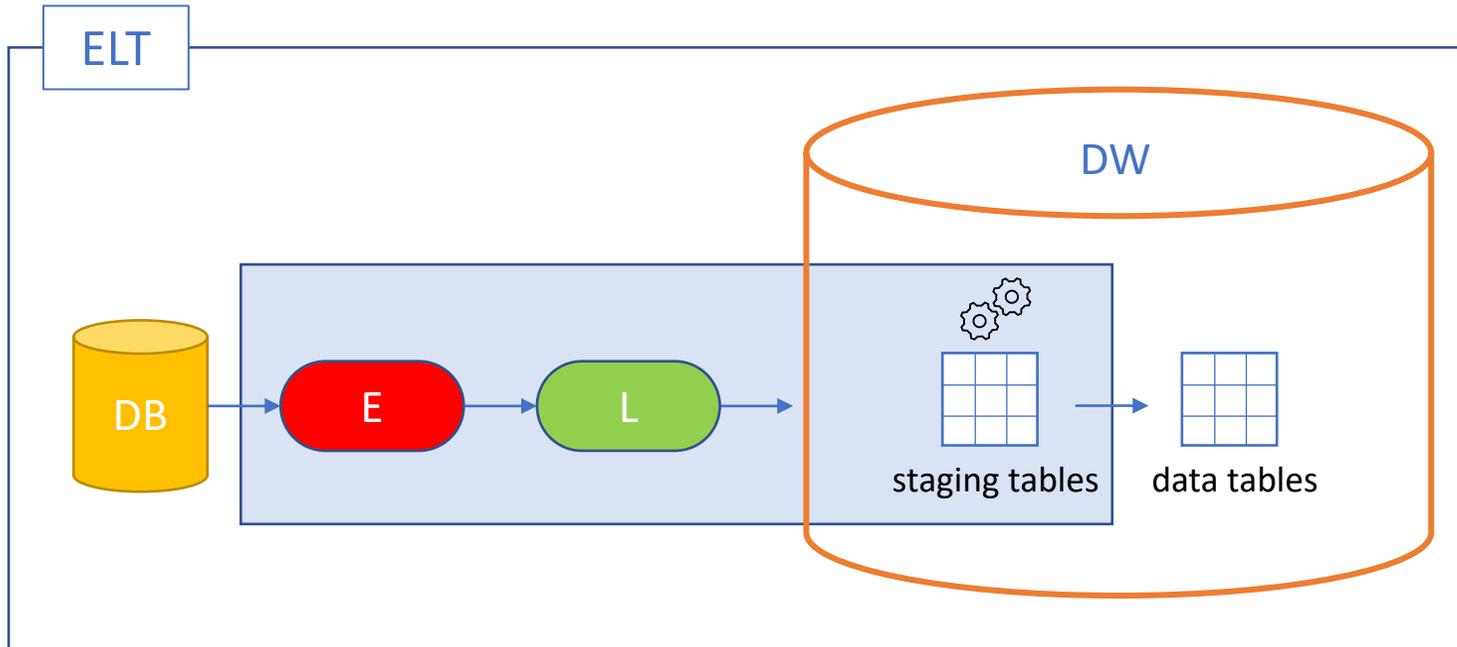
# Evolution of the ETL architecture



## 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, ...

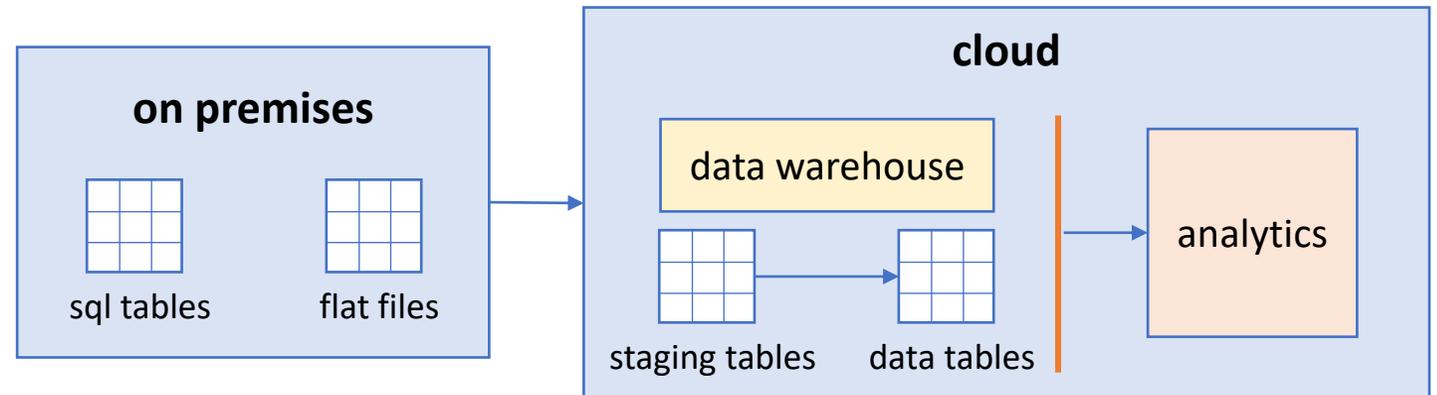
# Trends in ETL processing

- 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

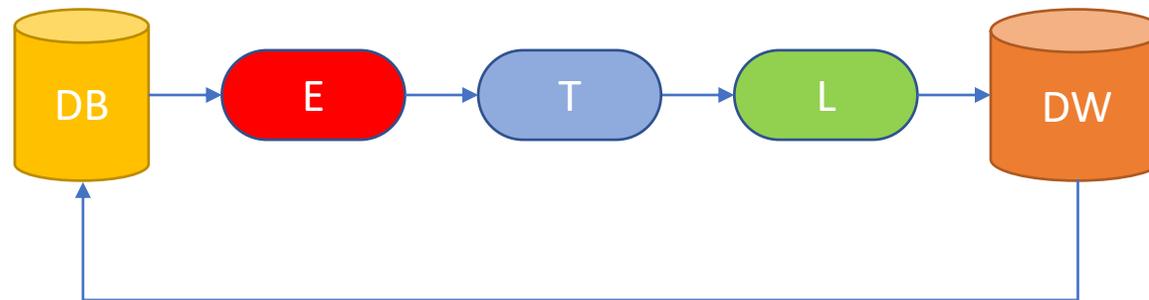
- elastic scalability
- massively parallel processing jobs
- ability of routinely start/stop jobs fast
- horizontal/vertical autoscaling
- run serverless pipelines
- dynamic work rebalancing
- flexible resource scheduling



# Trends in ETL processing

- Reverse ETL

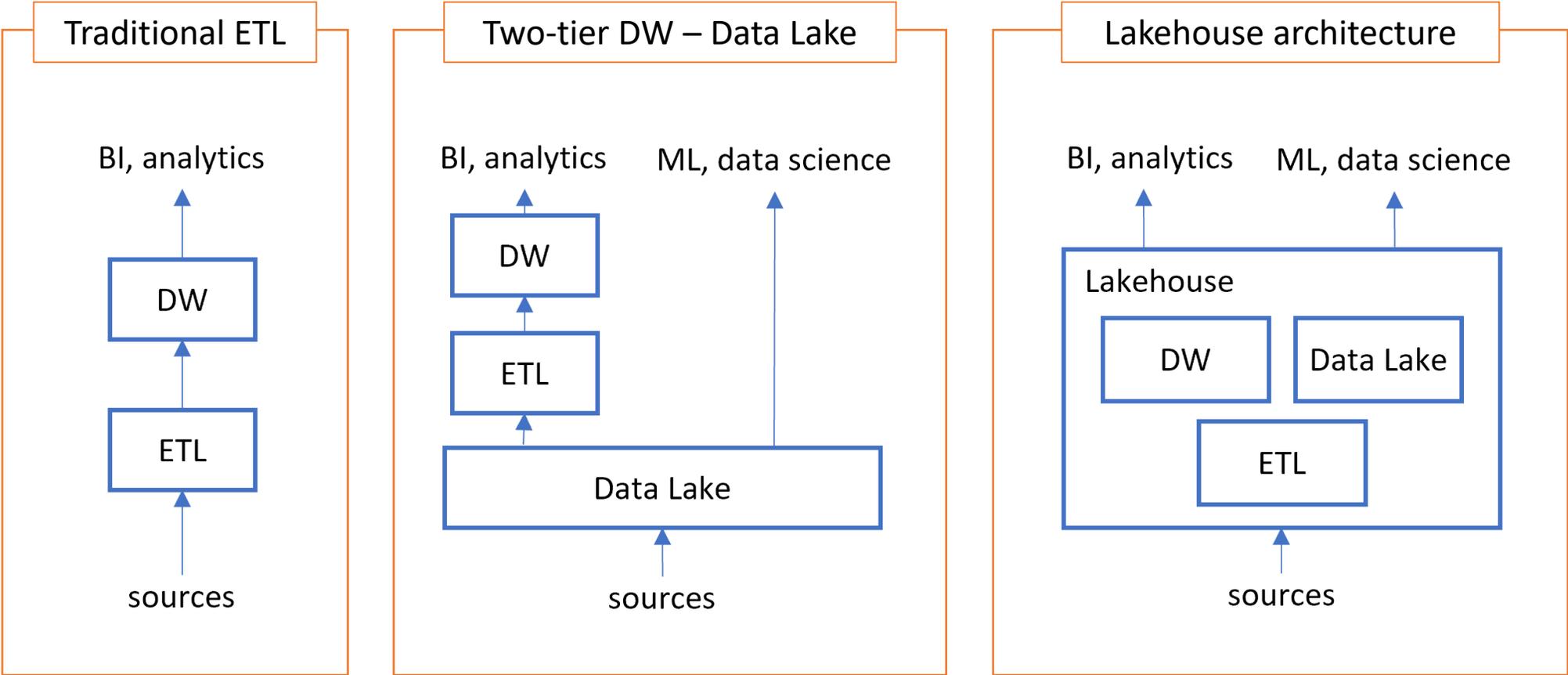
- Operational stores have isolated views of the domain
- DW has a global view
- Treat DW as an operational store that pushes insights back into the data sources



- 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