An Holistic Data Architecture:
From Source to Insight

Rick F. van der Lans
Industry analyst

CONSULTANCY

Rick F. van der Lans

Rick F. van der Lans is a highly-respected independent analyst, consultant, author, and internationally acclaimed lecturer specializing in data warehousing, business intelligence, big data, and database technology. He is managing director of R20/Consultancy BV.

He has presented countless seminars, webinars, and keynotes at industry-leading conferences. Rick helps clients worldwide to design their data warehouse, big data, and business intelligence architectures and solutions and assists them with selecting the right products. He has been influential in introducing the new logical data warehouse architecture worldwide which helps organizations to develop more agile business intelligence systems.

He is the author of several books on computing, including his new *Data Virtualization: Selected Writings* and *Data Virtualization for Business Intelligence Systems*. Some of these books are available in different languages. Books such as the popular *Introduction to SQL* is available in English, Dutch, Italian, Chinese, and German and is sold world wide. He also authored numerous whitepapers for vendors.

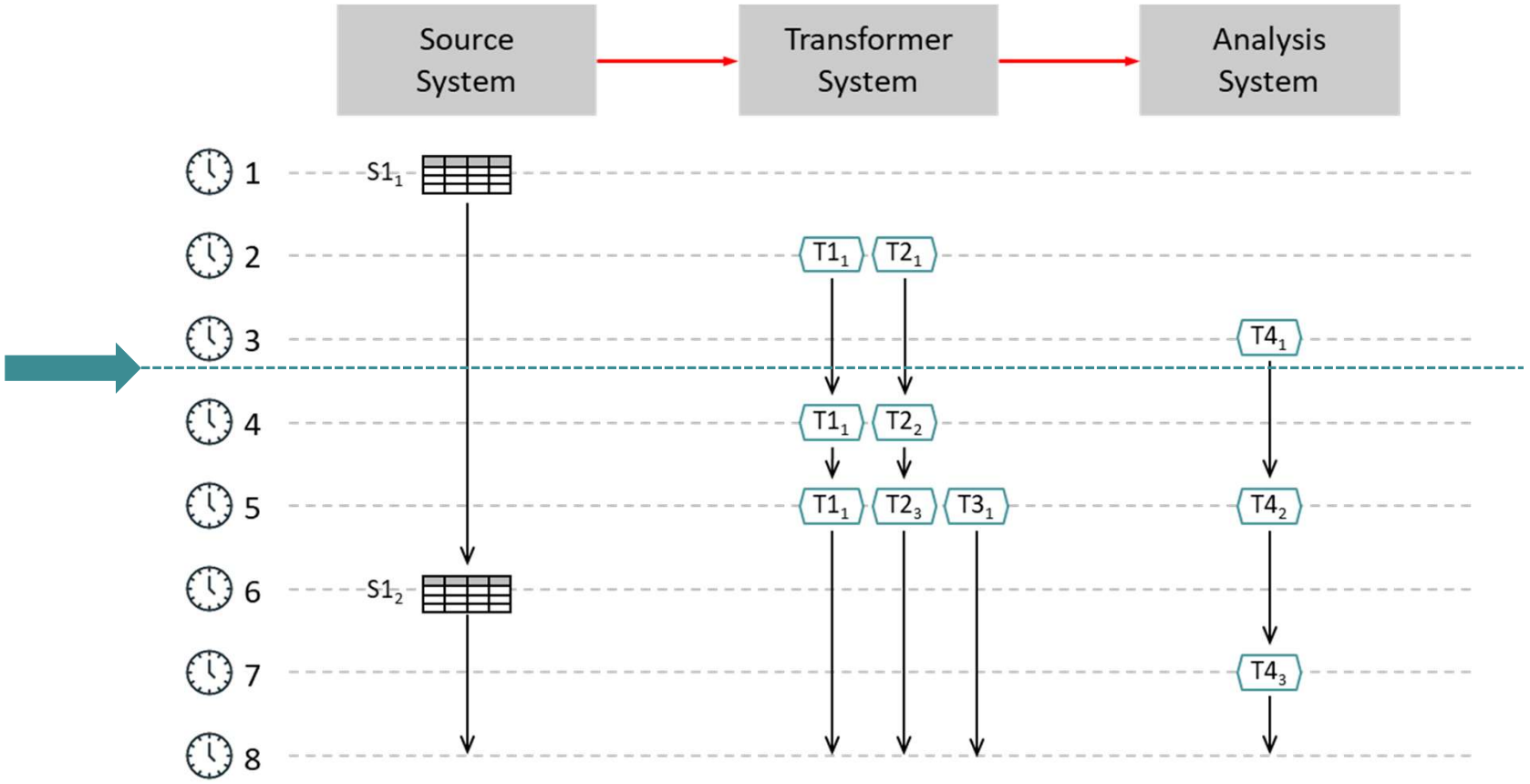
In 2018 he was selected the sixth most influential BI analyst worldwide by onalytica.com.

You can get in touch with Rick van der Lans via:
Email: rick@r20.nl
Website: www.r20.nl
LinkedIn: http://www.linkedin.com/pub/rick-van-der-lans

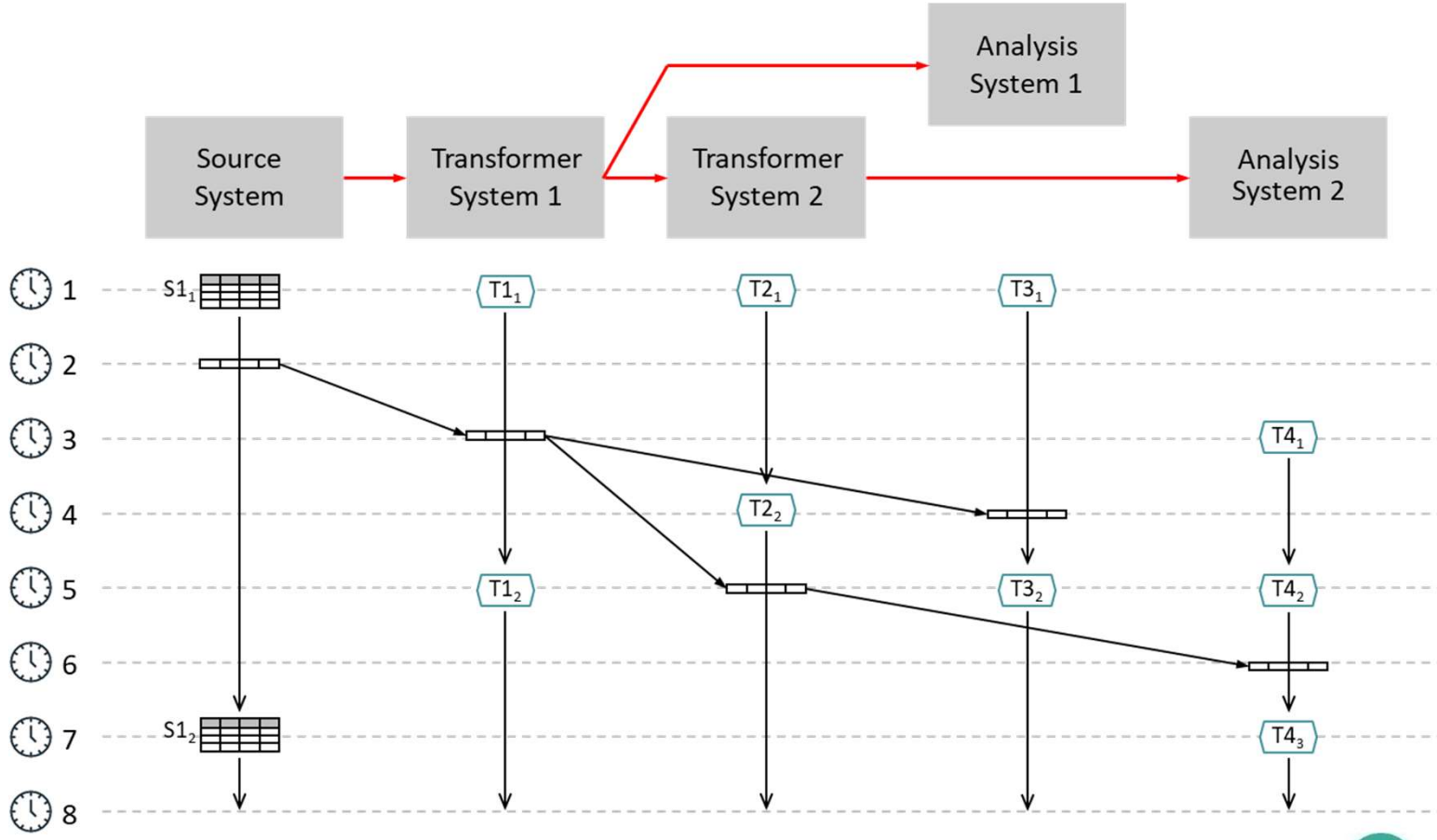
Challenges, Wishes and Requirements

- Speeding up development of *cross-system analysis systems*
- Speeding up development of *cross-organization analysis systems*
- Simplifying the *synchronization of source systems*
- Increasing data processing transparency for *process/decision reconstruction*
- Increasing the *discoverability* and clarity/meaning of data
- Improving *data quality*
- Strengthening *data security* and *data privacy*
- Strengthening *data history*
- Handling unstructured data: video, sound, text, and images
- Simplifying the data processing landscape
- Smarter uses of new technology

Example of Horizontal Lineage With Versions



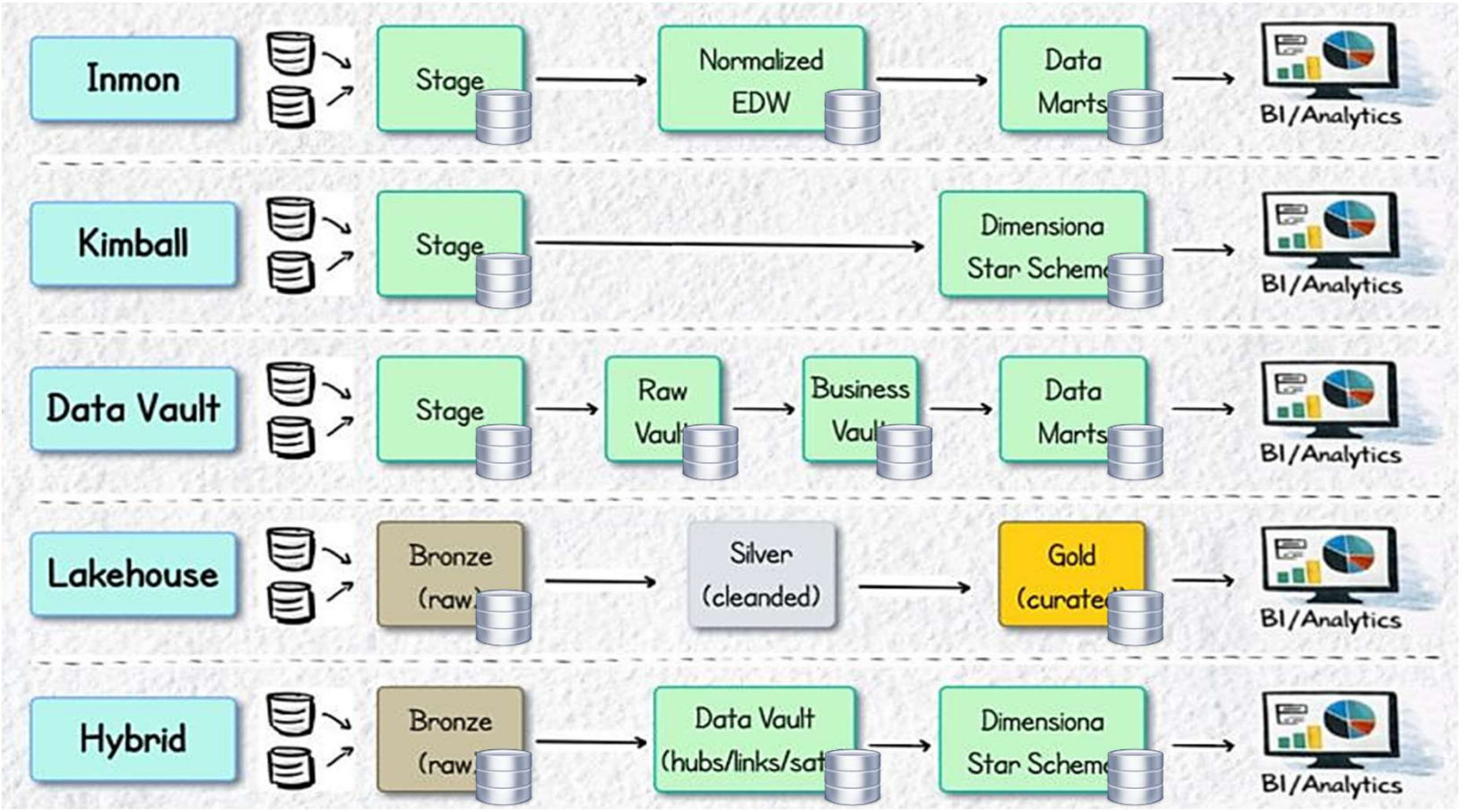
Example of Operational Lineage



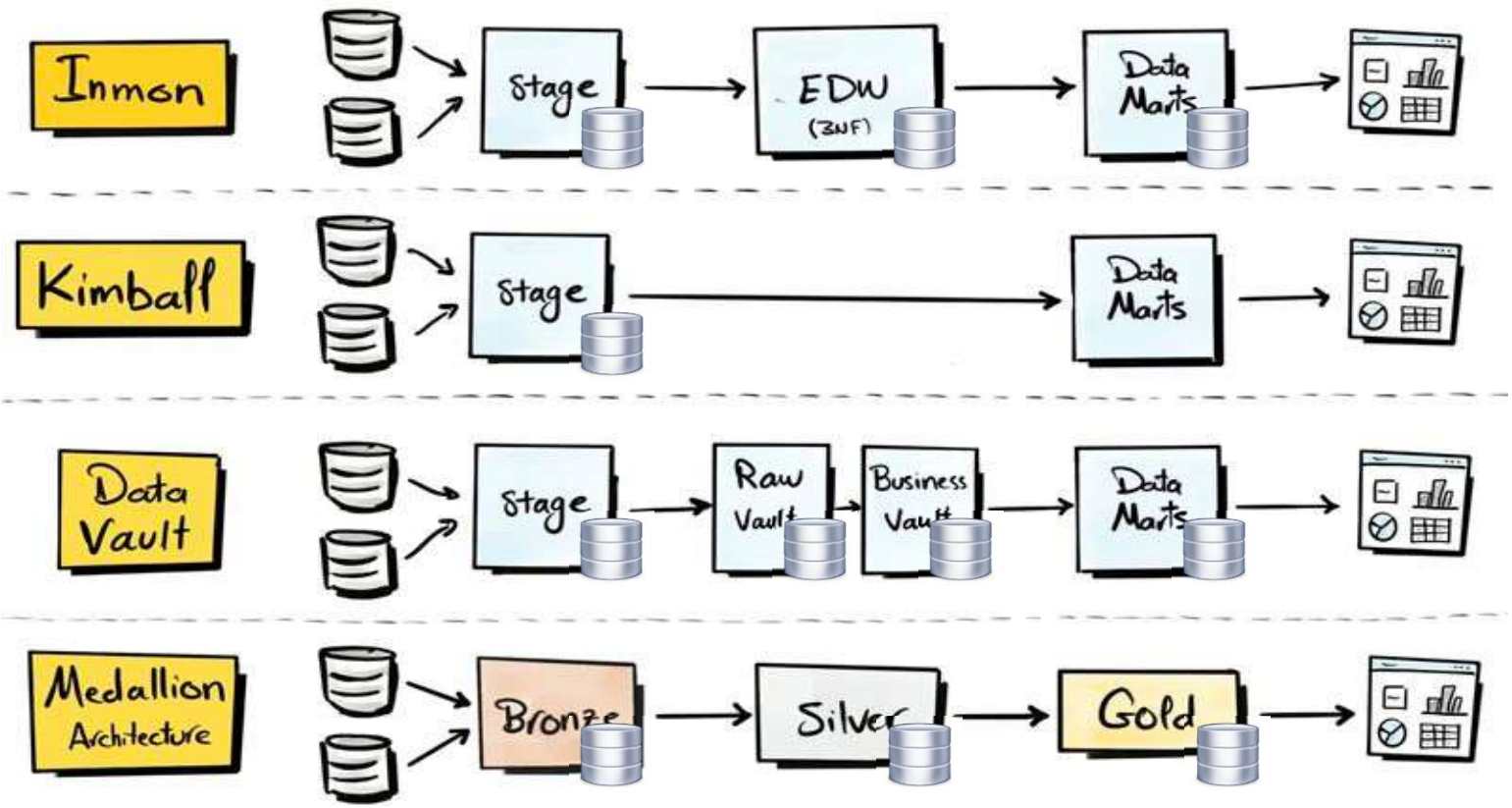
Challenges, Wishes and Requirements

- Speeding up development of *cross-system analysis systems*
- Speeding up development of *cross-organization analysis systems*
- Simplifying the *synchronization of source systems*
- Increasing data processing transparency for *process/decision reconstruction*
- Increasing the *discoverability* and clarity/meaning of data
- Improving *data quality*
- Strengthening *data security* and *data privacy*
- Strengthening *data history*
- Handling unstructured data: video, sound, text, and images
- Simplifying the data processing landscape
- Smarter uses of new technology

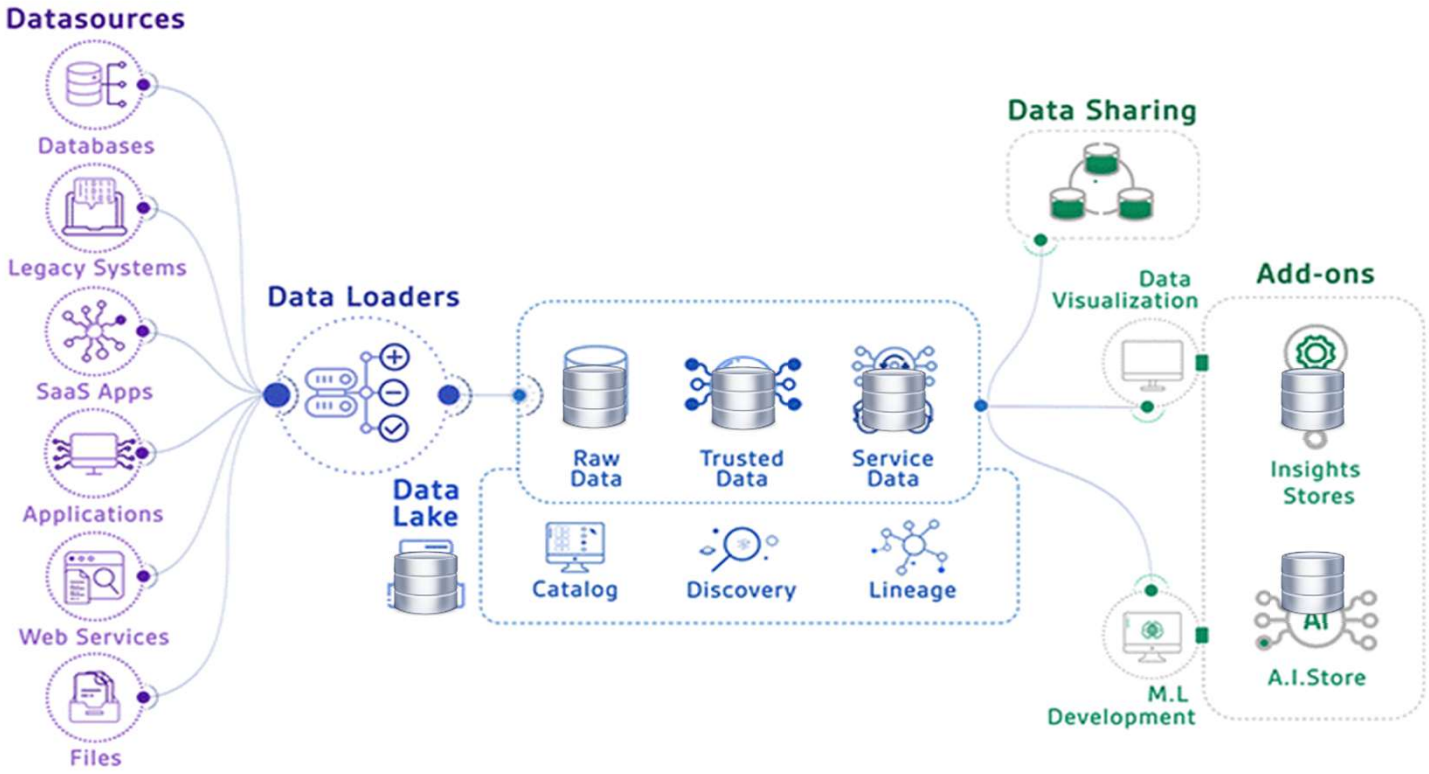
Domain Data Architectures (1)



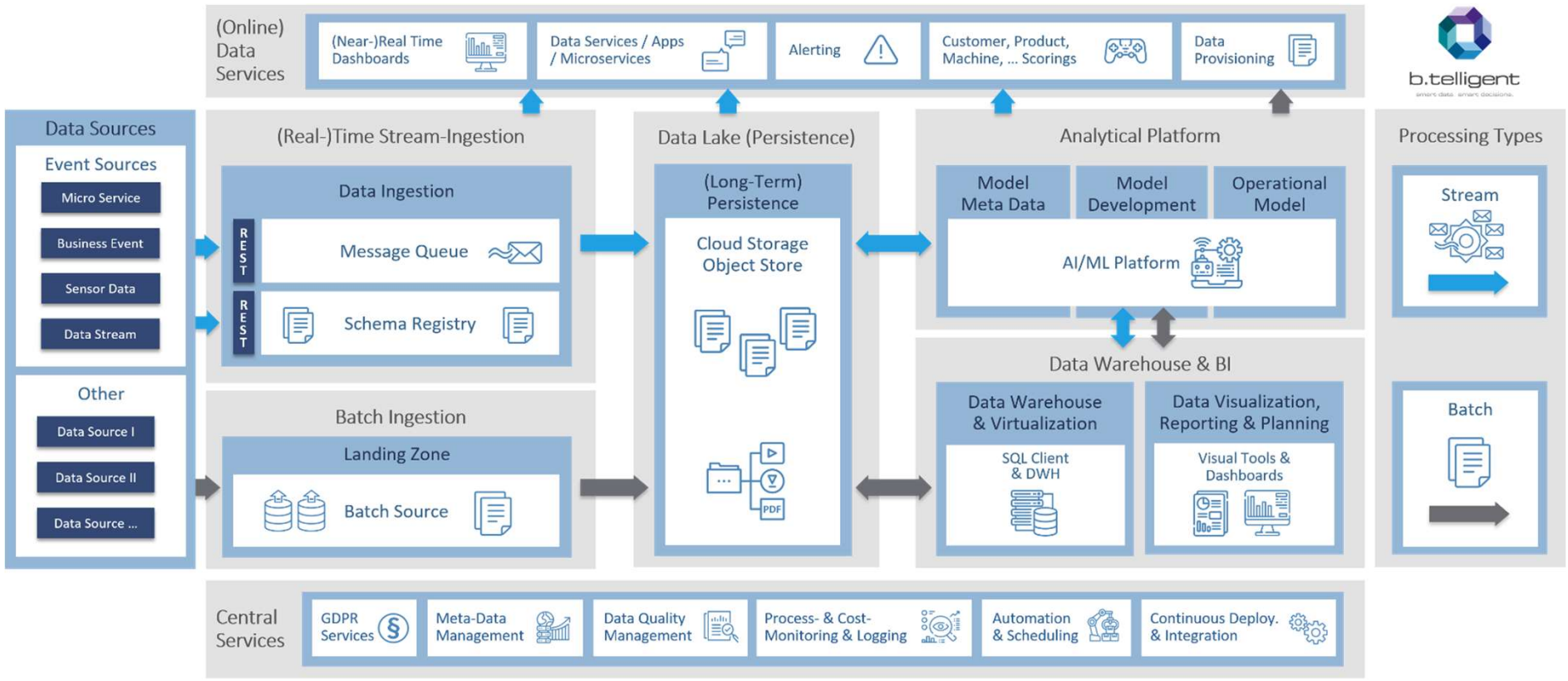
Domain Data Architectures (2)



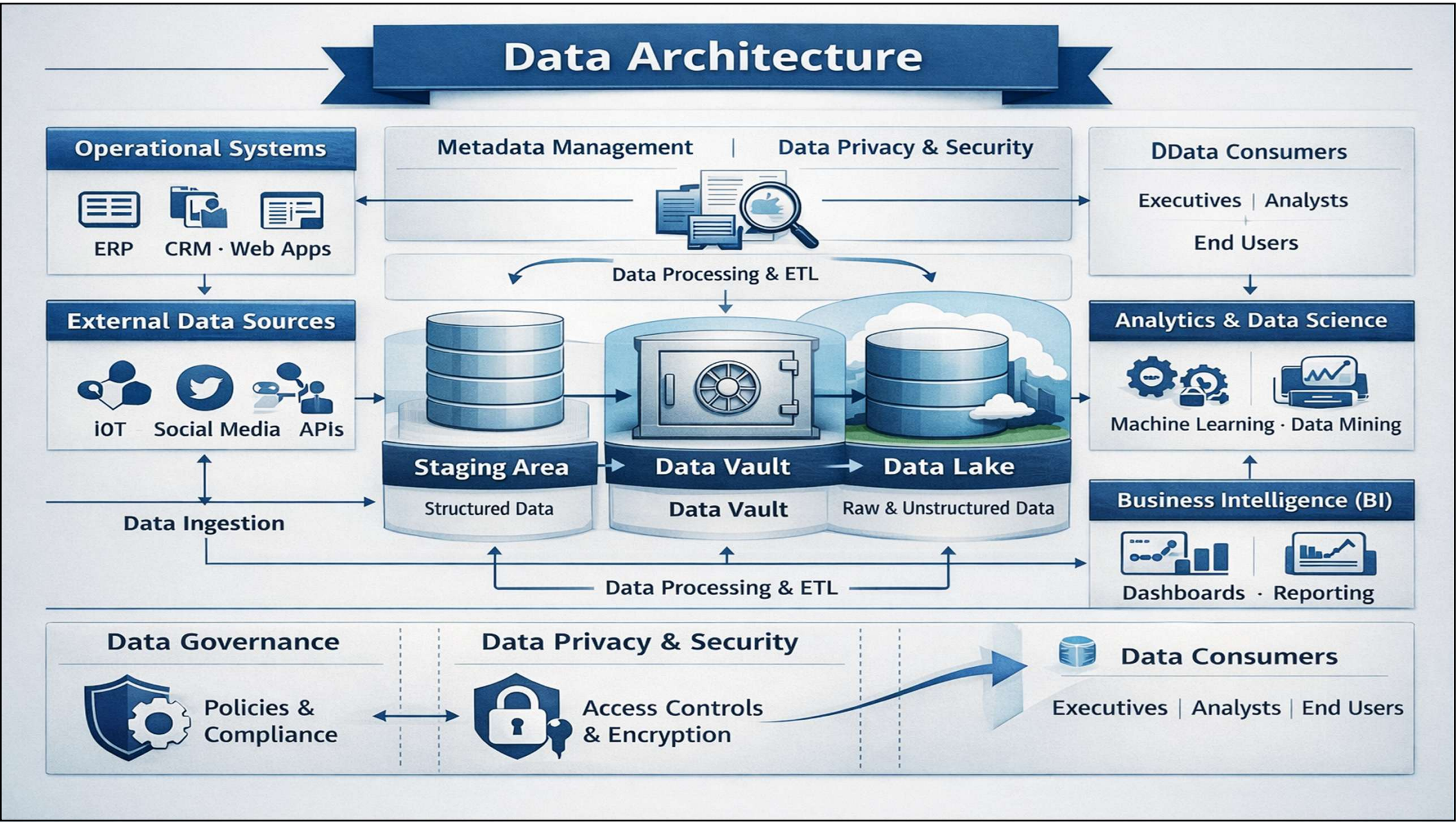
Domain Data Architectures (3)



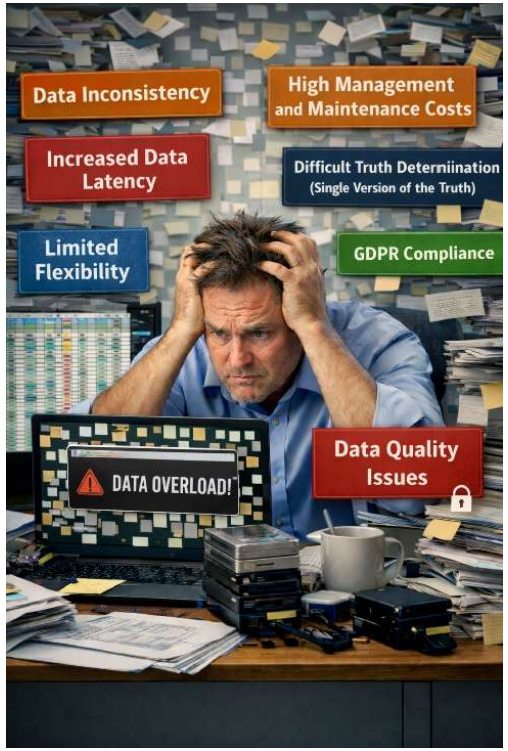
Data-copy Rich Data Architectures



Data Architecture

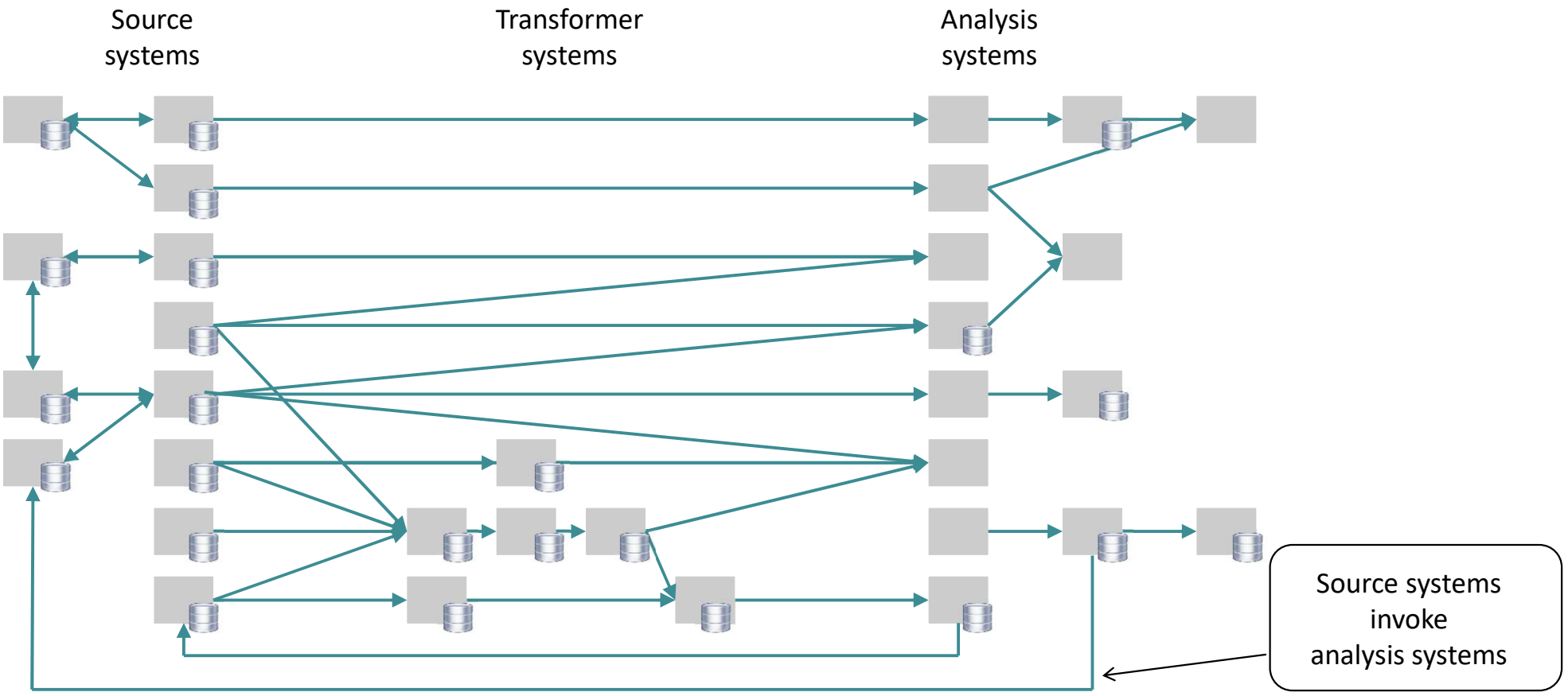


Disadvantages of Data Copying

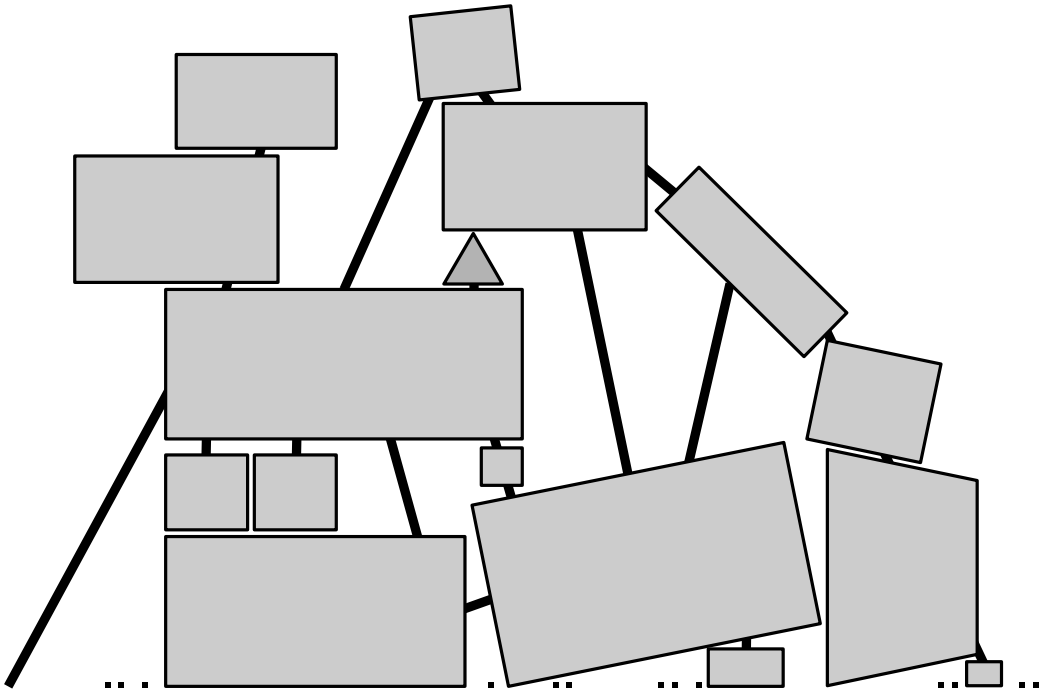


- Data inconsistency
- Difficult determination of the truth ("single version of the truth")
- Limited flexibility and agility
- High management, development, and maintenance costs
- More complex compliance with General Data Protection Regulation (GDPR)
- Complex data synchronization issues
- Data quality challenges
- More complex data security
- Outdated data (increased data latency)
- …

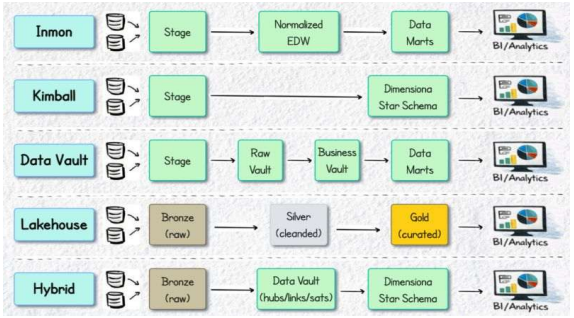
Evolution of the Data Processing Landscape

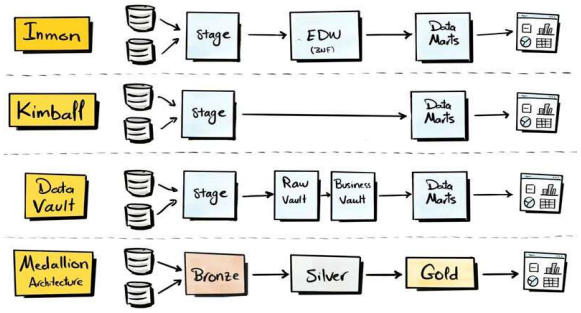


“Wobbly” Foundation of the Data Processing Landscape



Limitations of Some Data Architectures





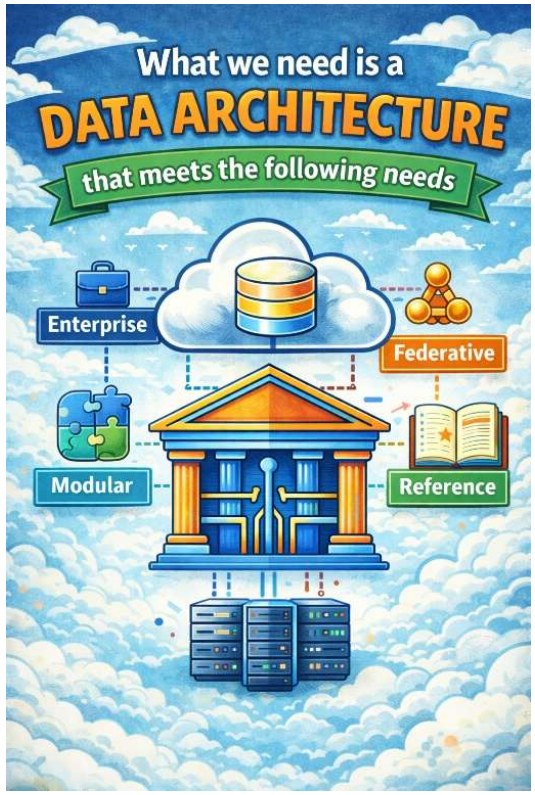
- Source systems are unjustly out of scope
  - Old and inflexible
  - Software packages
  - No IT sovereignty
  - No or limited data history
  - Poor data security
  - Complex data synchronization
- Analysis systems are unjustly out of scope
  - Many data copies
  - Not under control of IT
  - Analysis models input to source systems
- Many requirements are unjustly out of scope
  - Reconstruction
  - Metadata for everyone
  - Data-sharing versus data-exchange

Bring Your (Data)House in Order

- Data in order
- Metadata in order
- Lineage in order (origin and impact)
- Reconstruction in order
- Data security in order
- Technology in order

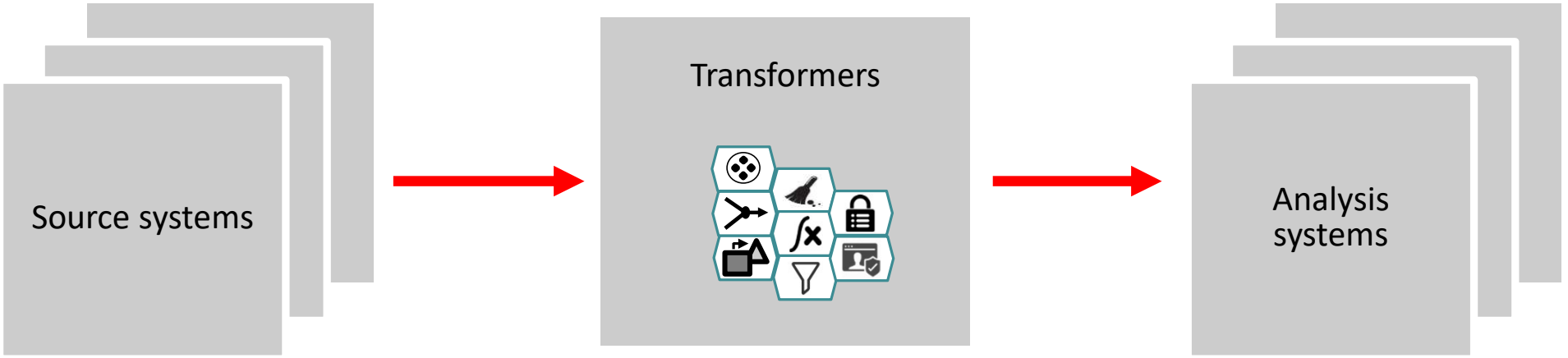
- It all starts with *data architecture in order*

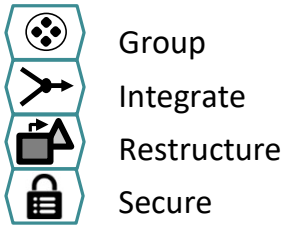
What We Need is a …



- *Enterprise* data architecture
  - Not a domain data architecture
  - Covers the entire data route/path
  - Not only focused on analysis and data science (and AI)
- *Modular* data architecture
  - Think big, act small
- *Federated* data architecture
  - Data storage minimization – less data copies
- *Reference* data architecture
  - Generic, reusable

Start with Transformers, Not With Databases





Group
Integrate
Restructure
Secure

Correct
Calculate
Filter
Anonymize

Examples of Transformers

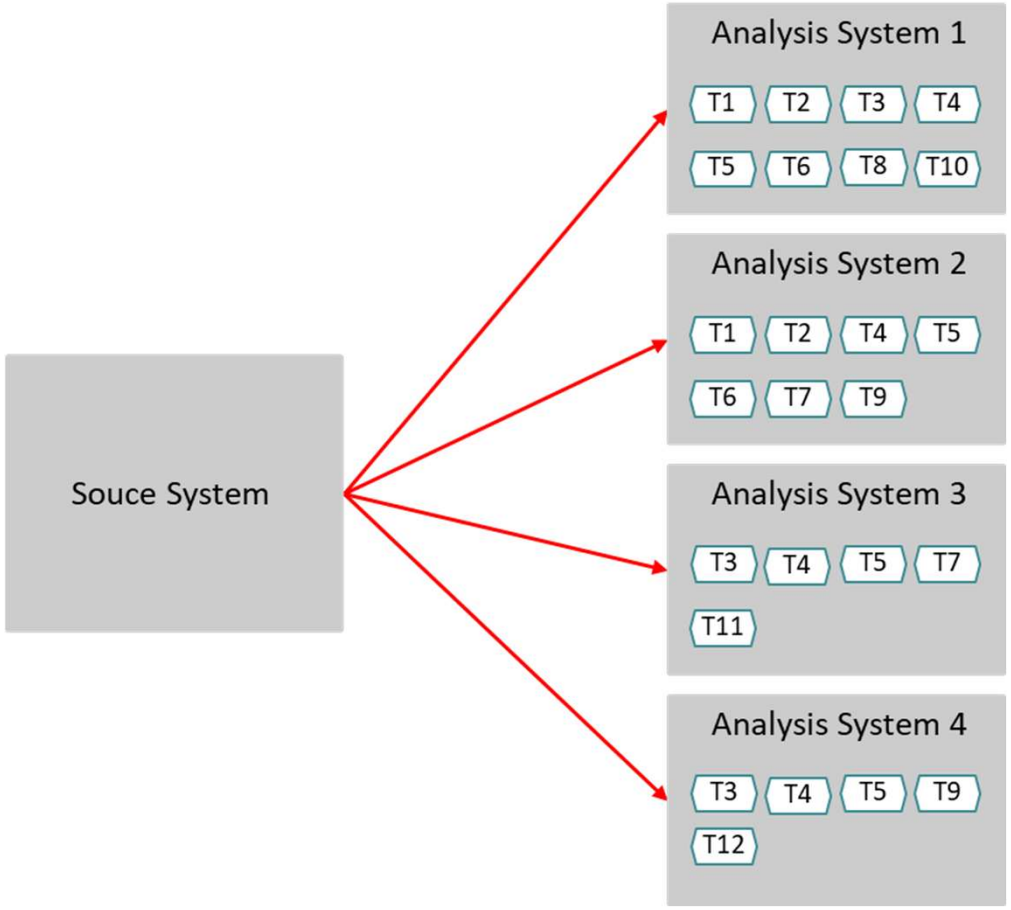
- Data value transformations
- Data structure transformations
- Aggregations
- Filters
- Groupings
- Calculations
- Integrations
- Technical corrections
- Functional corrections
- Anonymizations
- Historizations
- …

- Summarize
- Keywords extraction
- Translate
- Named Entity Recognition (NER)
- Optical Character Recognition (OCR)
- Text-to-speech
- Speech-to-text
- Data tagging
- Sentiment analysis
- Pattern recognition
- …

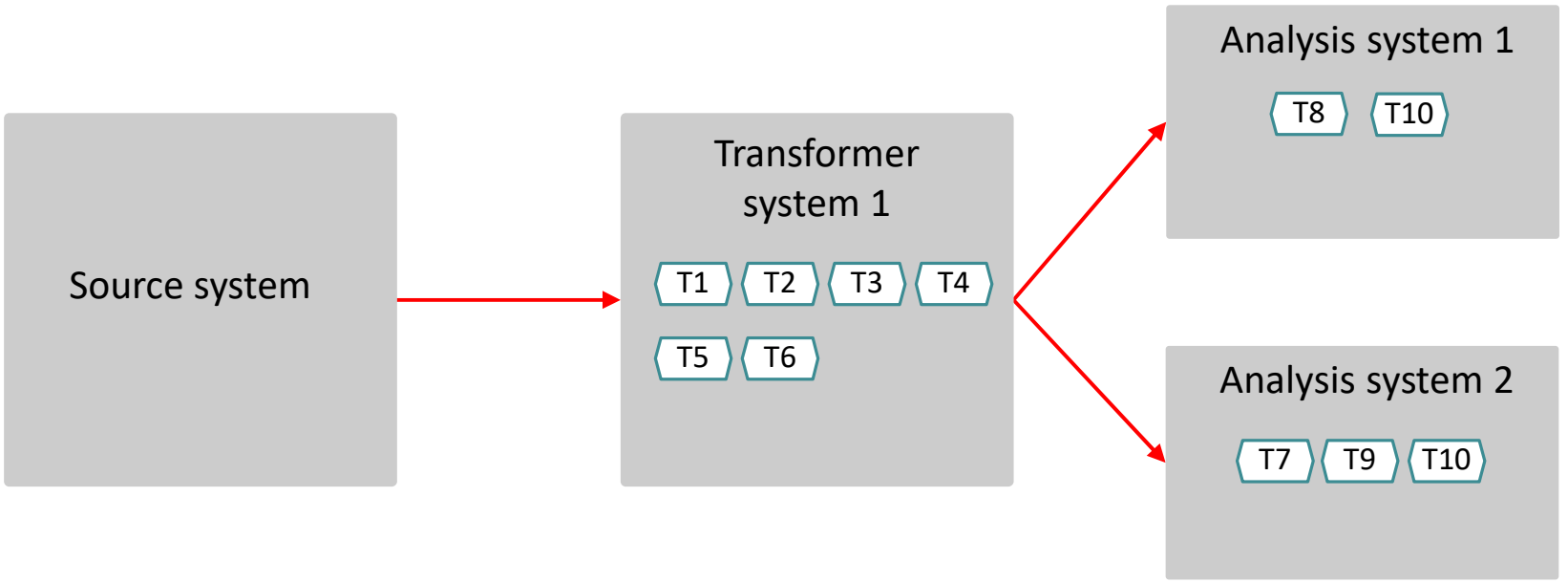
- Clustering
- Normalization
- Interpolation
- Extrapolation
- Binning
- Sampling
- Regression
- …

- Vector-to-Raster Conversion
- Containment
- Clipping
- Spatial intersection
- Buffering
- …

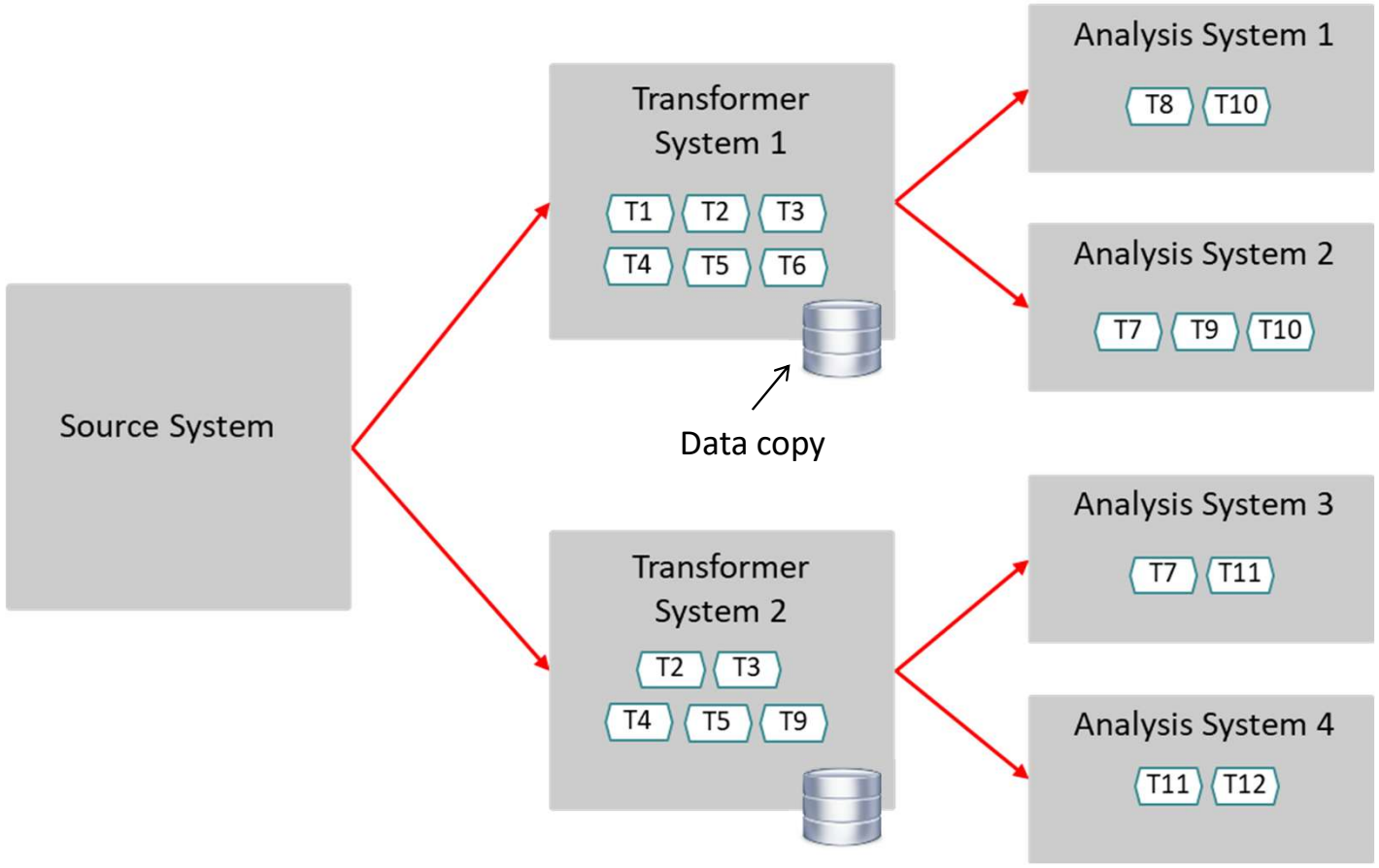
Duplicate Transformers in Analysis Systems



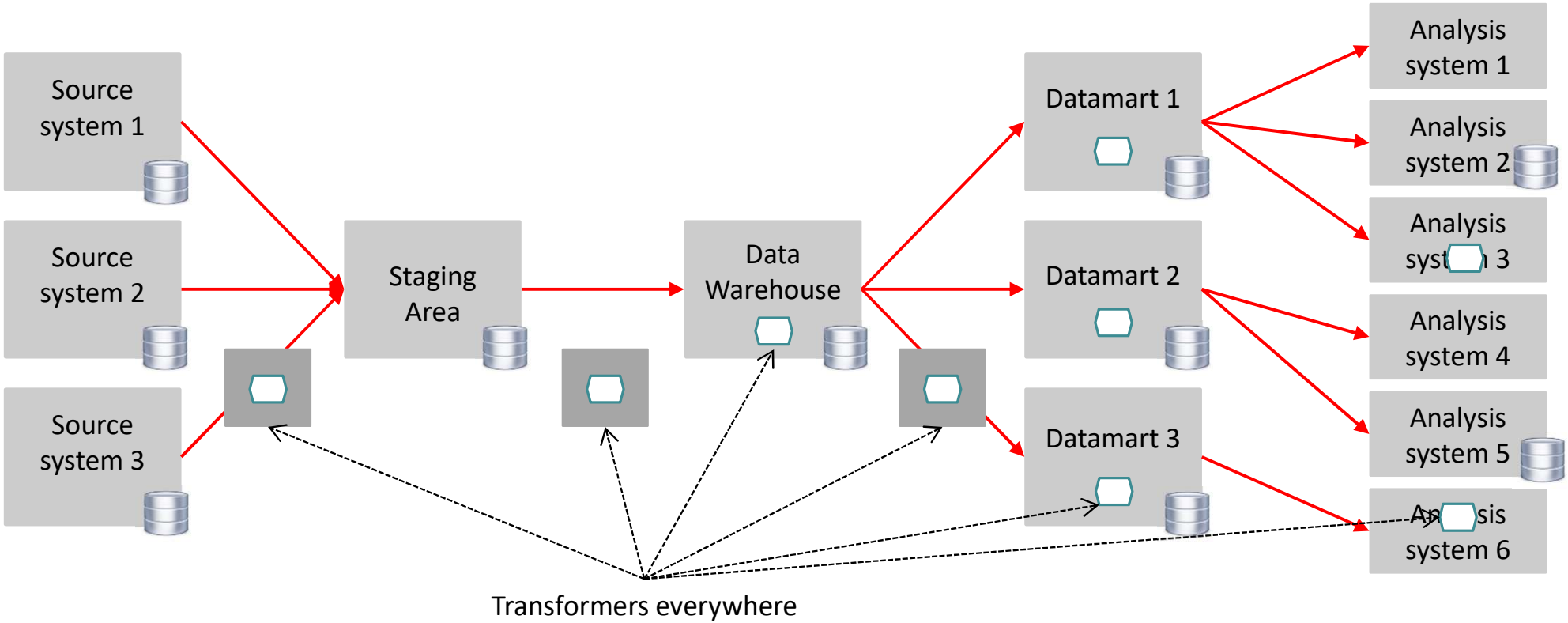
Centralizing Transformers



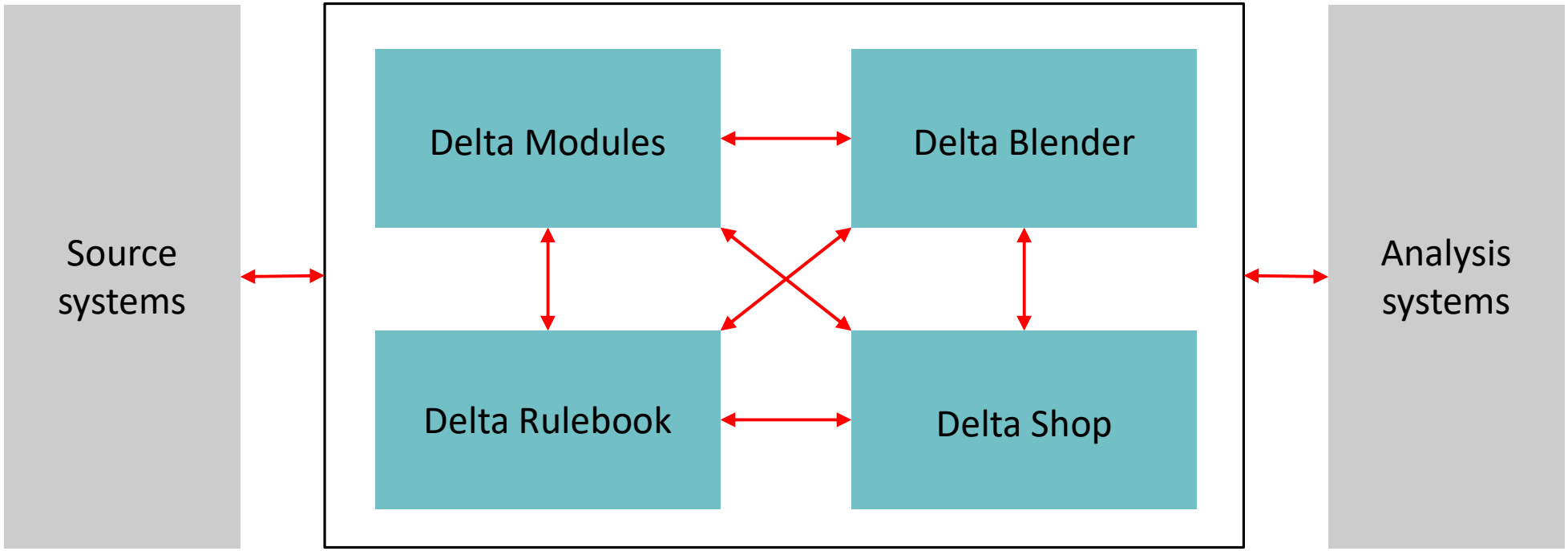
Dupulicate Transformers in Transformer Systems



A Traditional Data Warehouse Architecture



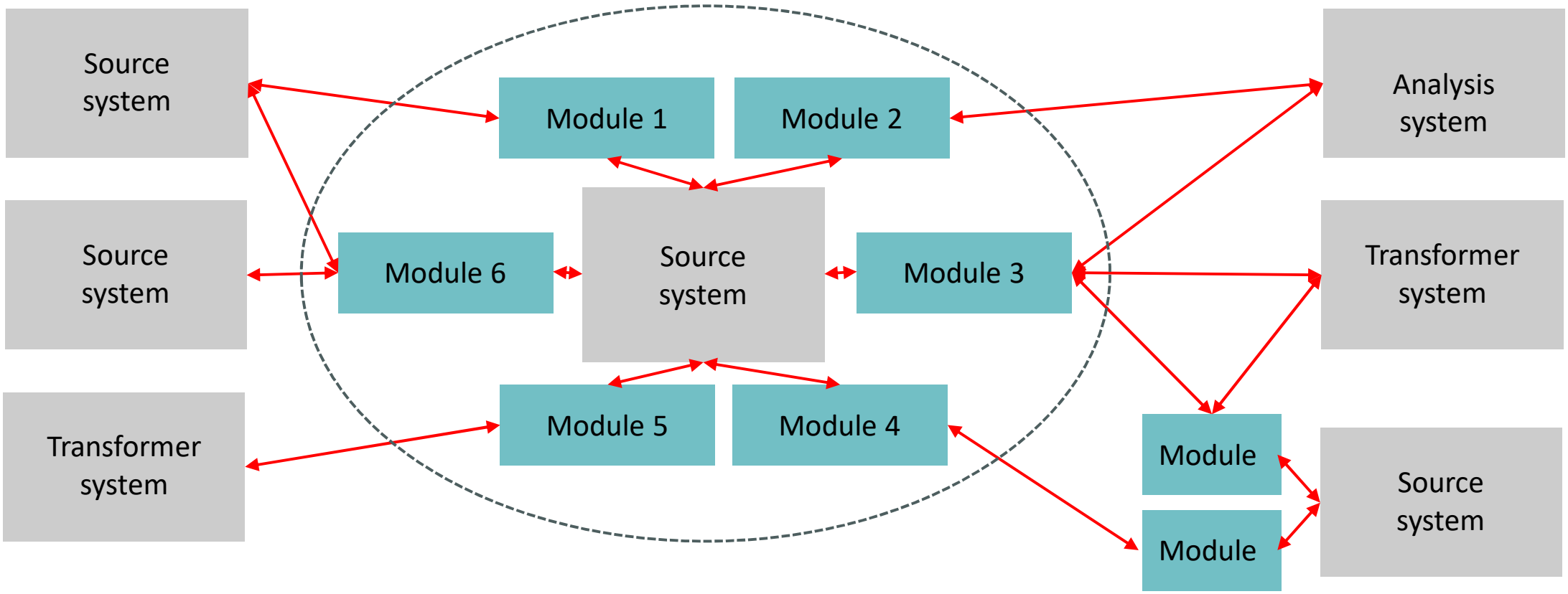
Delta: A Modern, Enterprise Data Architecture



Functions of Source Systems

- Entering, modifying, and deleting data
  - For many business objects
- Retrieving data
- Securing data
- Data quality
- Backup and recovery tasks
- Calculating results
- Handling workflow and processes
- Triggering events
- *Connections with other systems*
- …

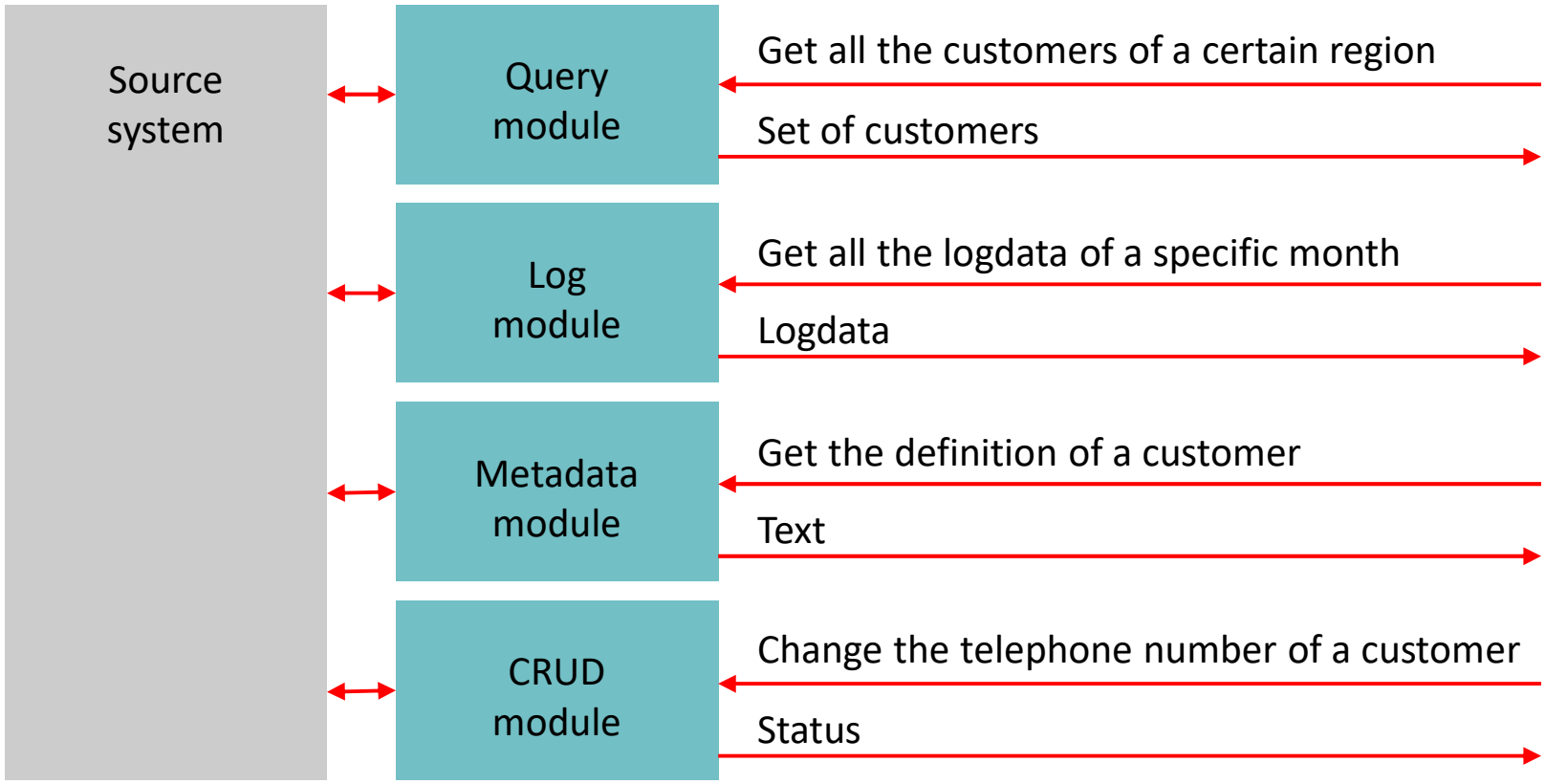
Wrapping the Source Systems - Abstraction



Six Delta Modules to Wrap Source Systems

- CRUD
- Query
- Metadata
- Log
- Messaging
- Data security

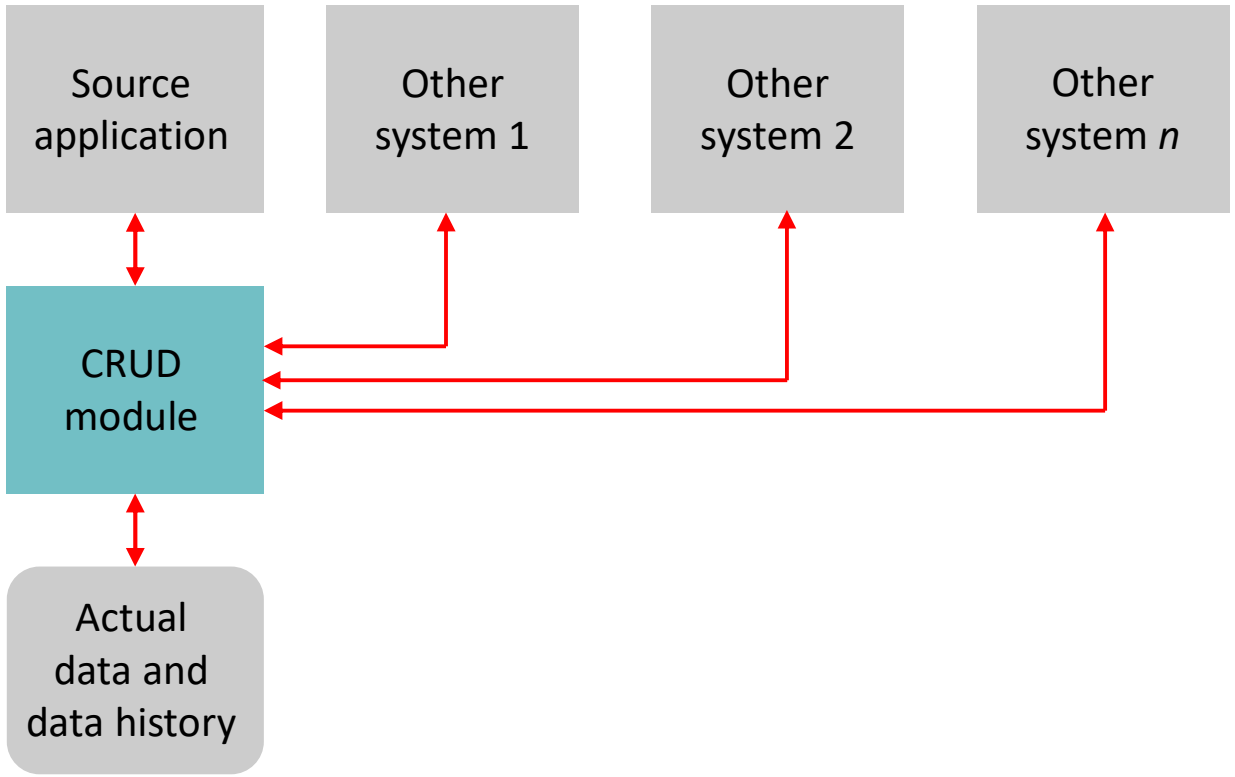
Modules and Functions



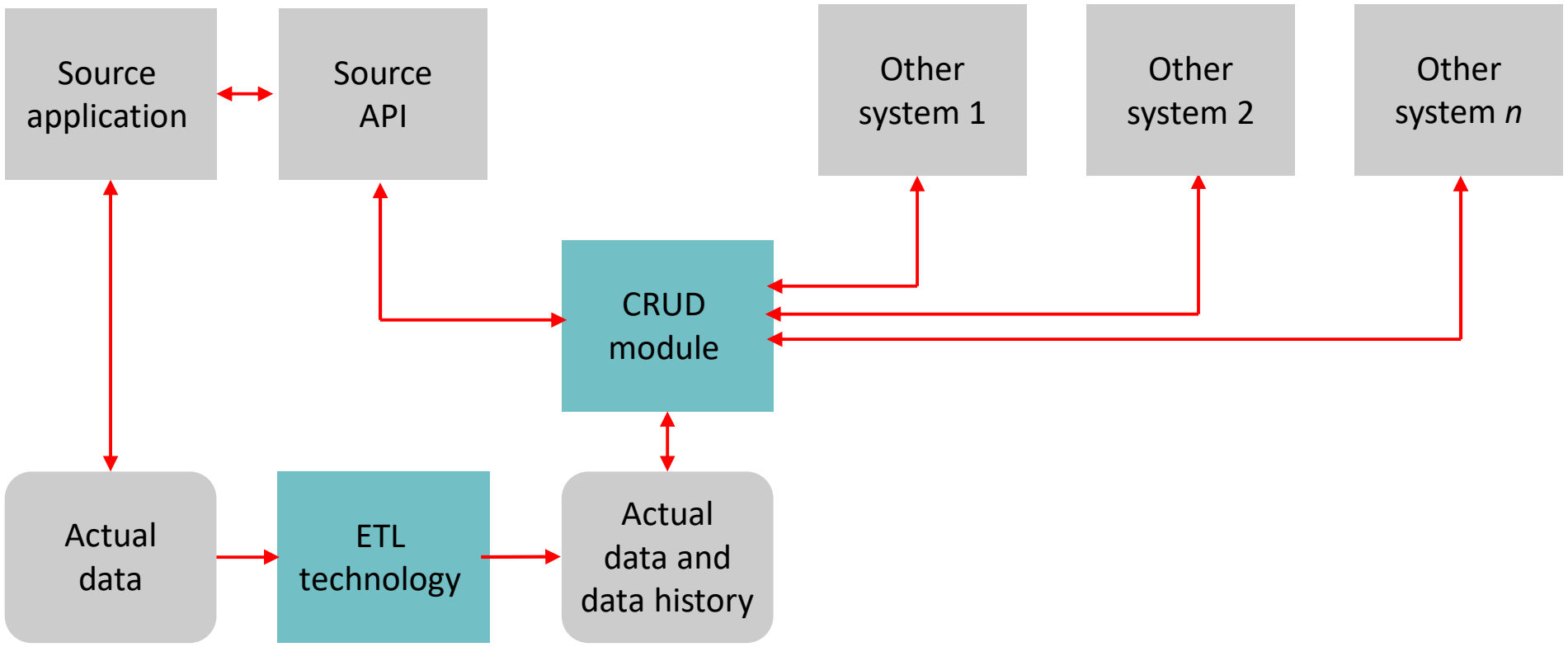
Functions of CRUD modules

- Offer functions to enter, view, modify, and delete data
- Hide internal and external source systems and files
- Data complies with the Enterprise Data Model
- Support time travel
  - Maintains data history if the source system does not
  - Two time dimensions: transaction time and real time
- Work with individual business objects
- Enforce data quality rules
- Involved in source system synchronization
- Replaces data exchange with data sharing
- …

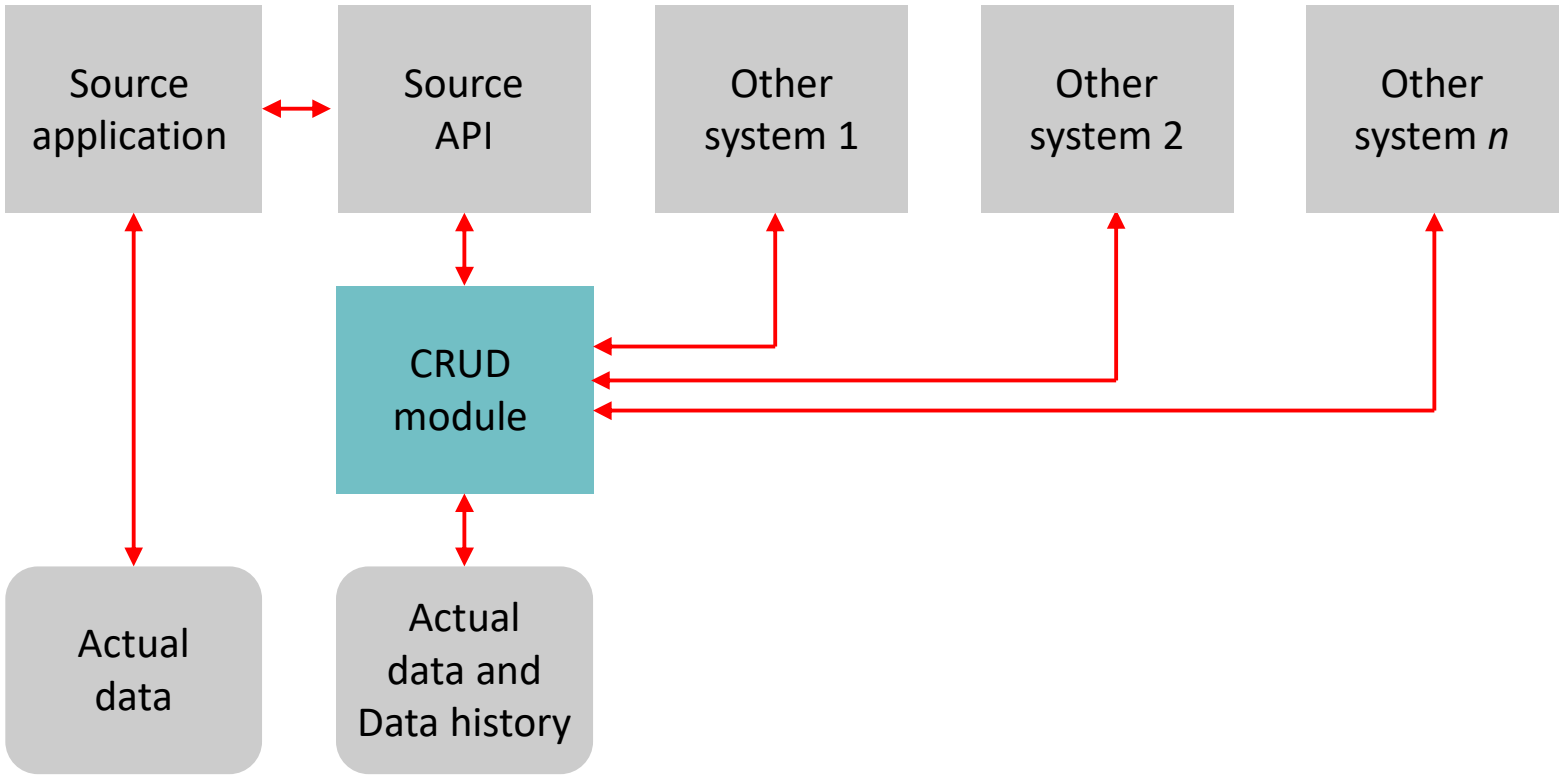
CRUD Module (1) The Dream



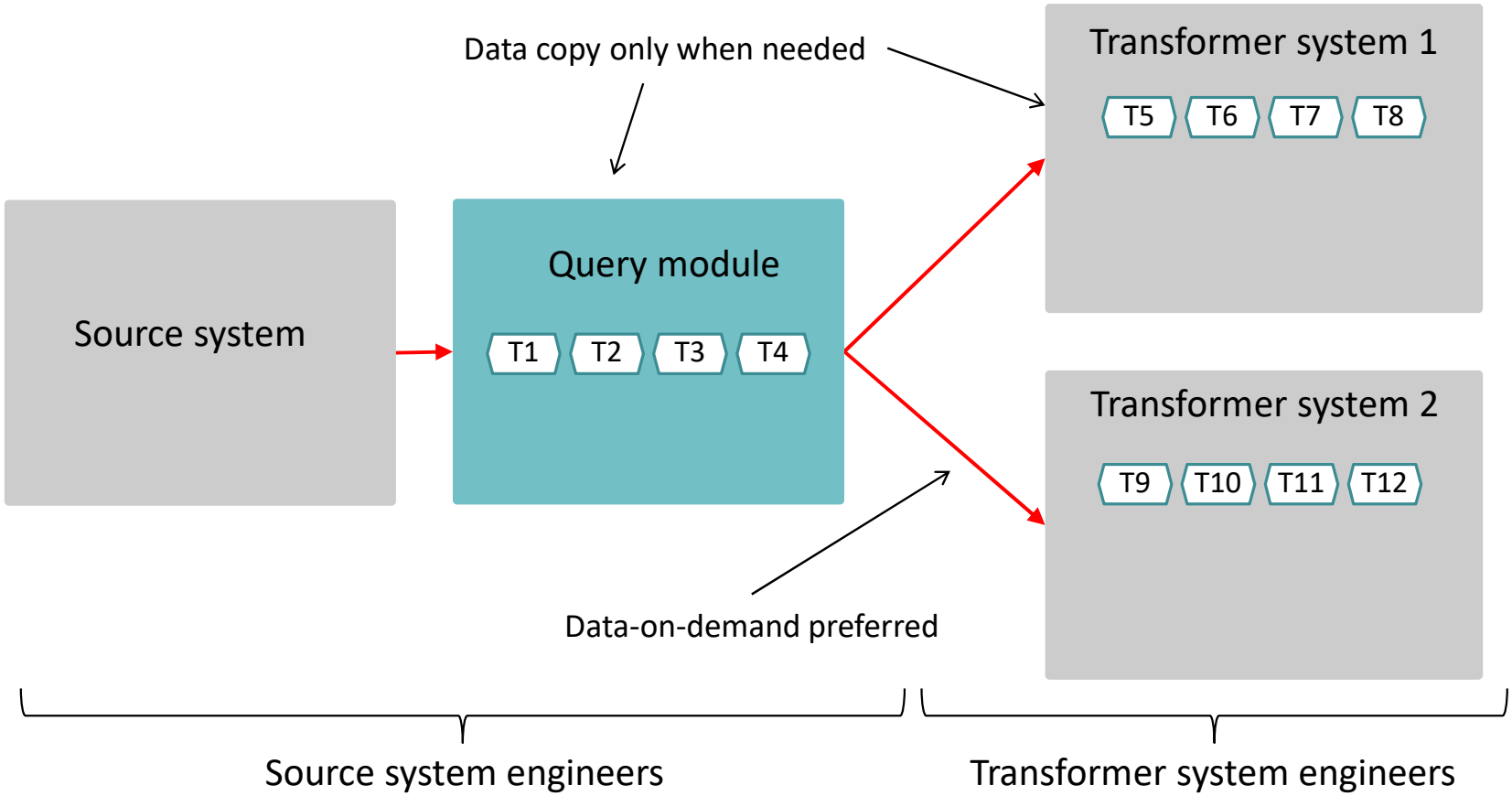
CRUD Module (2) Via API of Source System and ETL



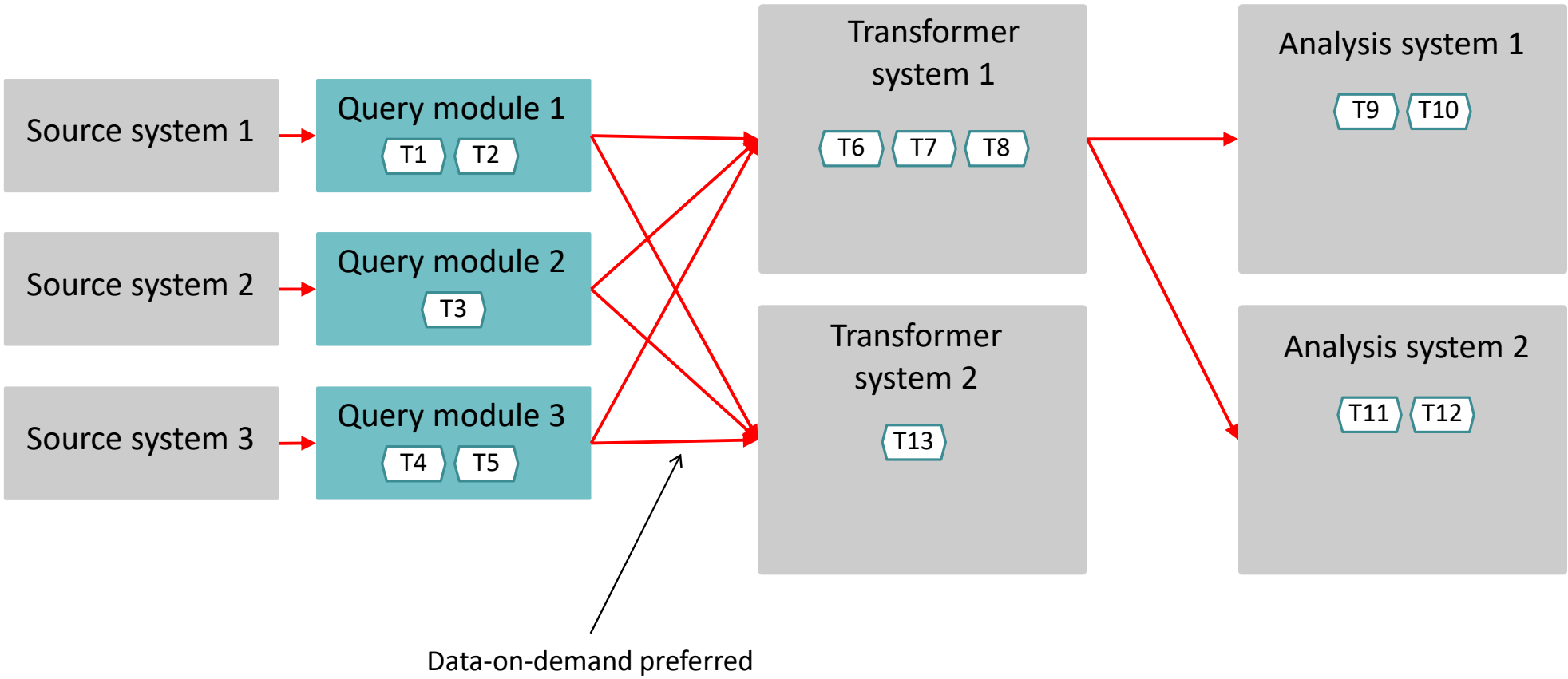
CRUD Module (3) Via API of Source System



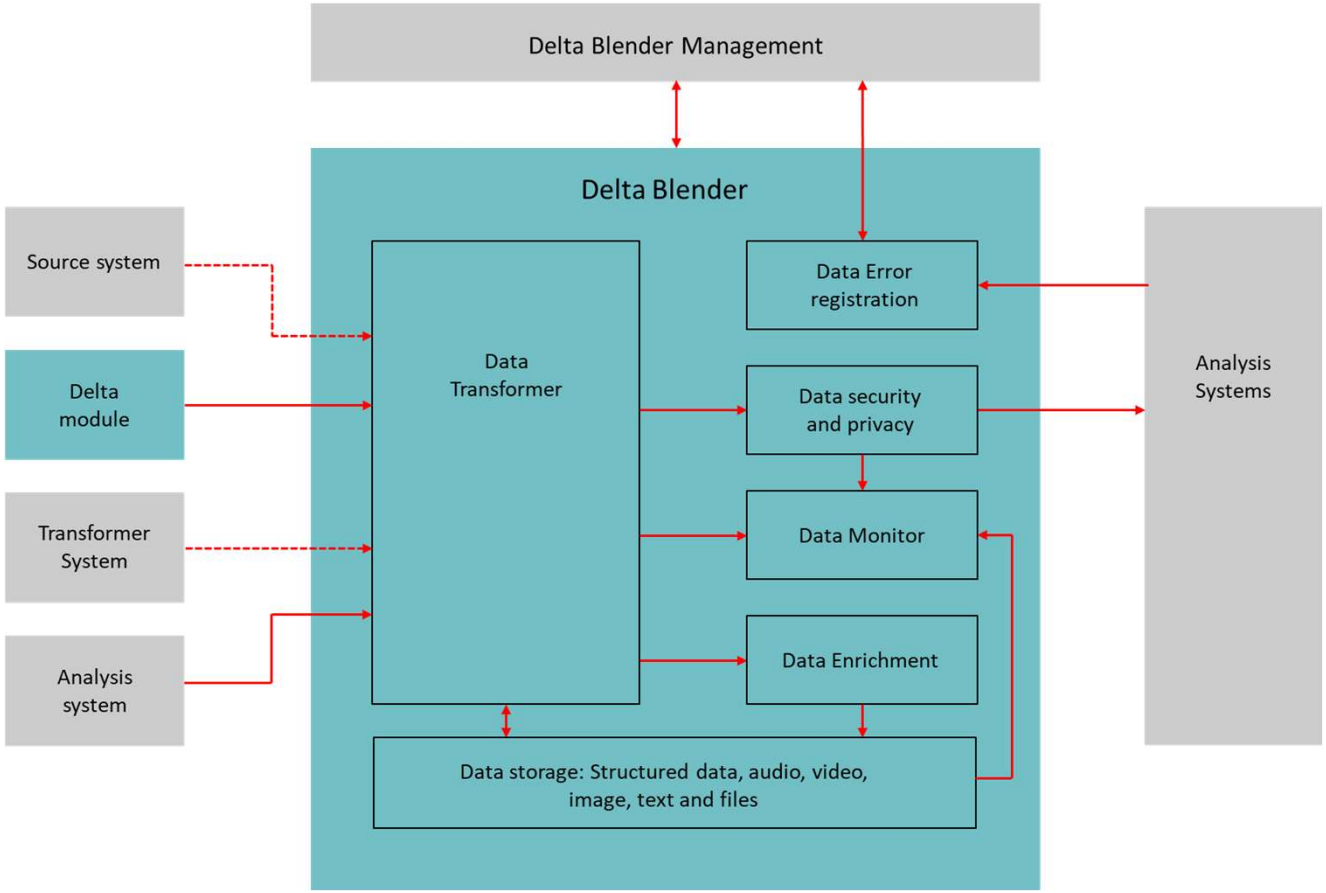
Transformers Closer to the Source Systems



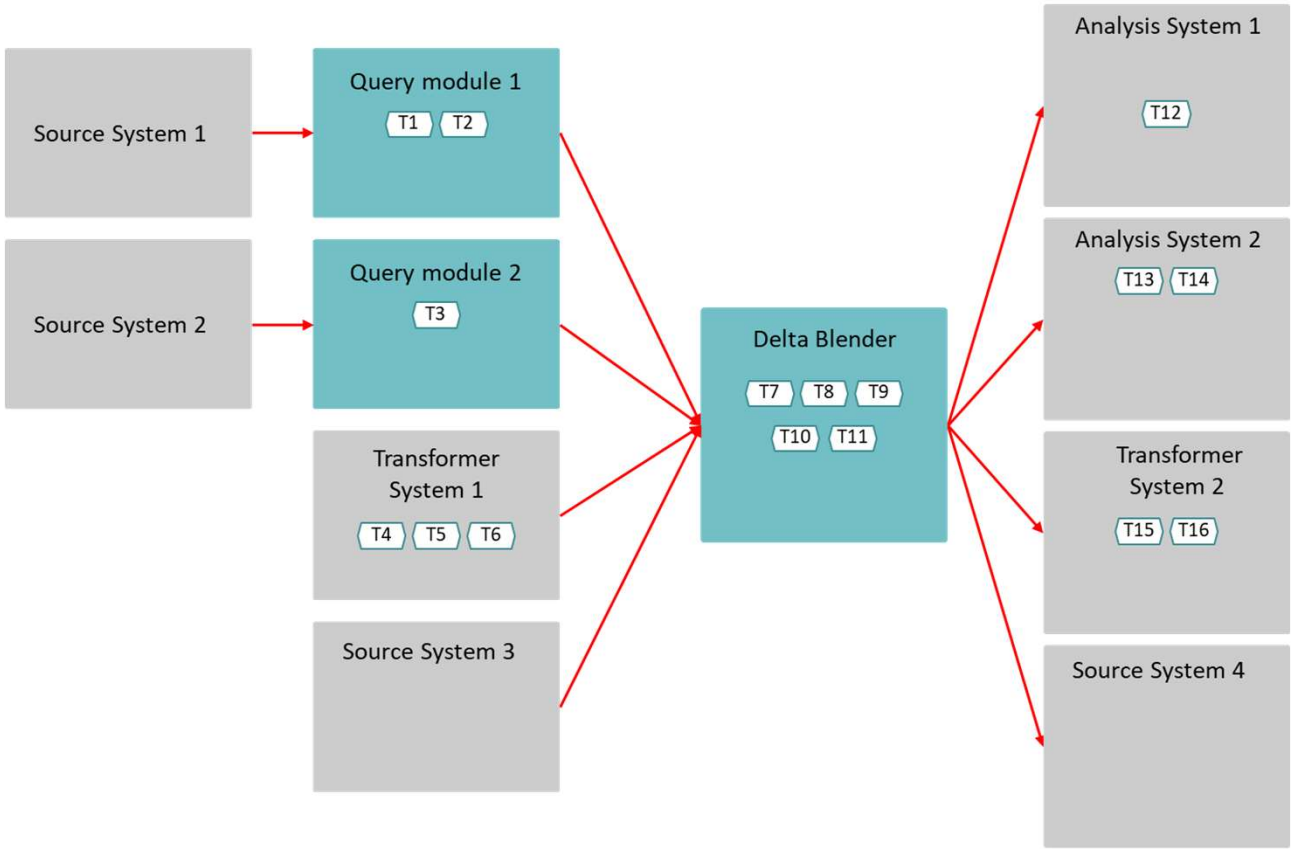
Transforming Data in Query Modules



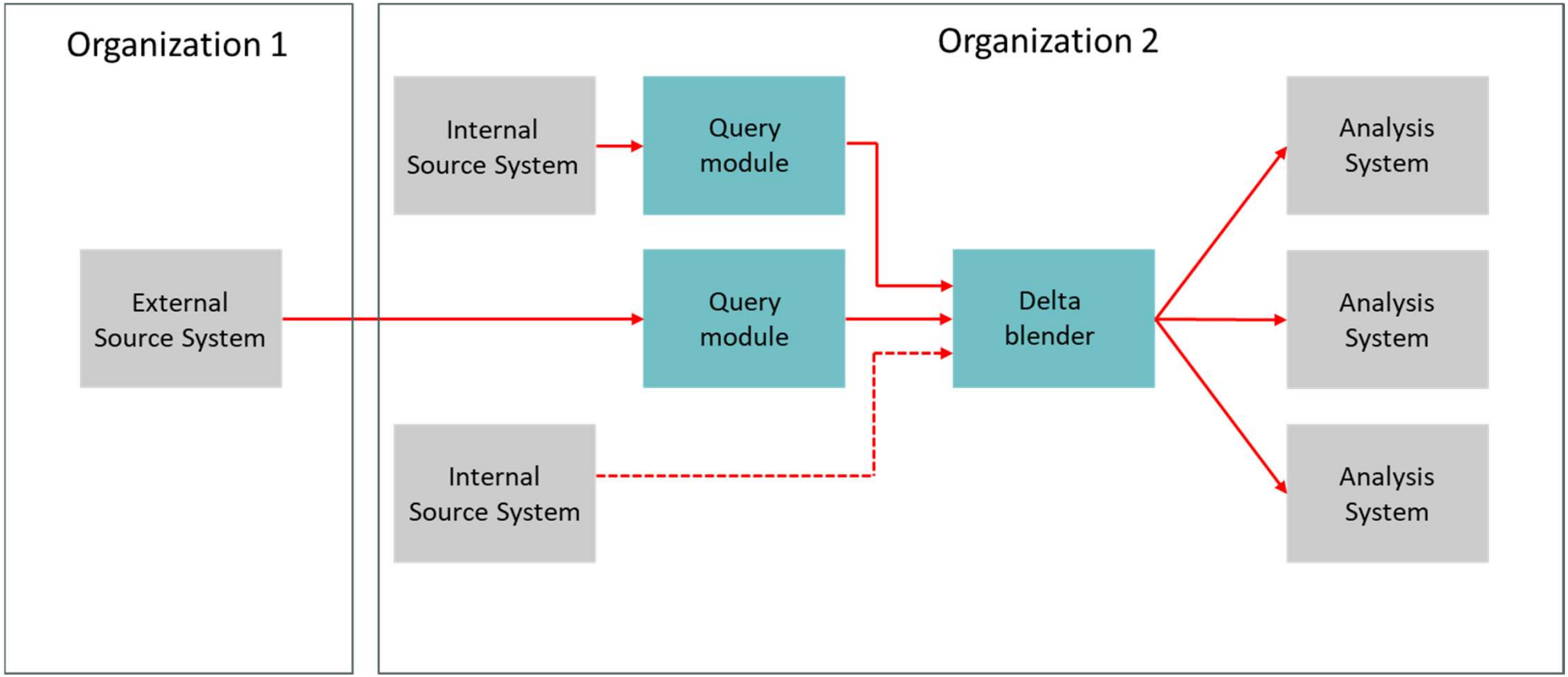
Delta Blender: Produces Ready-to-use Data



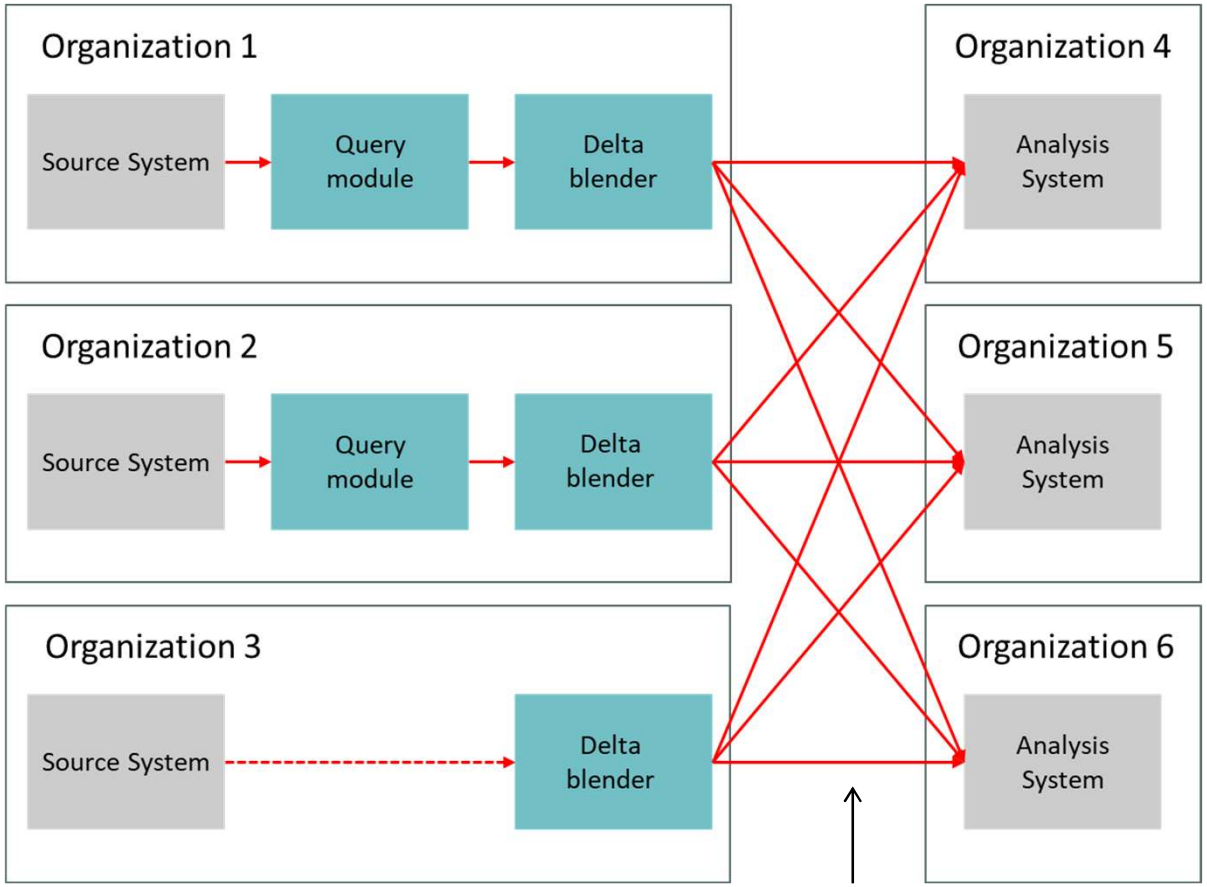
Delta Blender: Transformers



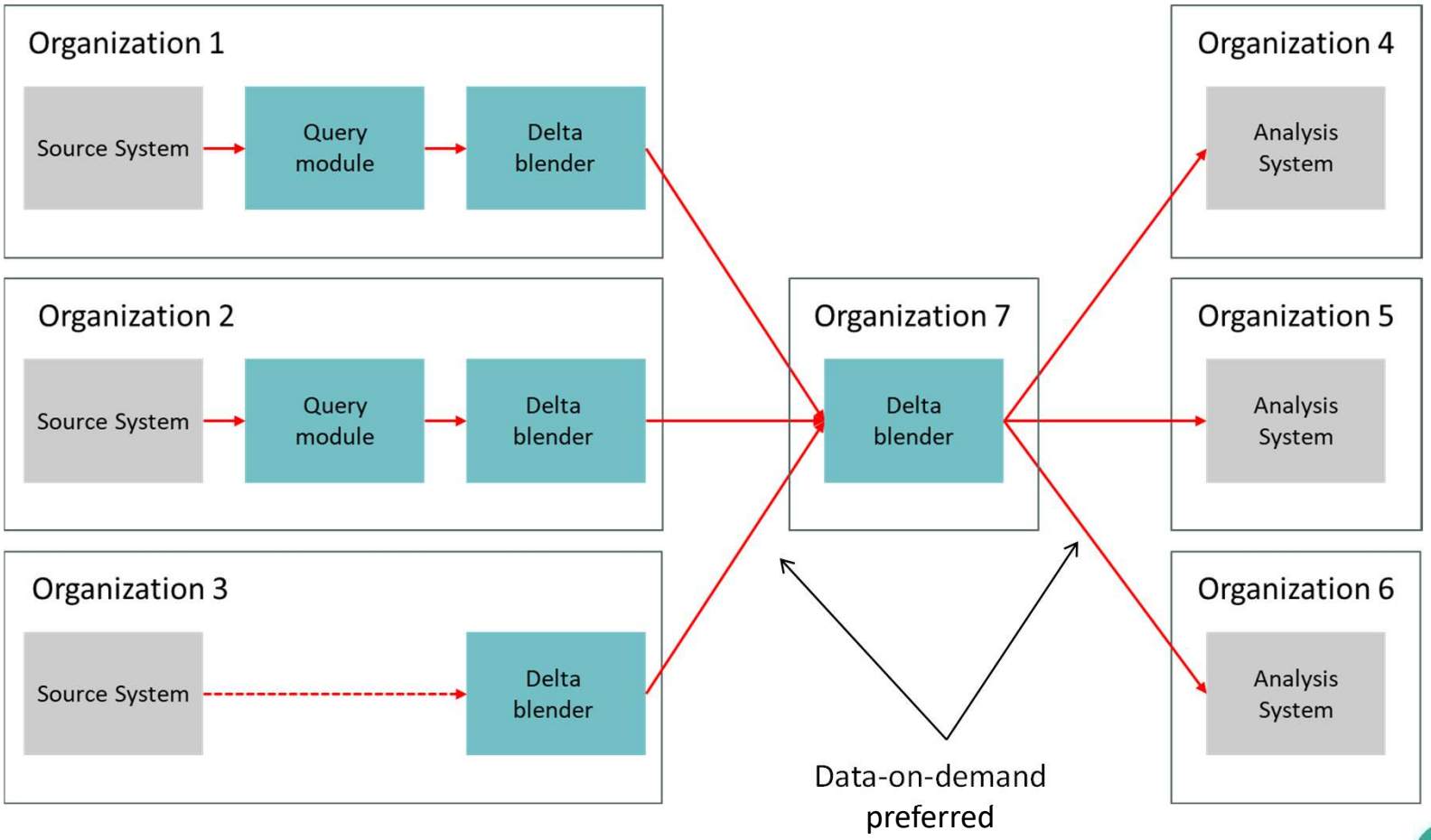
Local Federated Data Architecture



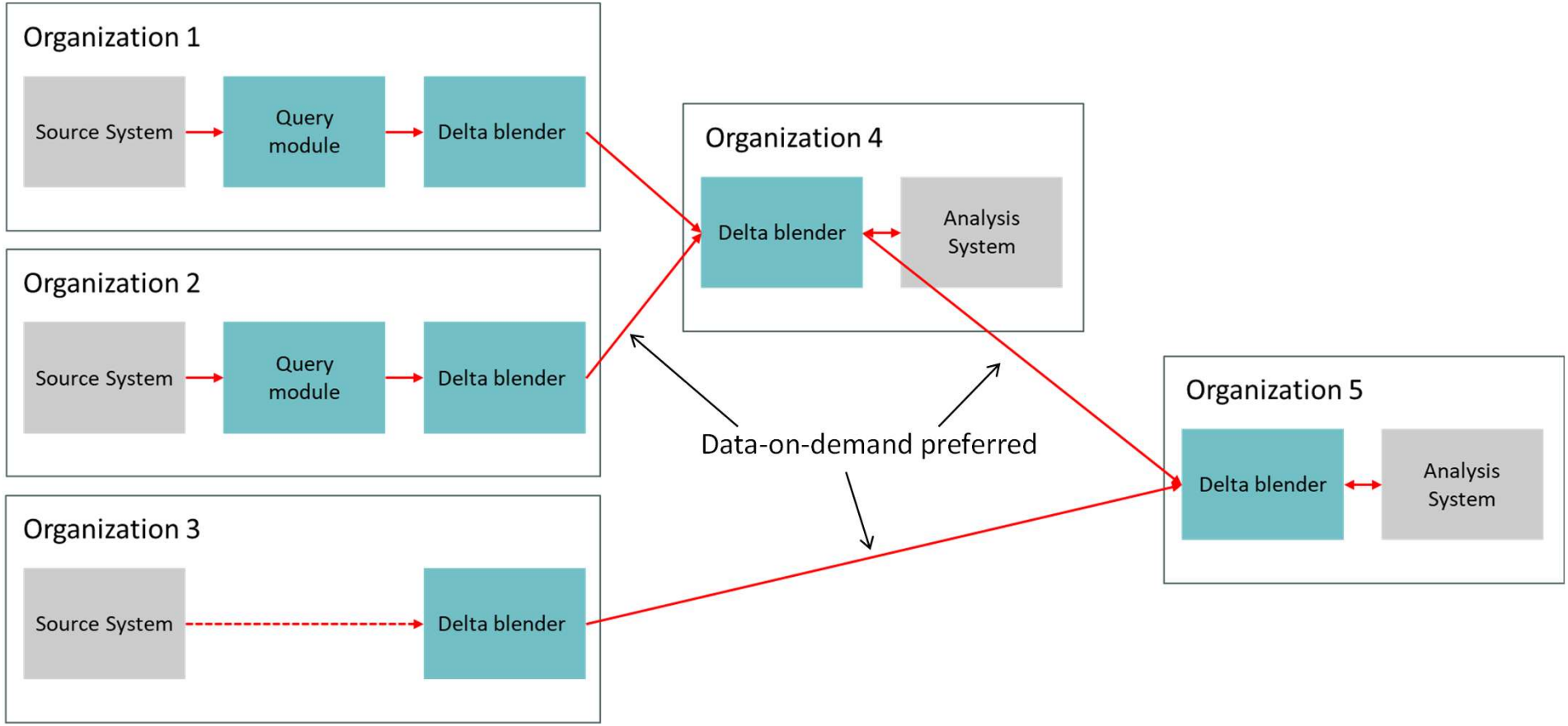
No Federated Data Architecture



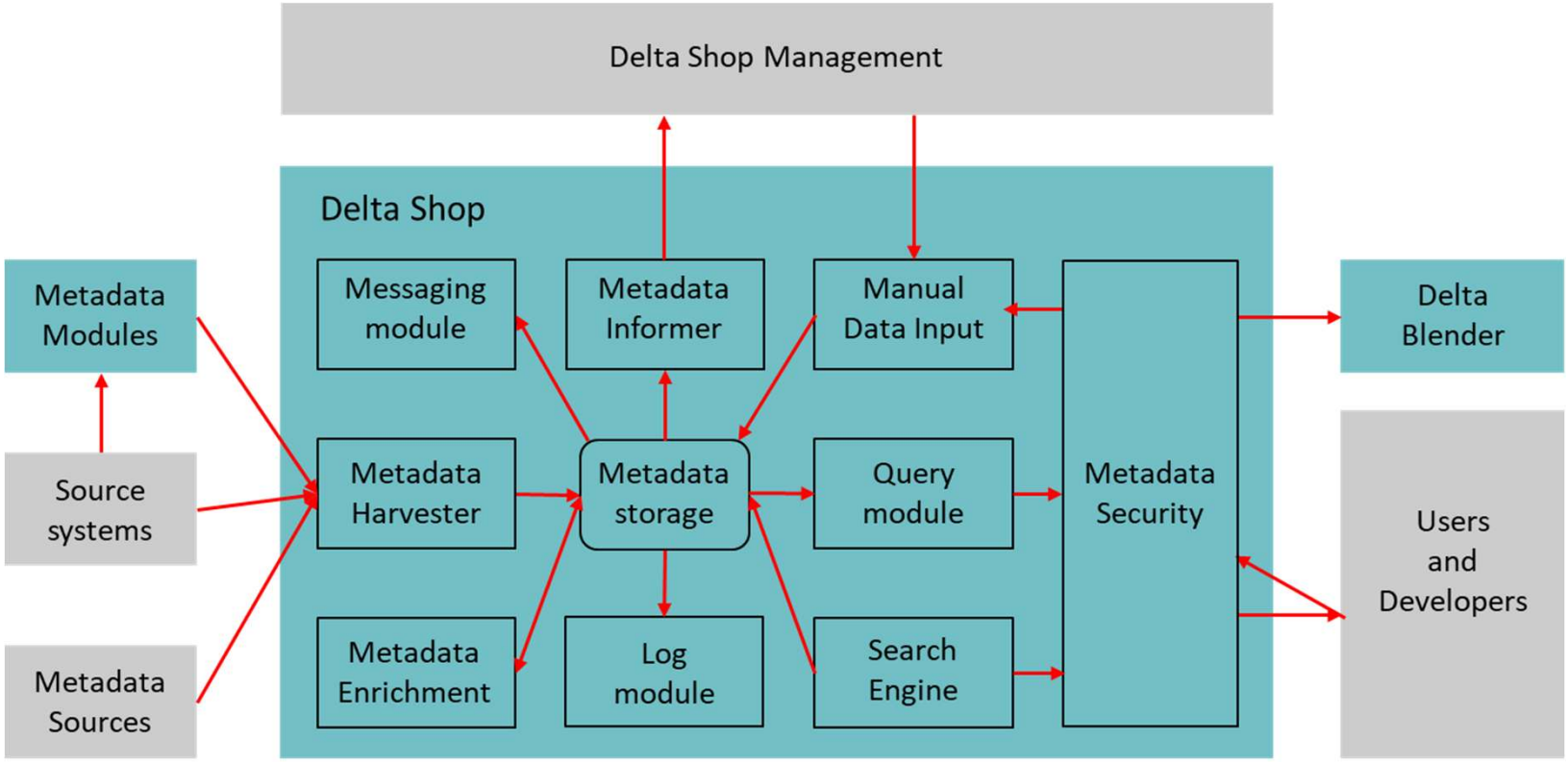
Global Federated Data Architecture



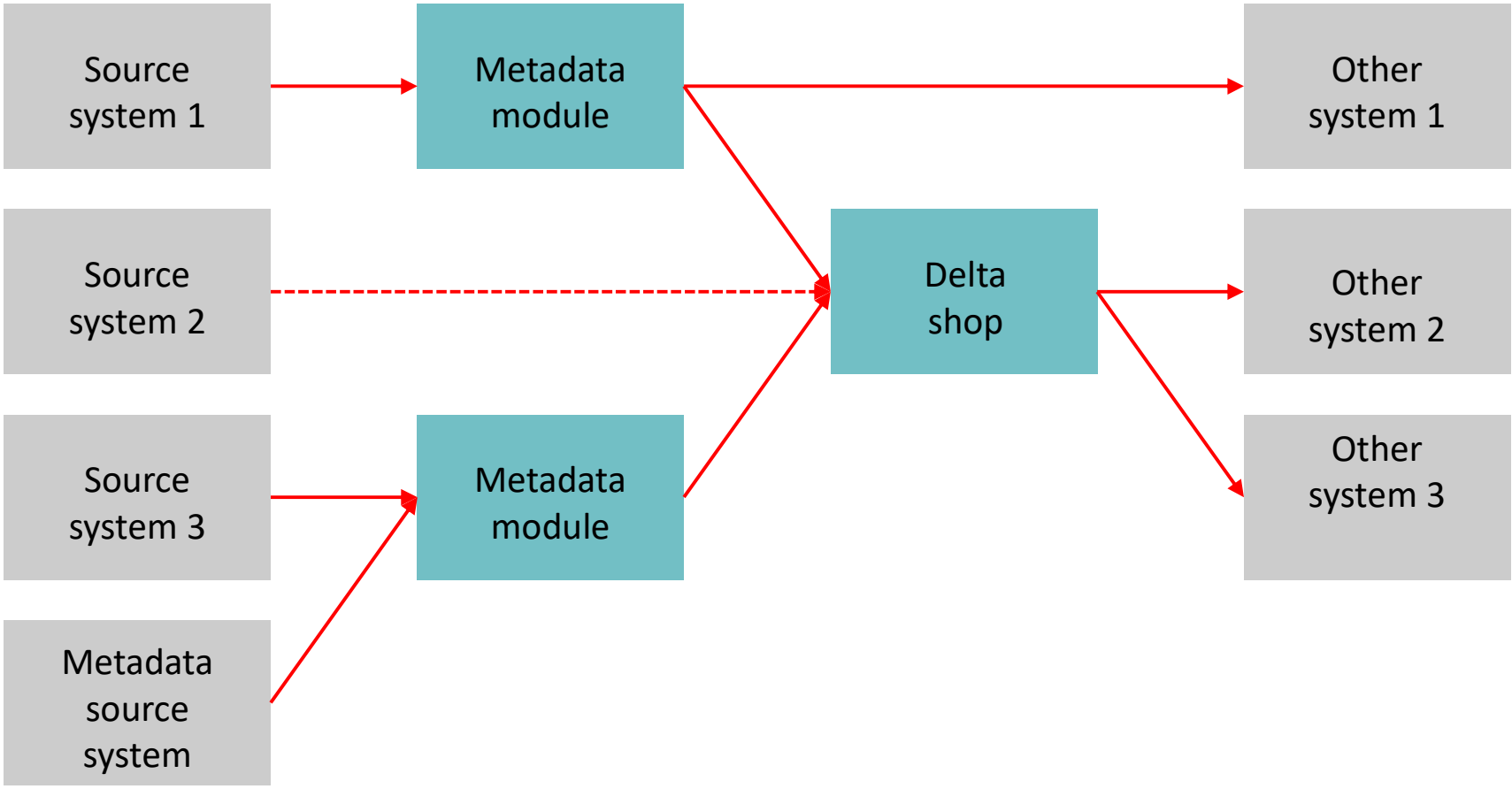
Layered Globale Federated Data Architecture



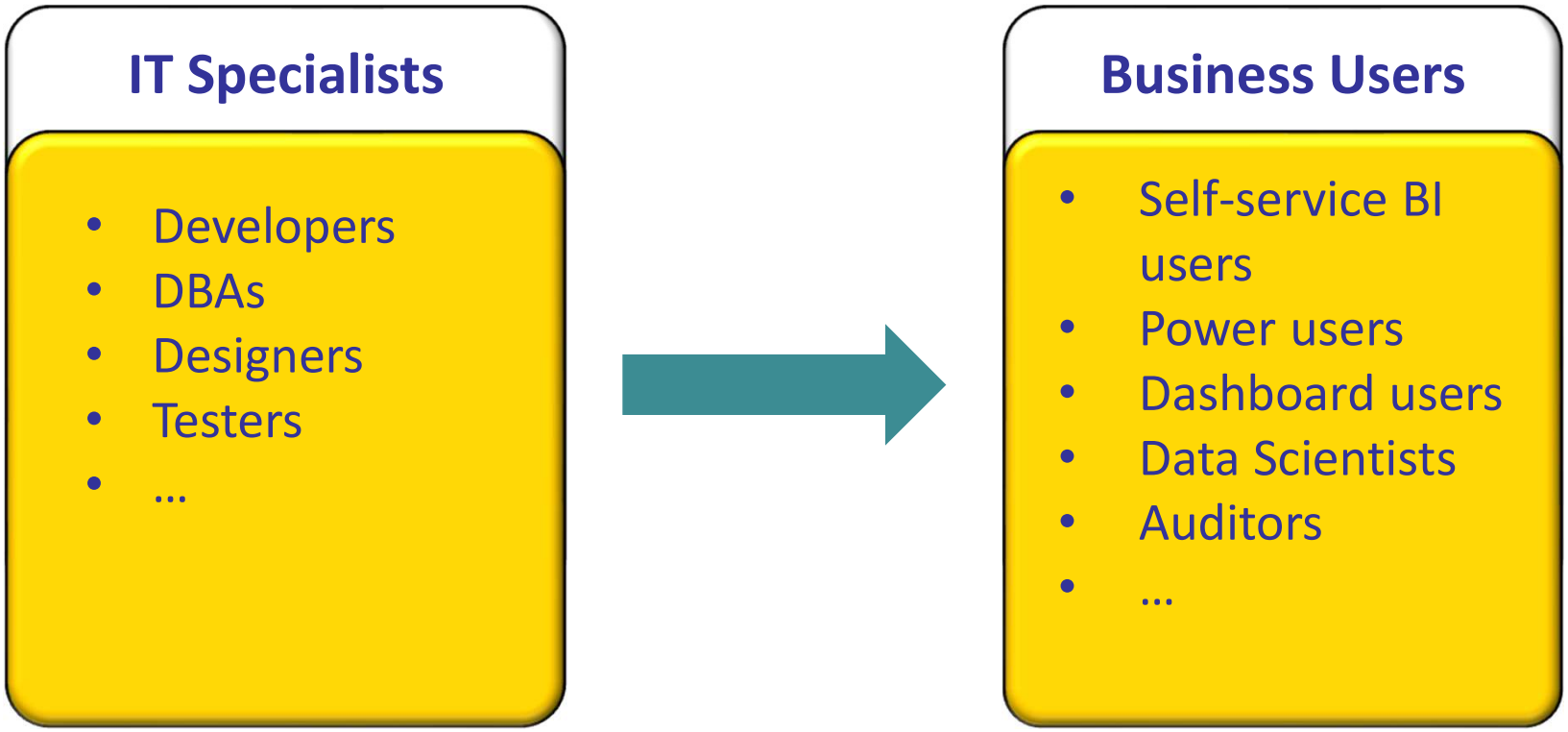
Delta Shop Components - Metadata



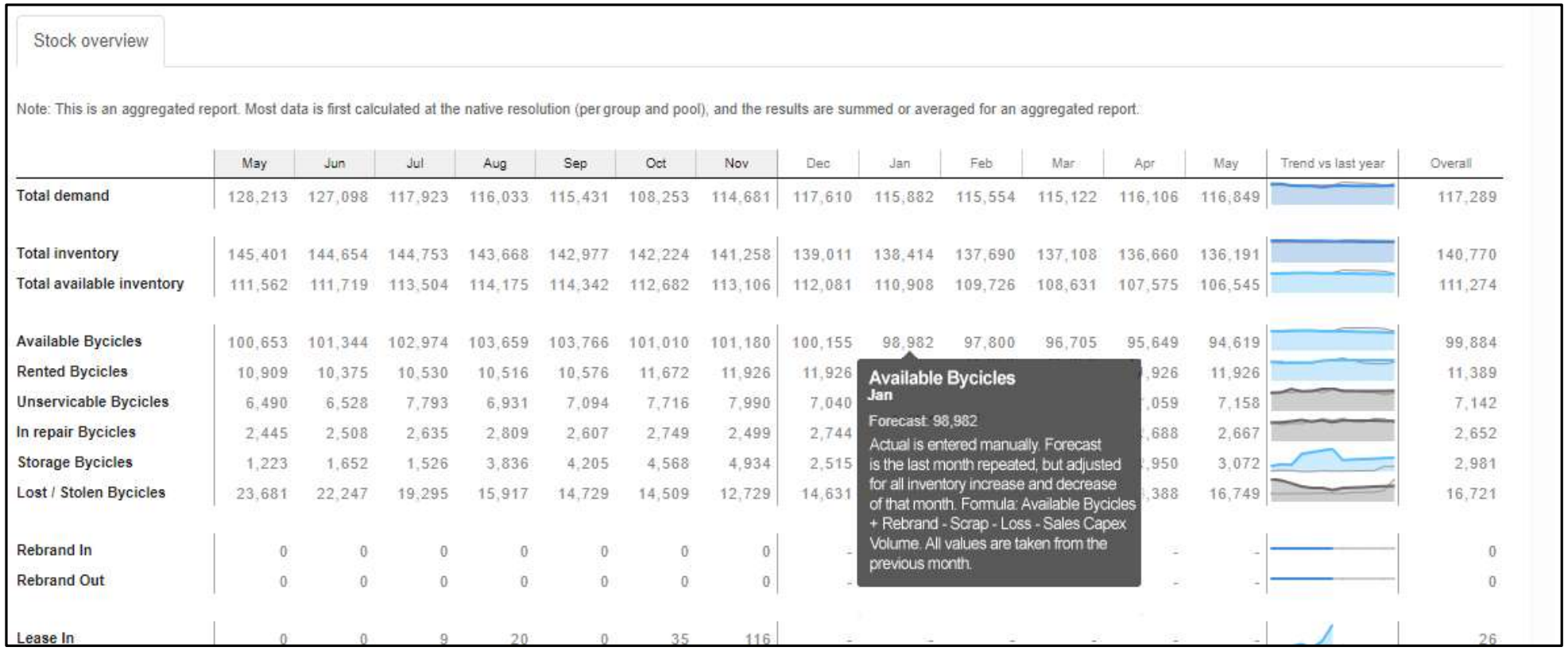
Metadata Modules and the Delta Shop



Target Audience of Metadata Has Changed



Instant Metadata



Stock overview

Note: This is an aggregated report. Most data is first calculated at the native resolution (per group and pool), and the results are summed or averaged for an aggregated report.

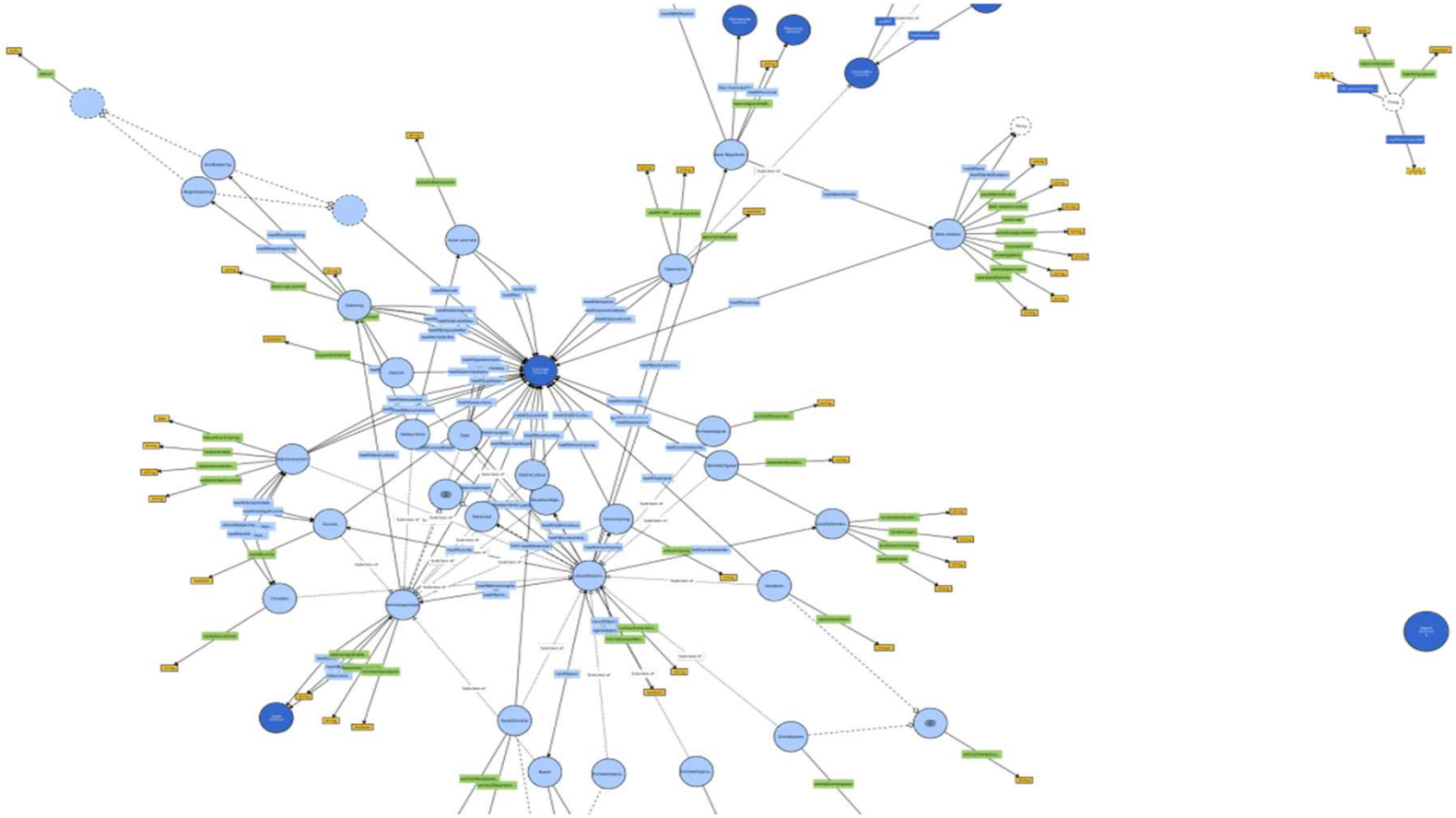
	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Trend vs last year	Overall
Total demand	128,213	127,098	117,923	116,033	115,431	108,253	114,681	117,610	115,882	115,554	115,122	116,106	116,849		117,289
Total inventory	145,401	144,654	144,753	143,668	142,977	142,224	141,258	139,011	138,414	137,690	137,108	136,660	136,191		140,770
Total available inventory	111,562	111,719	113,504	114,175	114,342	112,682	113,106	112,081	110,908	109,726	108,631	107,575	106,545		111,274
Available Bycicles	100,653	101,344	102,974	103,659	103,766	101,010	101,180	100,155	98,982	97,800	96,705	95,649	94,619		99,884
Rented Bycicles	10,909	10,375	10,530	10,516	10,576	11,672	11,926	11,926				,926	11,926		11,389
Unservicable Bycicles	6,490	6,528	7,793	6,931	7,094	7,716	7,990	7,040				,059	7,158		7,142
In repair Bycicles	2,445	2,508	2,635	2,809	2,607	2,749	2,499	2,744				,688	2,667		2,652
Storage Bycicles	1,223	1,652	1,526	3,836	4,205	4,568	4,934	2,515				,950	3,072		2,981
Lost / Stolen Bycicles	23,681	22,247	19,295	15,917	14,729	14,509	12,729	14,631				,388	16,749		16,721
Rebrand In	0	0	0	0	0	0	0	-				-	-		0
Rebrand Out	0	0	0	0	0	0	0	-				-	-		0
Lease In	0	0	9	20	0	35	116	-	-	-	-	-	-		26

Available Bycicles
Jan

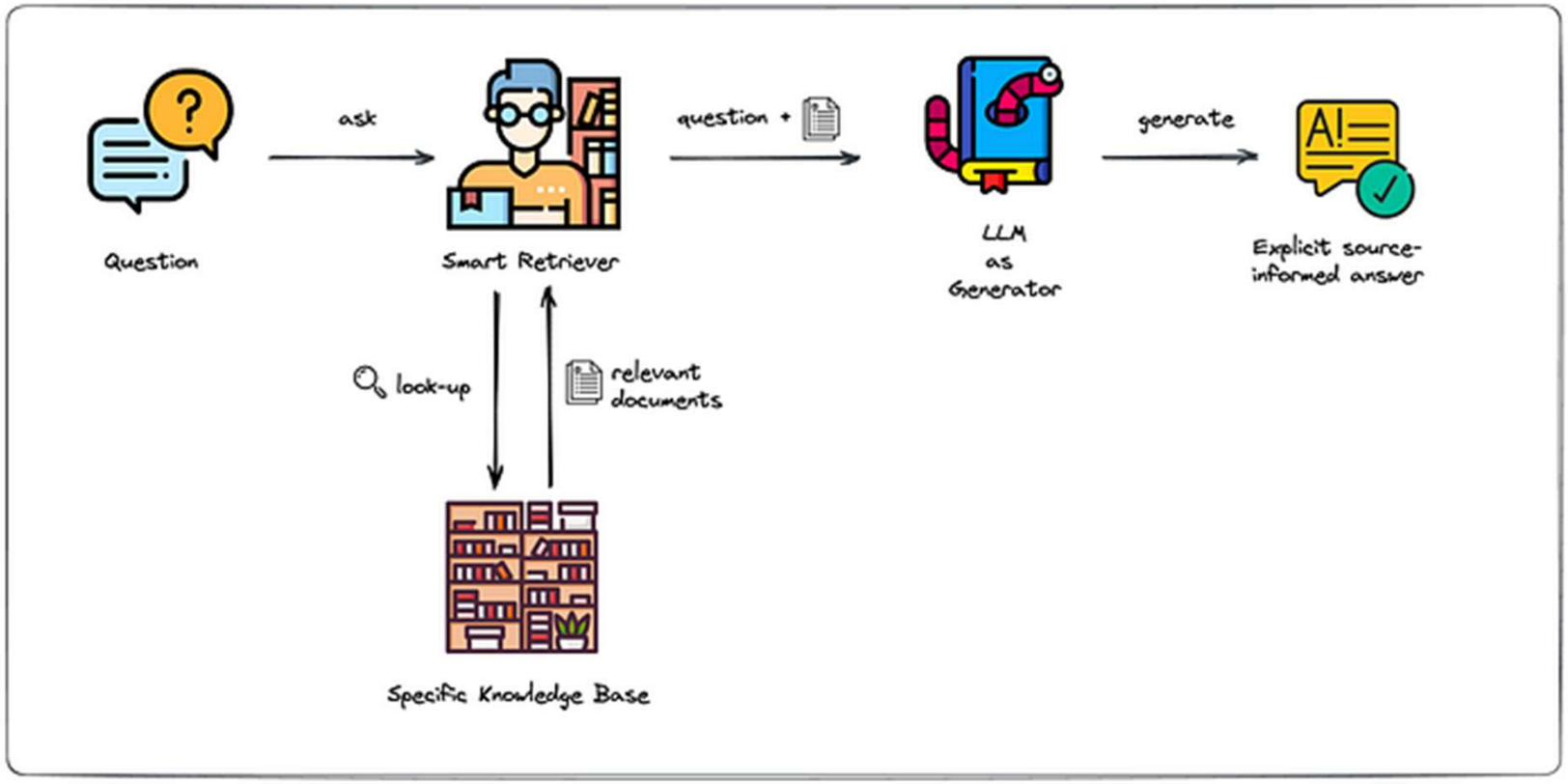
Forecast: 98,982

Actual is entered manually. Forecast is the last month repeated, but adjusted for all inventory increase and decrease of that month. Formula: Available Bycicles + Rebrand - Scrap - Loss - Sales Capex Volume. All values are taken from the previous month.

Generating an Onthology (for AI?)



Retrieval Augmented Generation



Summary: A Modern Data Architecture is Holistic

- Enterprise data architecture
  - From source to insight
  - Data must be ready-to-use, discoverable, and insightful
  - It's about the transformers, not the databases
  - Metadata for everyone
- Federated data architecture
  - Unbridled creation of data copies must stop
  - Data-on-demand, not data-by-copy
  - Avoid duplicate transformers
- Transparent data architecture
  - Operational lineage for reconstruction, transparency, and auditability
- Modular data architecture
  - Co-exist with existing solutions


De Delta
data-architectuur:
van bron tot inzicht
Rick F. van der Lans