mimacom

Modern Data Engineering 101

Von Data Warehouse bis Data Mesh

Agenda

Introduction

OLTP vs. OLAP

Data Storage Paradigms

Data Architectures

Choosing the right approach

Apache Spark

Julia Riedel

Background

- Data Engineer
- Specialized on Azure / Databricks
- In D&A for 3,5 years

At Mimacom

- Design and Implement new Data Platforms for Customers

Thomas Konstantinides

Background

- In the IT since 2002
- Backend / frontend / tech lead / hands-on architect
- Now focus on data engineering

At Mimacom

- Part of a customer's data lakehouse team
- Support transition to a data mesh and development of the governance layer

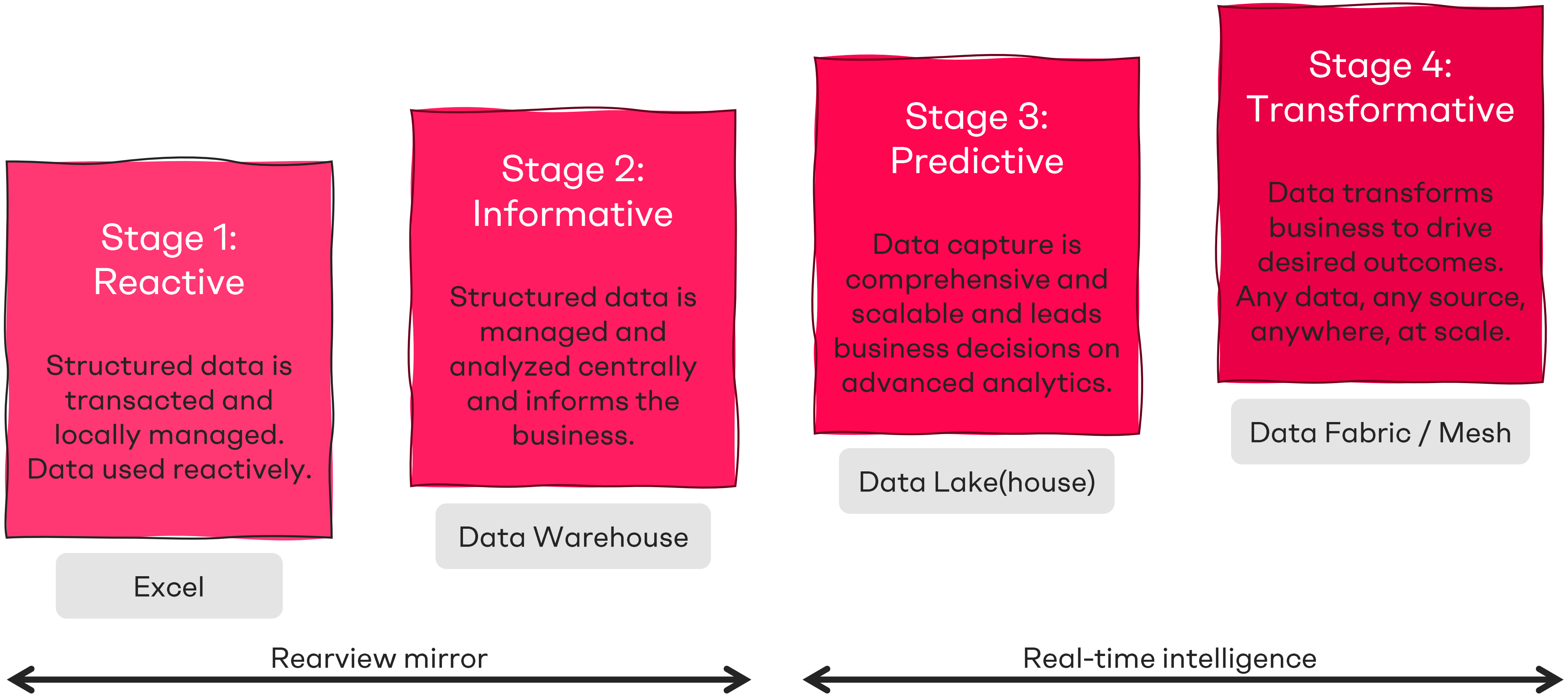
Data Engineering <-> Java Forum

Why should we care?

- Software Development is changing
- Data is becoming more important
- Newer architectures bring data and software development closer together
- A “normal” software developer product team will often also take care about data topics

The Goal

Becoming a data-driven company



Stage 1: Reactive

Structured data is transacted and locally managed. Data used reactively.

Excel

Stage 2: Informative

Structured data is managed and analyzed centrally and informs the business.

Data Warehouse

Stage 3: Predictive

Data capture is comprehensive and scalable and leads business decisions on advanced analytics.

Data Lake(house)

Stage 4: Transformative

Data transforms business to drive desired outcomes. Any data, any source, anywhere, at scale.

Data Fabric / Mesh

Rearview mirror

Real-time intelligence

Source: Enterprise data maturity stages from "Deciphering Data Architectures" by James Serra

OLTP vs. OLAP

- What’s the difference?
- Why process the data?
- ETL: From operational data to analytical data

OLAP vs. OLTP

What’s the difference?

OLTP

Online Transactional Processing

- Transactional
- Normalized
- Simple Queries (Read, Insert, Update)
- Current Data

OLAP

Online Analytical Processing

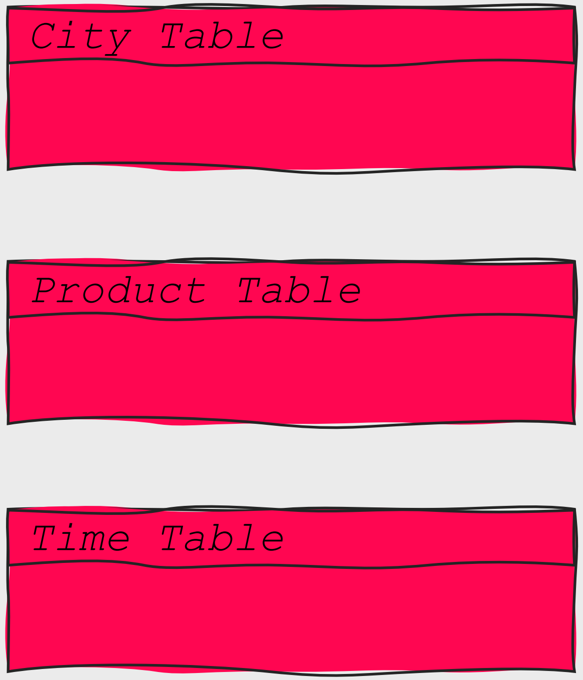
- Analytical
- Denormalized
- Complex Queries (joins and aggregation)
- Historical Data

OLAP vs. OLTP

What’s the difference?

OLTP

- Transactional
- Normalized
- Simple Queries (Read, Insert, Update)
- Current Data



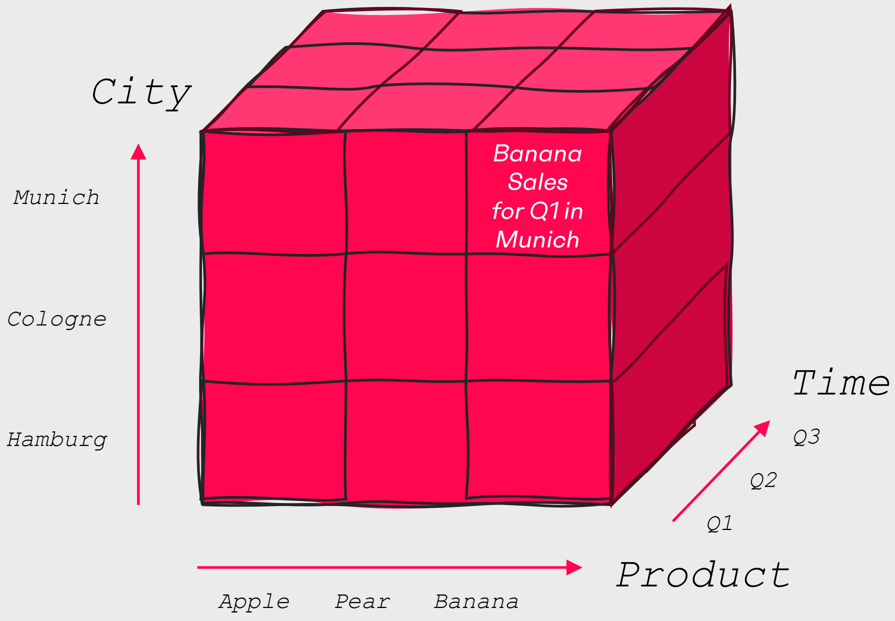
City Table

Product Table

Time Table

OLAP vs. OLTP

What’s the difference?



OLAP

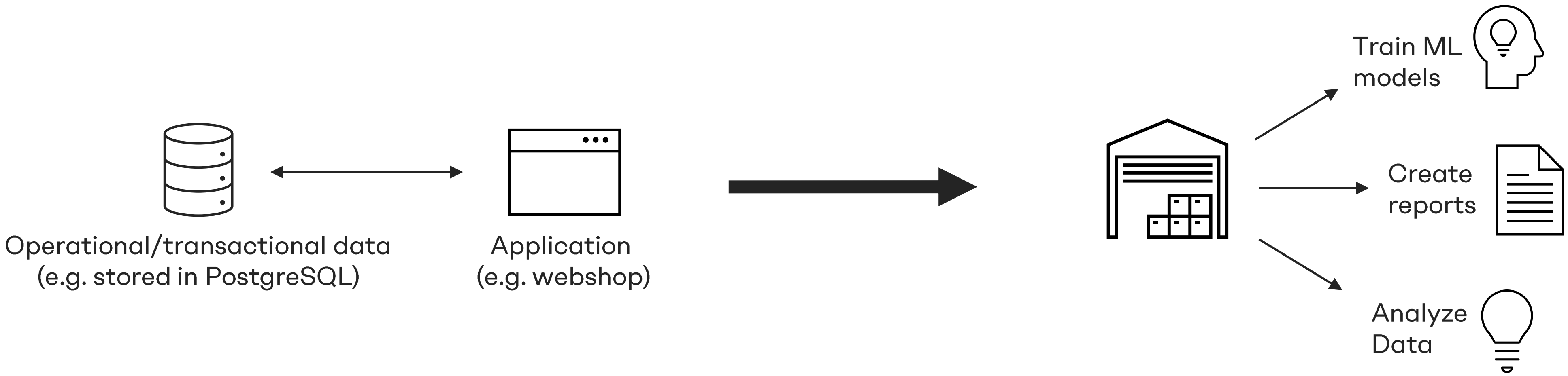
- Analytical
- Denormalized
- Complex Queries (joins and aggregation)
- Historical Data

OLAP vs. OLTP

Why process the data?

OLTP

OLAP



Operational/transactional data
(e.g. stored in PostgreSQL)

Application
(e.g. webshop)

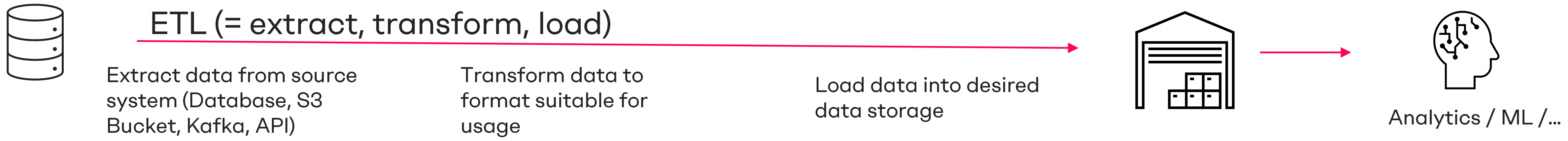
Train ML
models

Create
reports

Analyze
Data

OLAP vs. OLTP

ETL: From operational data to analytical data



ETL (= extract, transform, load)

Extract data from source system (Database, S3 Bucket, Kafka, API)

Transform data to format suitable for usage

Load data into desired data storage

Analytics / ML /…

ELT (= extract, load, transform)

Data Storage Paradigms

- Data Warehouse
- Data Lake
- Data Lakehouse
- Technologies

Data Warehouse

Goal: Support decision making process and reporting & visualizations

- Schema-on-write (ETL)
- Relational schema
- Structured data

Advantages / Disadvantages

○ Optimized for downstream BI consumption
○ Pay for the peak of user load
○ No support for unstructured data
○ Limited use-case Support

Data Lake

Goal: provide a cheap storage for data

- Schema-on-Read (ELT)
- Structured and unstructured data
- Data in generic and open file formats
- Often combined with a data warehouse

Advantages / Disadvantages

○ Low-cost storage systems with file API

○ Lack of basic management features

○ When used with an additional DWH double costs for storage

○ It’s hard to use the data in the lake

Data Storage Paradigms

Data Lakehouse

Goal: combine advantages of Data Warehouse and Data Lake

- Open direct-access data formats
- Open table formats like Delta Lake
- First-class support for machine learning and data science workloads

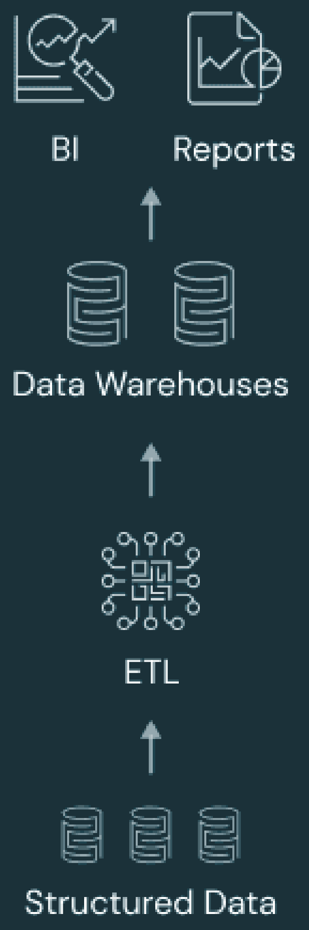
Advantages / Disadvantages

○ Performance and management features of data warehouses

○ Fast, direct I/O for advanced analytics workloads

○ Less performance than DWH for some use cases

Data Warehouse



BI

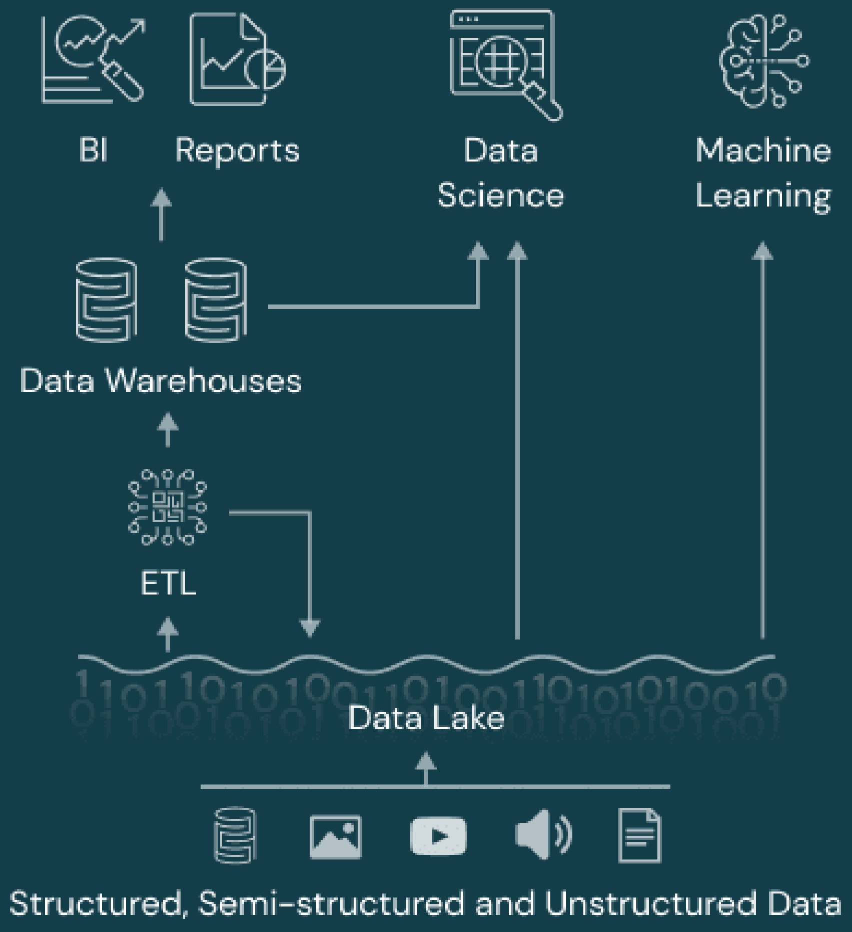
Reports

Data Warehouses

ETL

Structured Data

Data Lake



BI

Reports

Data Science

Machine Learning

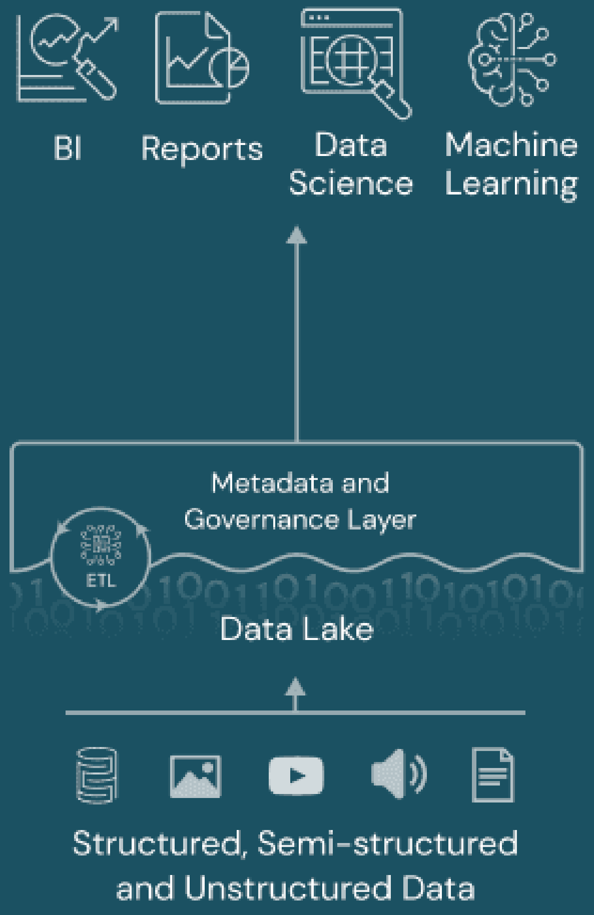
Data Warehouses

ETL

Data Lake

Structured, Semi-structured and Unstructured Data

Data Lakehouse



BI

Reports

Data Science

Machine Learning

Metadata and Governance Layer

ETL

Data Lake

Structured, Semi-structured and Unstructured Data

Quelle: https://www.databricks.com/de/glossary/data-lakehouse

Technologies

APACHE Spark™

Multi language engine for data engineering

- Supports Scala, Java, Python and R
- Can make use of SQL
- Integrates with most relevant frameworks and formats

Azure Synapse Analytics

databricks

amazon S3

snowflake®

Microsoft® SQL Server®

Data Storage Paradigms

Technologies

Azure Synapse Analytics

Azure ELT / Lakehouse Solution

- ELT Pipelines
- Serverless / Dedicated SQL Pools
- Read/Write to Data Lake

APACHE Spark

databricks

amazon S3

snowflake

Microsoft SQL Server

Data Storage Paradigms

Technologies

databricks

Lakehouse Solution

- Data transformations via Spark
- Read/Write to Lakehouse
- Read/Write to Data Lake
- Combination with Azure Synapse possible

APACHE Spark™ Azure Synapse Analytics amazon S3 snowflake® Microsoft® SQL Server®

Data Storage Paradigms

Technologies

amazon S3

Data Lake Solution

- Combination with Azure Synapse, AWS Glue possible
- Storage for Data Lake or Lakehouse

APACHE Spark

Azure Synapse Analytics

databricks

snowflake

Microsoft SQL Server

Data Storage Paradigms

Technologies

snowflake

Lakehouse Solution

- Data transformations via SQL
- All on one solution or only DWH
- Combination with Azure Synapse possible

APACHE Spark

Azure Synapse Analytics

databricks

amazon S3

Microsoft SQL Server

Technologies

Microsoft® SQL Server®

DWH Solution

- On-prem
- SQL Server Integration Services (SSIS) as ETL Solution
- traditional solution

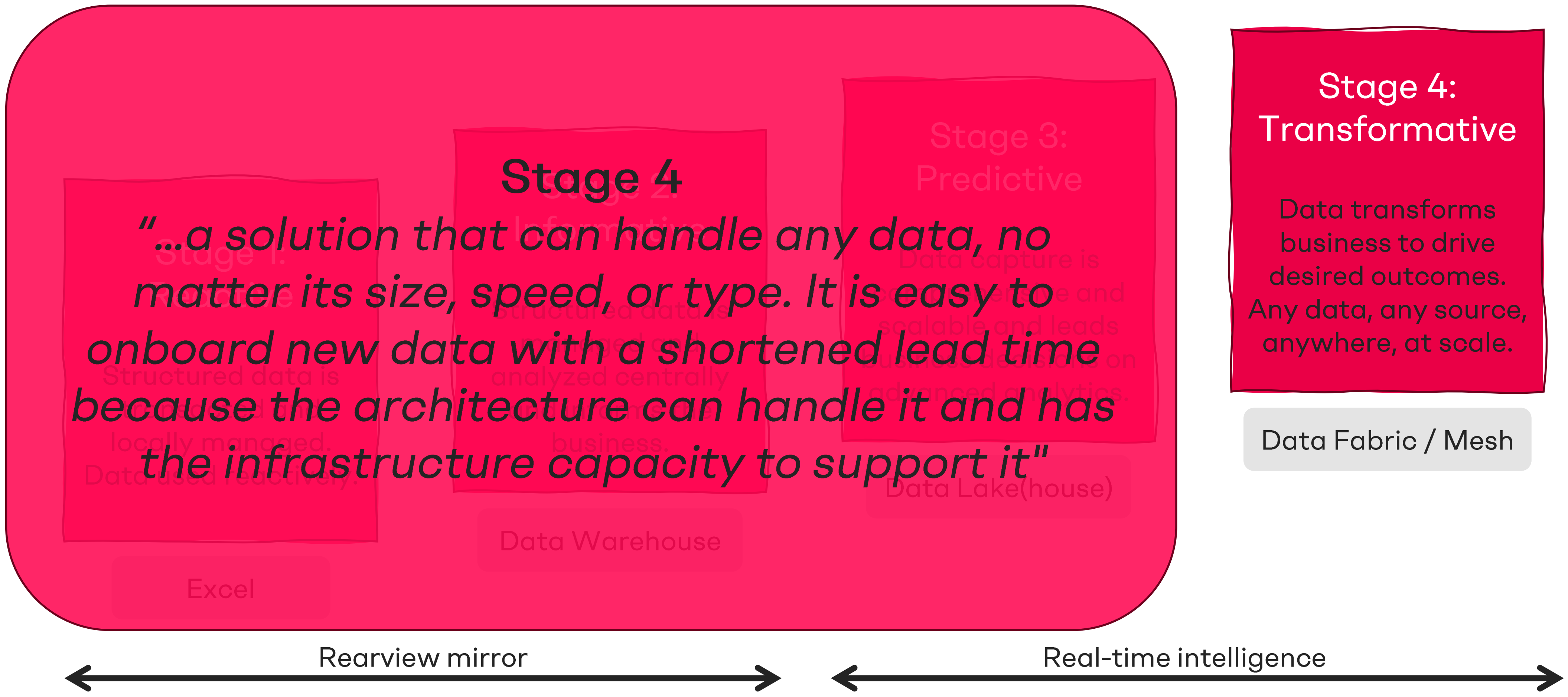
Apache Spark™ Azure Synapse Analytics databricks amazon S3 snowflake®

Data Architecture Patterns

- Data Fabric
- Data Mesh

Data Architecture Patterns

Why more?



Stage 4

"...a solution that can handle any data, no matter its size, speed, or type. It is easy to onboard new data with a shortened lead time because the architecture can handle it and has the infrastructure capacity to support it"

Stage 4: Transformative

Data transforms business to drive desired outcomes. Any data, any source, anywhere, at scale.

Data Fabric / Mesh

Enterprise data maturity stages from "Deciphering Data Architectures" by James Serra

Data Architecture Patterns

Scalability issues

- Source data stored in different clouds and on-premise
- Different ways to access data sources
- Integrating new data sources takes long
- Hard to find needed data and know who can grant access
- Governance processes can create bottlenecks
- Data Warehouses cannot serve all use cases
- Data Lake can become "Data Swamp"

Data Integration

Data Governance

Data Democratization

Scalability

Image from https://medium.com/@armandovazquez/navigating-the-waters-designing-a-data-lake-to-avoid-the-murky-depths-of-a-data-swamp-d67f5600c27c

Data Architecture Patterns

Scalability issues

- Source data stored in different clouds and on-premise
- Different ways to access data sources
- Integrating new data sources takes long
- Hard to find needed data and kno ess
- Governance proc checks
- D not serve all use cases
- ake can become “Data Swamp”

Decreases business agility

Data Integration

Data Governance

Data Democratization

Scalability

Image from https://medium.com/@armandovazquez/navigating-the-waters-designing-a-data-lake-to-avoid-the-murky-depths-of-a-data-swamp-d67f5600c27c

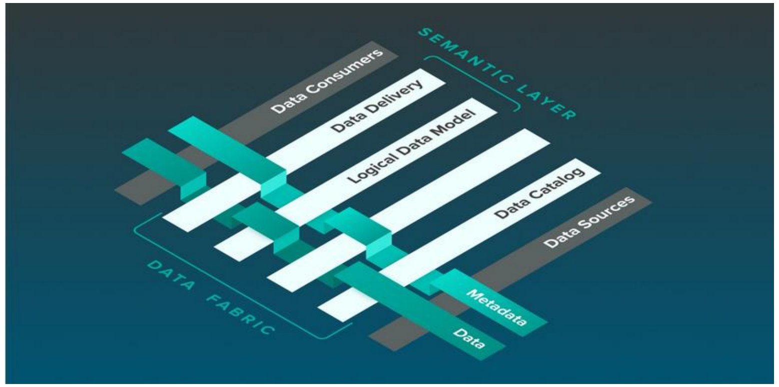
Data Fabric

Data Architecture Patterns

What is data fabric?

Data fabric is an architecture that facilitates the end-to-end integration of various data pipelines and cloud environments through the use of intelligent and automated systems.

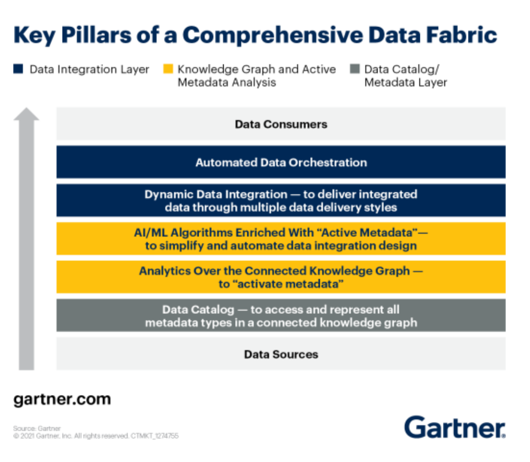
From https://www.ibm.com/topics/data-fabric



From https://www.atscale.com/blog/what-is-a-data-fabric/

…a design concept that serves as an integrated layer (fabric) of data and connecting processes. A data fabric utilizes continuous analytics over existing, discoverable and inferenced metadata assets to support the design, deployment and utilization of integrated and reusable data across all environments, including hybrid and multi-cloud platforms.

From https://www.gartner.com/smarterwithgartner/data-fabric-architecture-is-key-to-modernizing-data-management-and-integration



Key Pillars of a Comprehensive Data Fabric

■ Data Integration Layer ■ Knowledge Graph and Active Metadata Analysis ■ Data Catalog/ Metadata Layer

Data Consumers

Automated Data Orchestration

Dynamic Data Integration — to deliver integrated data through multiple data delivery styles

AI/ML Algorithms Enriched With "Active Metadata"— to simplify and automate data integration design

Analytics Over the Connected Knowledge Graph — to "activate metadata"

Data Catalog — to access and represent all metadata types in a connected knowledge graph

Data Sources

gartner.com

Source: Gartner
© 2021 Gartner, Inc. All rights reserved. CTMKT_1274755

Gartner

Data Architecture Patterns

Takeaways

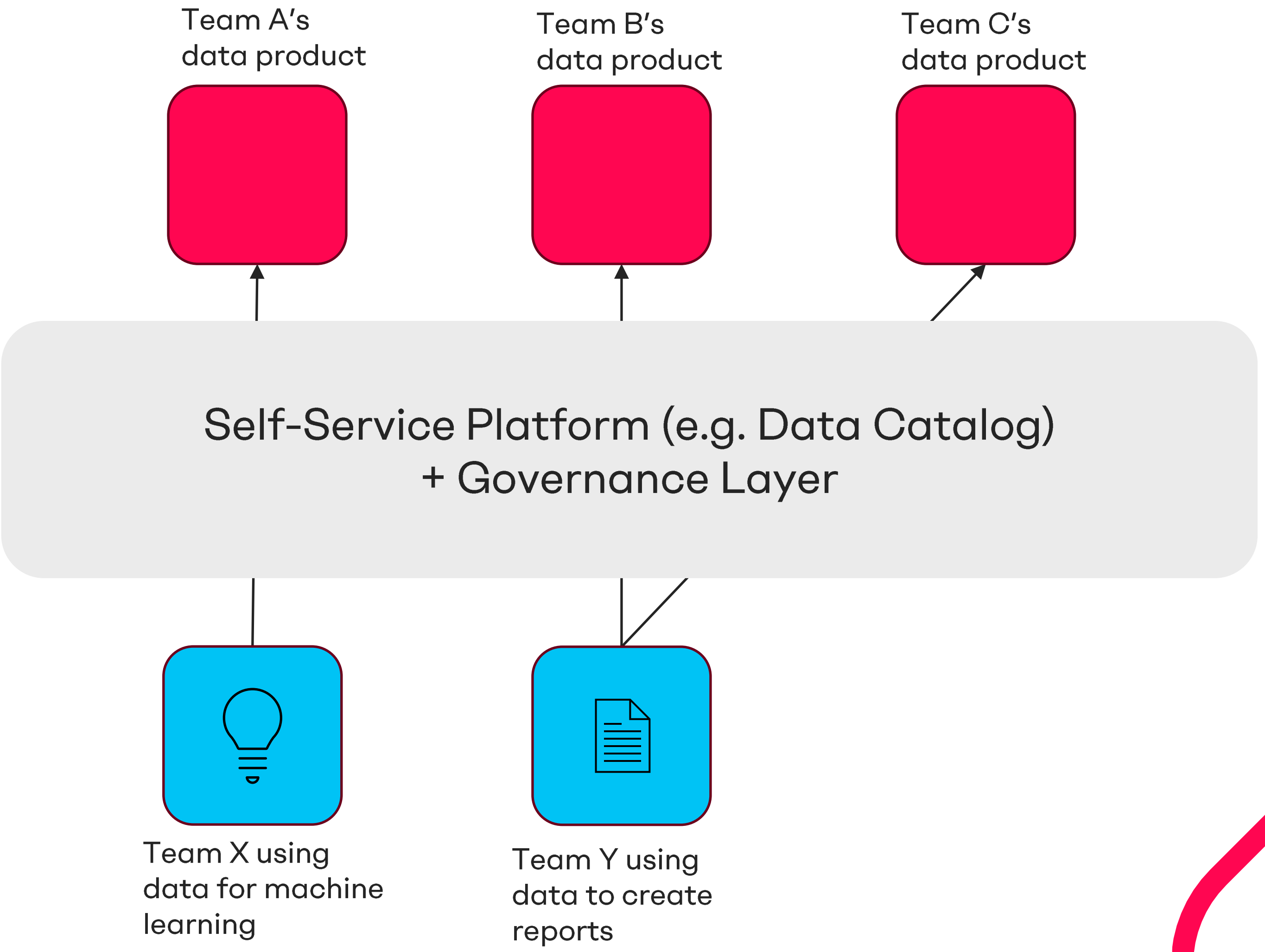
It needs more than just storing data and processing it

Seamless integration and governance across the entire data landscape

- Integration over different data sources
- Data virtualization
- Unifying data access over APIs
- Enabling data discovery
- Meta data catalog
- Centralized governance and security
- Real-time support

Data Mesh

Data Mesh idea


Team A's data product
Team B's data product
Team C's data product
Self-Service Platform (e.g. Data Catalog) + Governance Layer
Team X using data for machine learning
Team Y using data to create reports

Data Architecture Patterns

Addressed issues

Ownership: The source team provides the data product

- Domain-driven, decentralized ownership
- Management of data products within domain-specific teams

Quality: The source team knows the data best

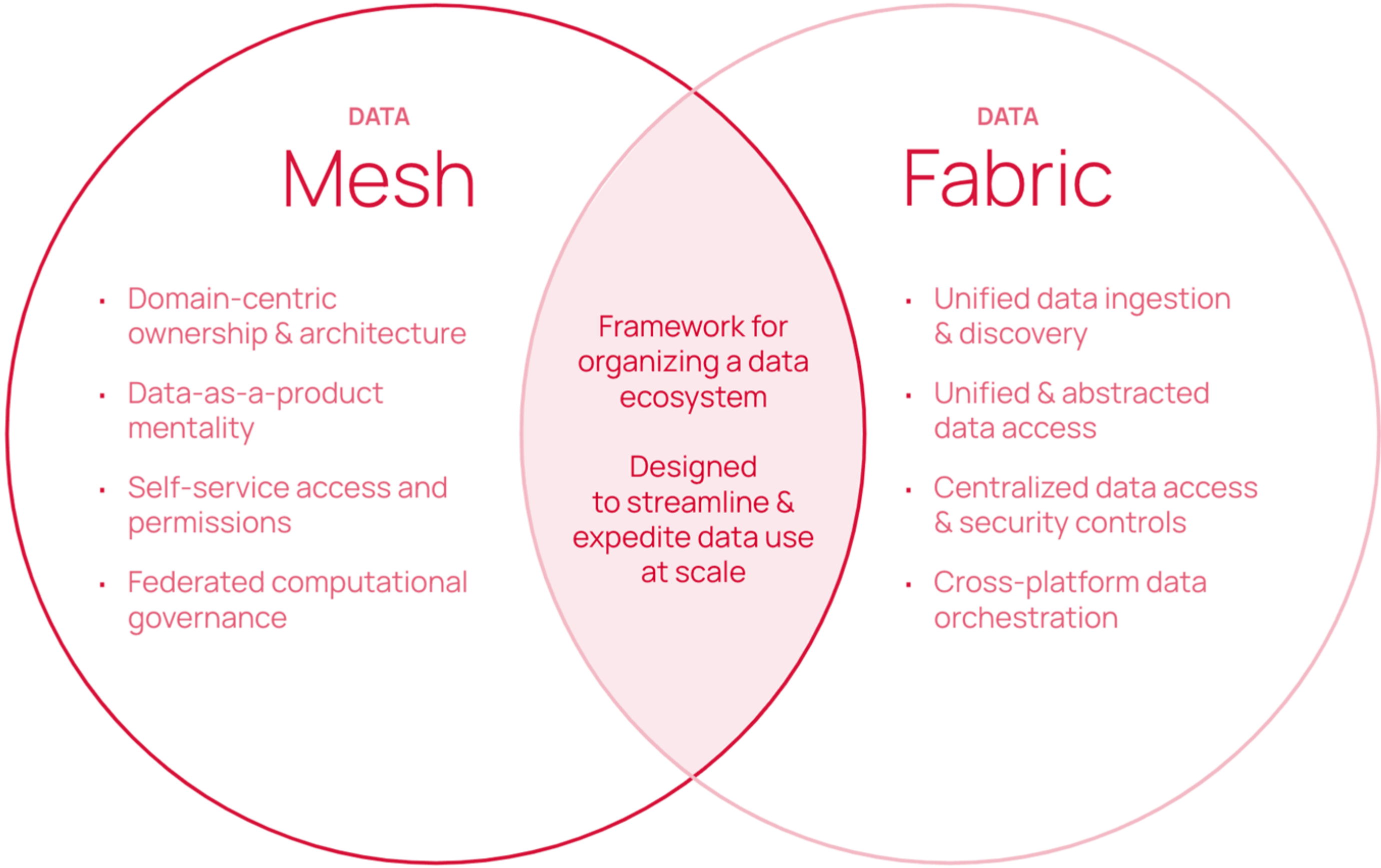
- Product thinking

Scaling: The central team can become a bottleneck

- Self-service platform
- Federated Governance

Data Architecture Patterns

Data Mesh vs. Data Fabric



DATA

Mesh

- Domain-centric ownership & architecture
- Data-as-a-product mentality
- Self-service access and permissions
- Federated computational governance

Framework for organizing a data ecosystem

Designed to streamline & expedite data use at scale

DATA

Fabric

- Unified data ingestion & discovery
- Unified & abstracted data access
- Centralized data access & security controls
- Cross-platform data orchestration

From https://www.immuta.com/blog/data-mesh-vs-data-fabric/

What should we use?

It depends (of course) 😁

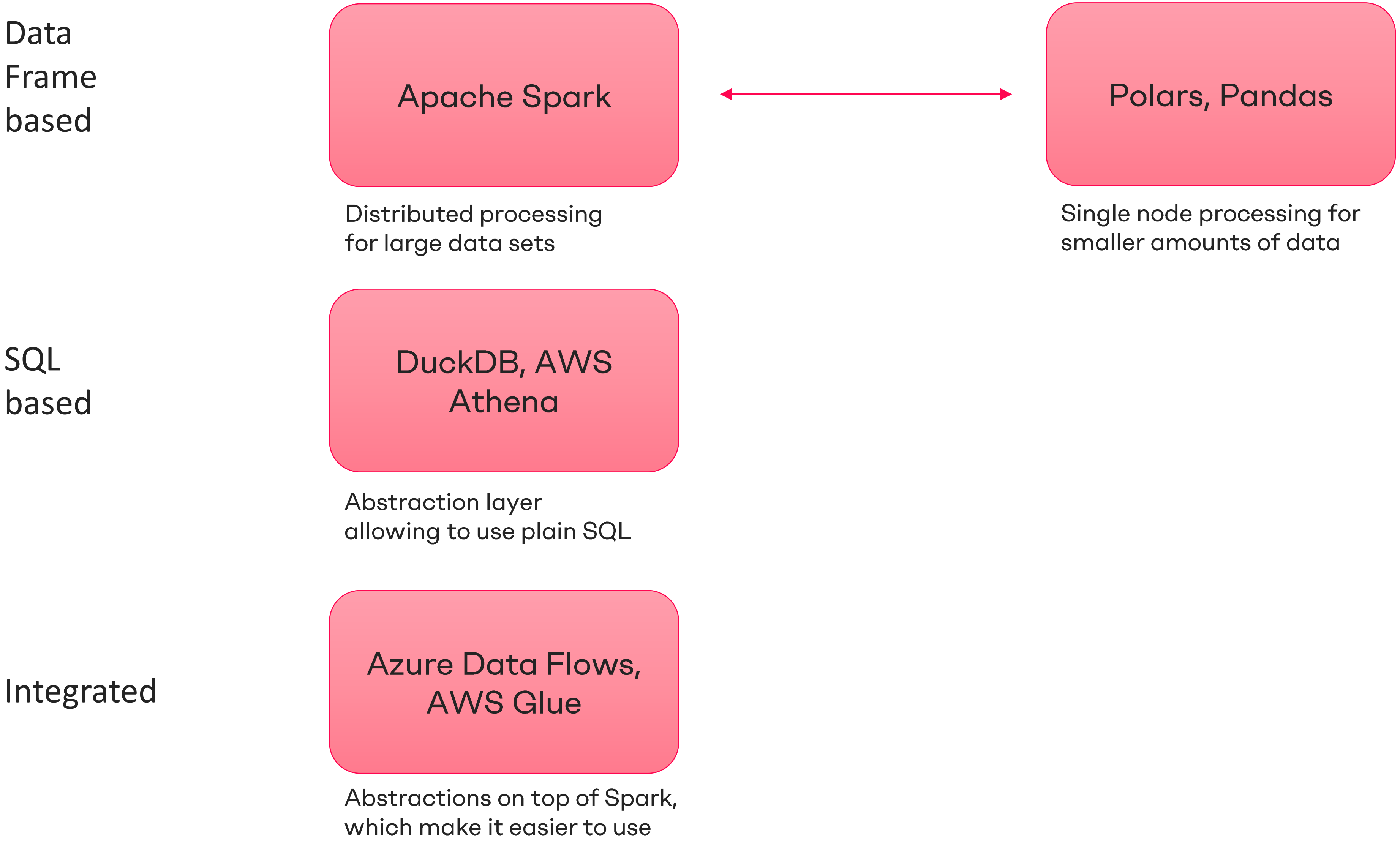
What should we use?

Topics to consider

- Current organizational data maturity level
- The 6 vs of (big) data processing: volume, velocity, variety, variability, veracity, value
- Structured/semistructured/unstructured/binary data
- Number of data sources
- Experience and amount of data engineers/analysts
- Short-term goals vs. strategic vision
- You are (most likely) not Facebook, Google, Microsoft or NASA

Apache Spark

Tools for ETL and Data Analysis



Data Frame based

Apache Spark

Distributed processing
for large data sets

Polars, Pandas

Single node processing for
smaller amounts of data

SQL based

DuckDB, AWS Athena

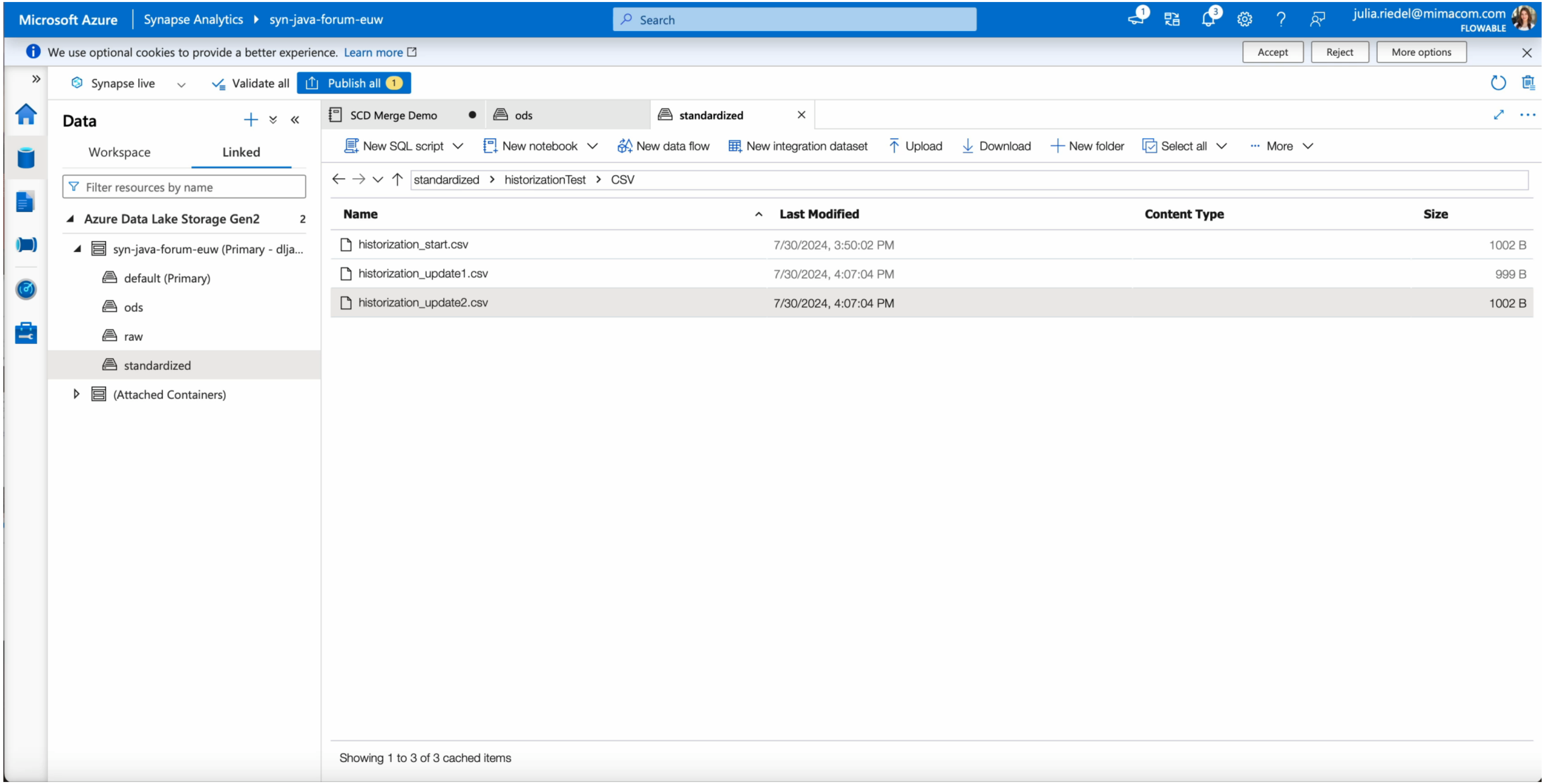
Abstraction layer
allowing to use plain SQL

Integrated

Azure Data Flows, AWS Glue

Abstractions on top of Spark,
which make it easier to use

Demo



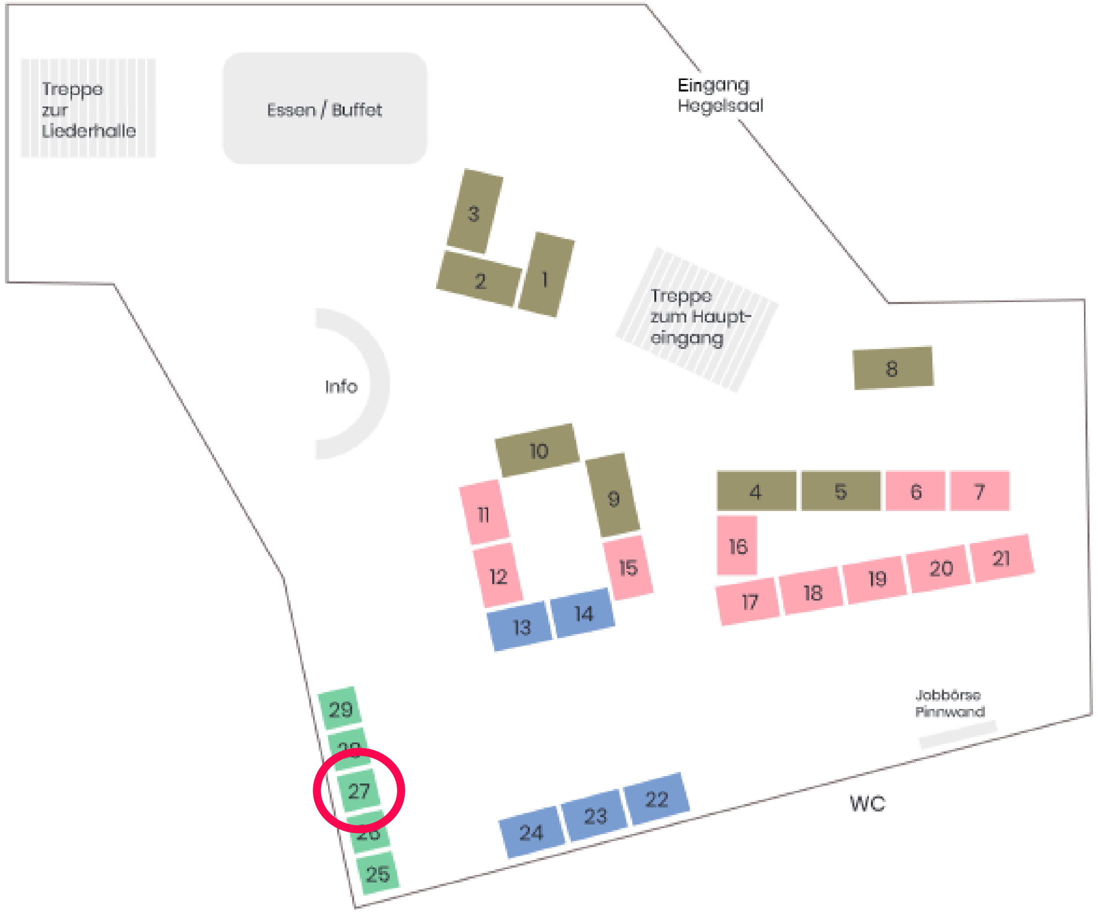
standardized › historizationTest › CSV

Name	Last Modified	Content Type	Size
historization_start.csv	7/30/2024, 3:50:02 PM		1002 B
historization_update1.csv	7/30/2024, 4:07:04 PM		999 B
historization_update2.csv	7/30/2024, 4:07:04 PM		1002 B

Showing 1 to 3 of 3 cached items

Kommt uns an unserem Stand besuchen!

Hegel Foyer – Stand 27


Treppe zur Liederhalle
Essen / Buffet
Eingang Hegelsaal
Treppe zum Haupteingang
Info
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
27
29
Jobbörse Pinnwand
WC