Databricks & Fabric
Better Together

Gerhard Brueckl, paiqo
Marcel Franke, Microsoft

Marcel Franke

Microsoft

- Technical Specialist Manager for Data & AI @Microsoft
- Team of 15 Technical Experts around Data Platform, Databases, AI & Machine Learning
- We cover all industries in Germany

Marcel Franke
Data & AI Team Lead | CTO Office | Startup Advisor | Mentor & Coach

https://github.com/marcelfranke

https://www.linkedin.com/in/marcelfranke/

marcel.franke@microsoft.com

20 SQL KONFERENZ 2024

Gerhard Brueckl

paiqo

www.paiqo.com

- Data Engineer and Solution Architect
- Azure Databricks & Fabric
- Passionate VSCode extension developer
  - Databricks Power Tools
  - Power BI Studio
  - Fabric Studio (preview)

AUSTRIA PASS

MVP Microsoft® Most Valuable Professional

DELTA LAKE

https://github.com/gbrueckl

https://www.linkedin.com/in/gbrueckl

gerhard@gbrueckl.at

20 SQL KONFERENZ 2024

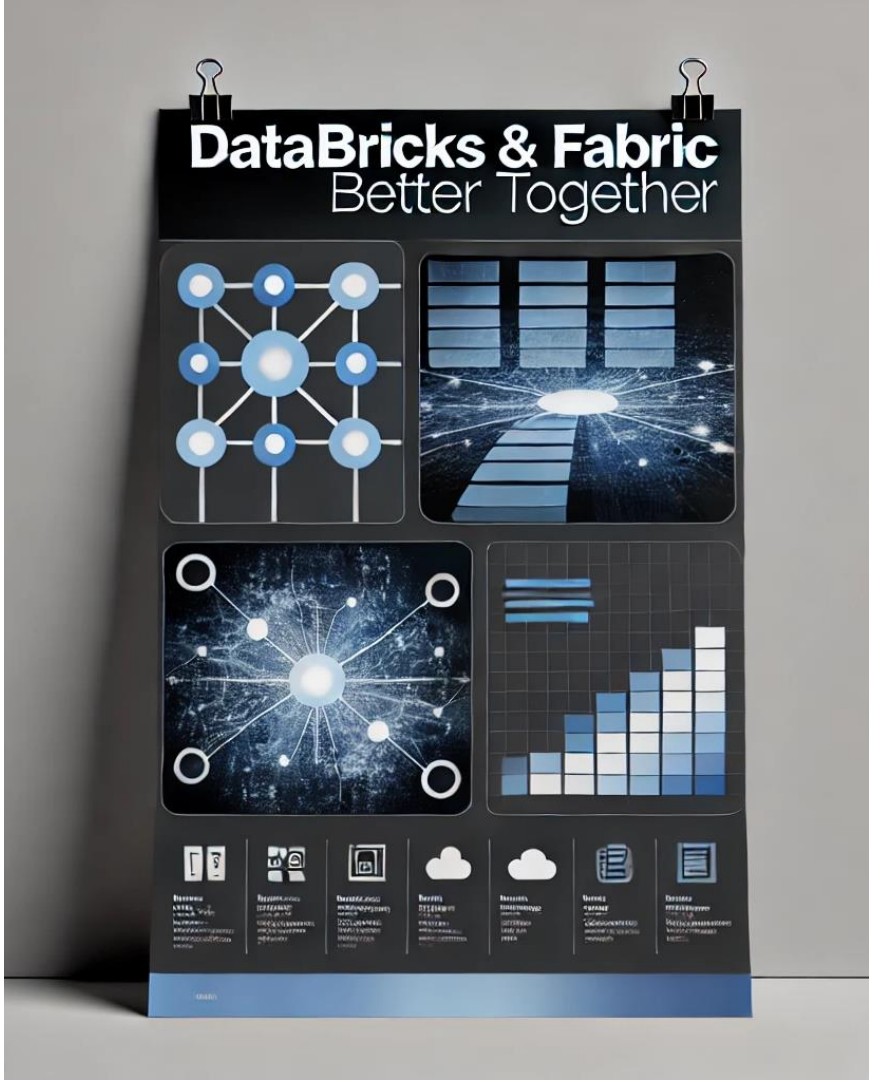
DATA SATURDAYS DATA COMMUNITY AUSTRIA DAY 2025

24-January-2025

- 6 parallel tracks
- 35 national and international speakers
- Sessions in German and English
- 300+ attendees
- **FREE**

Agenda

- Introduction
- Short overview Databricks
- Short overview Fabric
- Typical Scenarios & Considerations
- Outlook & Summary



Session Expectations

- We don’t provide a deeper introduction into Databricks or Fabric
- We compare concepts and architecture patterns
- We look at the similarities and differences
- We show how both worlds can be integrated together

Introduction Databricks

What is Apache Spark?

Open-source Cluster computing framework

- Massive Parallel Processing with linear scale

Built for:

- Speed/Scalability
- Ease-of-Use
- Extensibility

Support for multiple languages

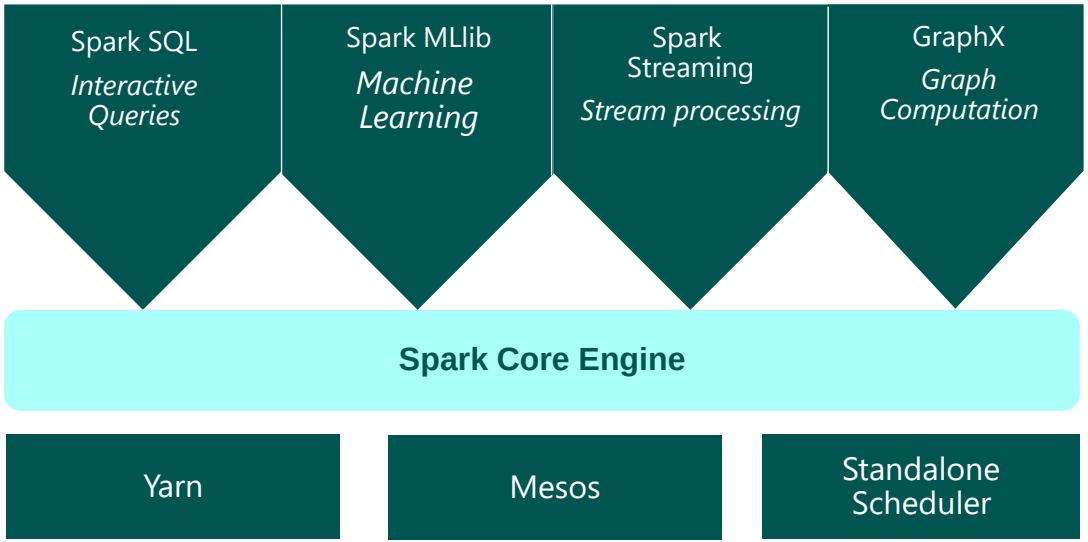
- Java, Scala, Python, R, SQL, (.Net)

APACHE SPARK

An unified, open source, parallel, data processing framework for Big Data Analytics

Spark Unifies:

- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing



What is Databricks?

databricks

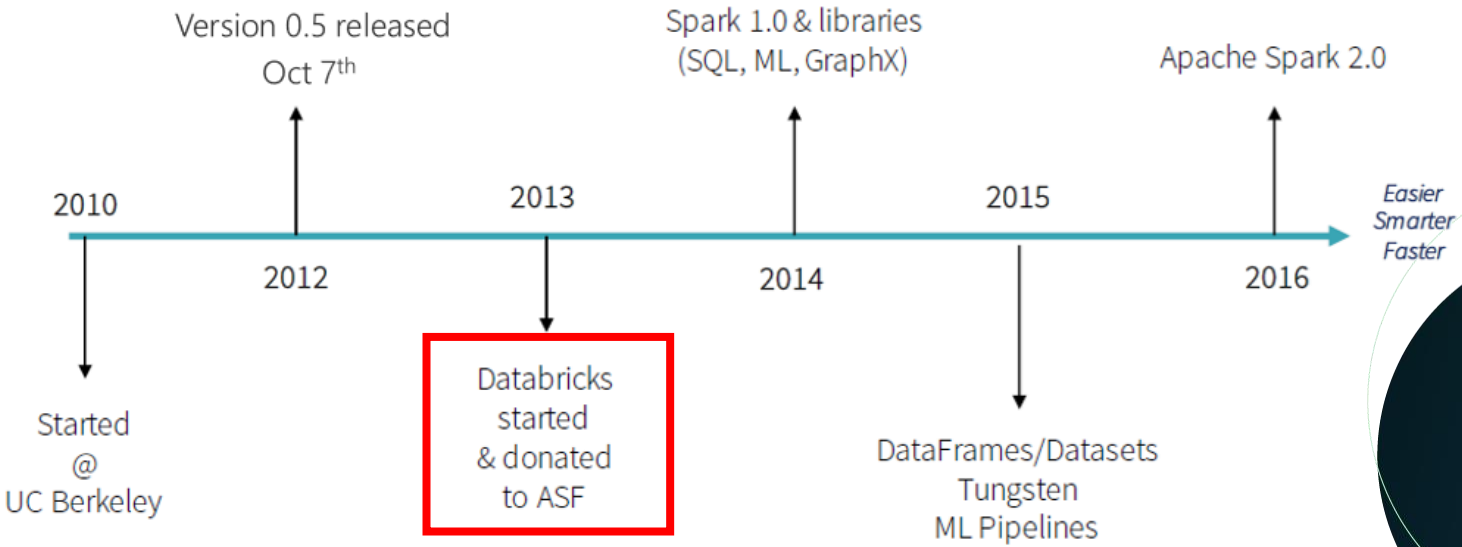
Company that provides a Big Data processing solution in the Cloud based on Apache Spark

- Databricks on AWS
- Azure Databricks
- Databricks on Google Cloud
- **No on-prem solution!**

Creators of Apache® Spark™

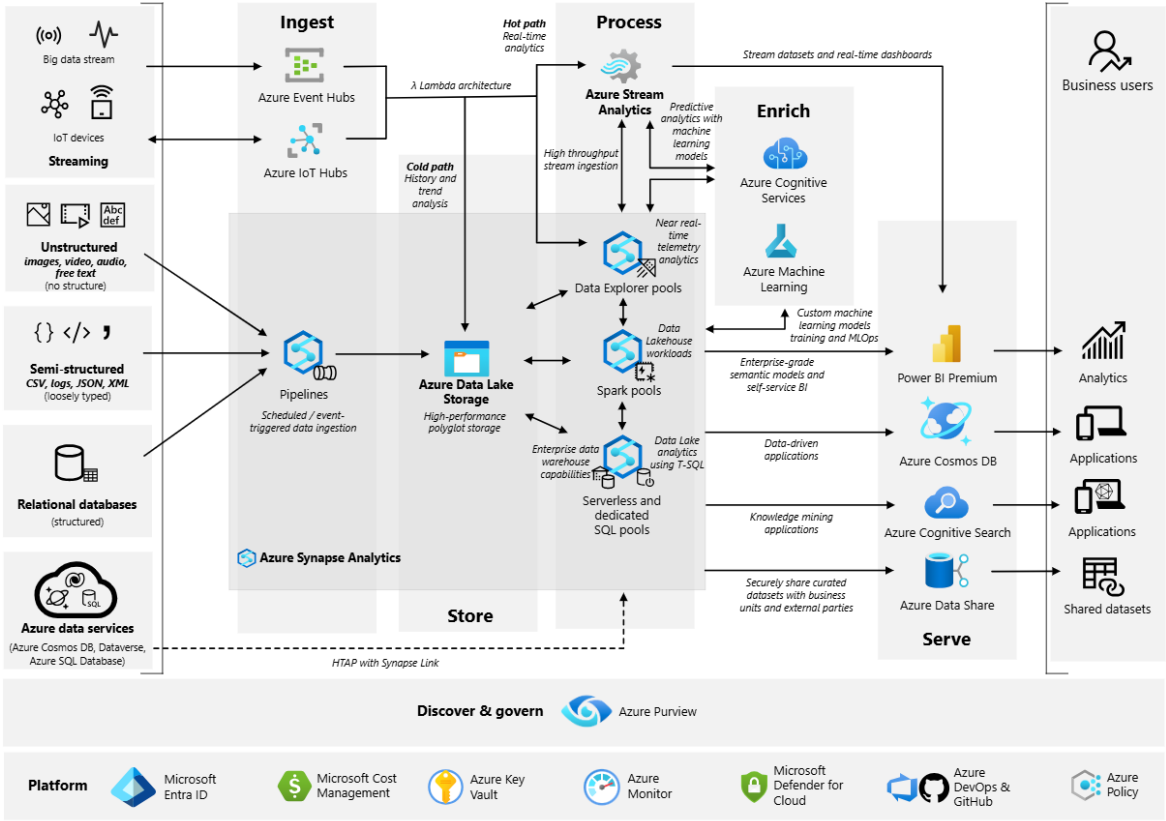
Creators of Delta Lake

Spark & Databricks – A Brief History



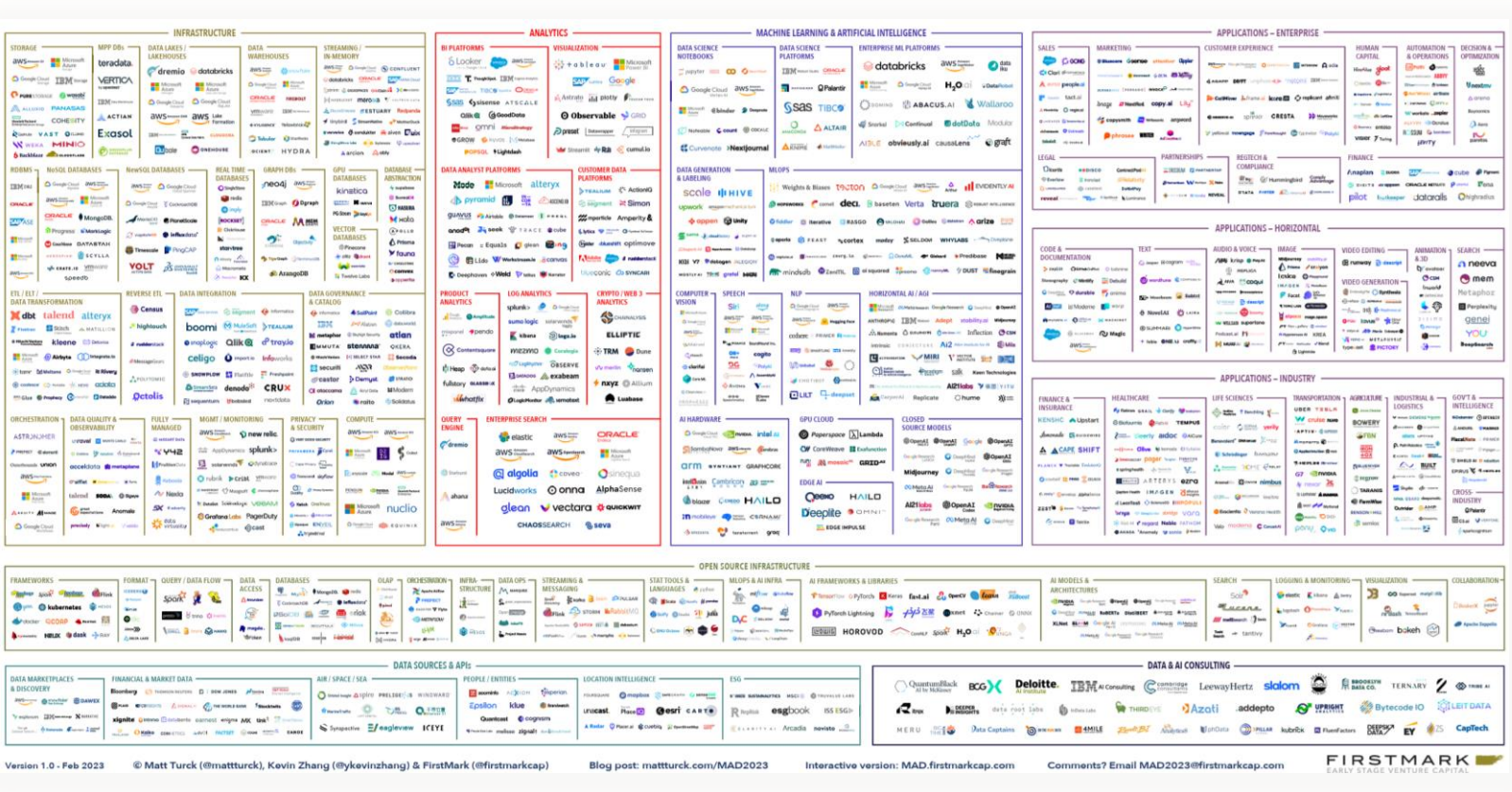
What is Microsoft Fabric?

Reference Architecture Microsoft Data Platform



Microsoft Azure

Customers enhancing their data estate face immense complexity

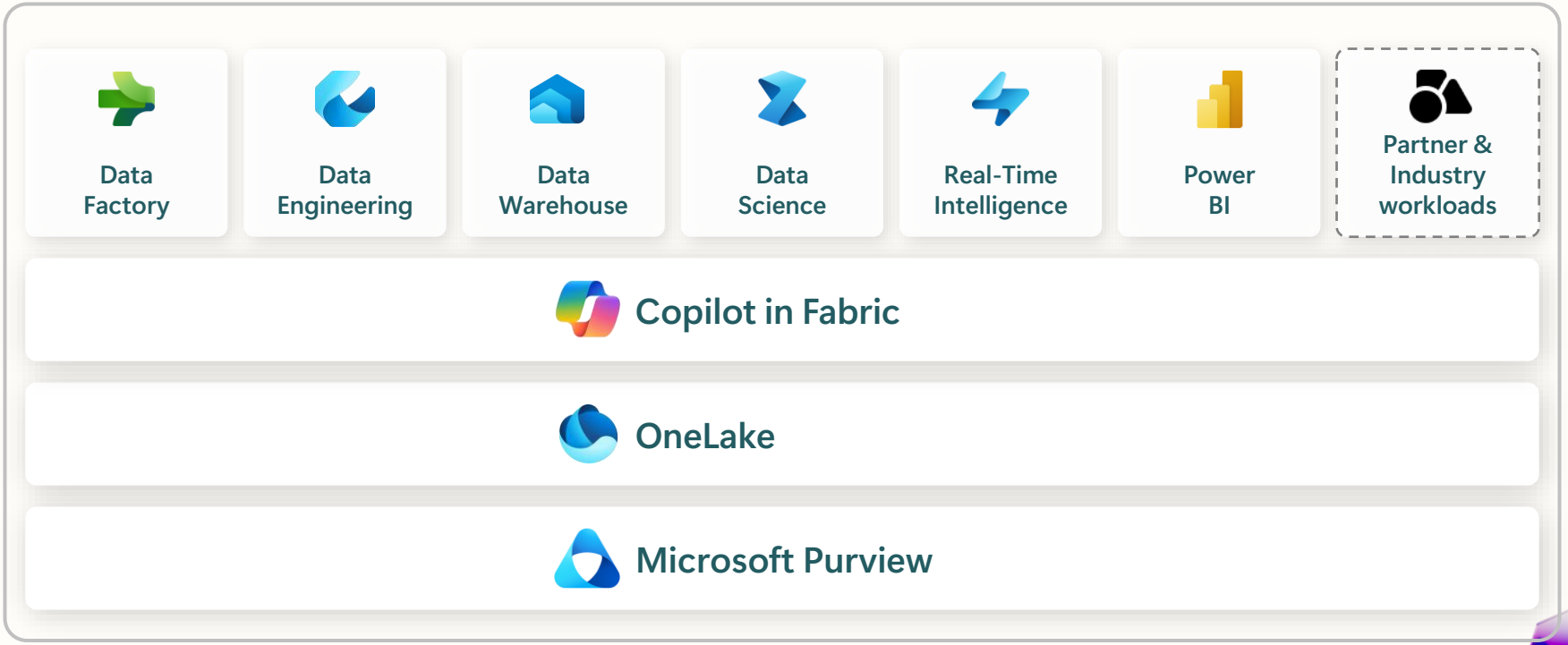


Microsoft Fabric

The data platform for the era of AI

From		To
Multiple analytics services	»	Unified stack
Disconnected data sources	»	All the data in one place
Isolated application	»	Entire estate
Gen AI built on	»	Gen AI built in

Microsoft Fabric



What is needed for a Data Platform?

What is needed for a Data Platform?

- Storage Engine: Data Lake, Open Storage Format
- Processing Engine: Data Preparation, Transformations, Cleaning
- Analytics Engine: BI, Reporting, Querying
- Metastore / Catalog: Data Access Mgmt, Data Discovery, Data Sharing
- Governance & Security: User Access Management, Policies, Data Protection
- Machine Learning & AI: Feature Store, Model Hub & Registry, Logging

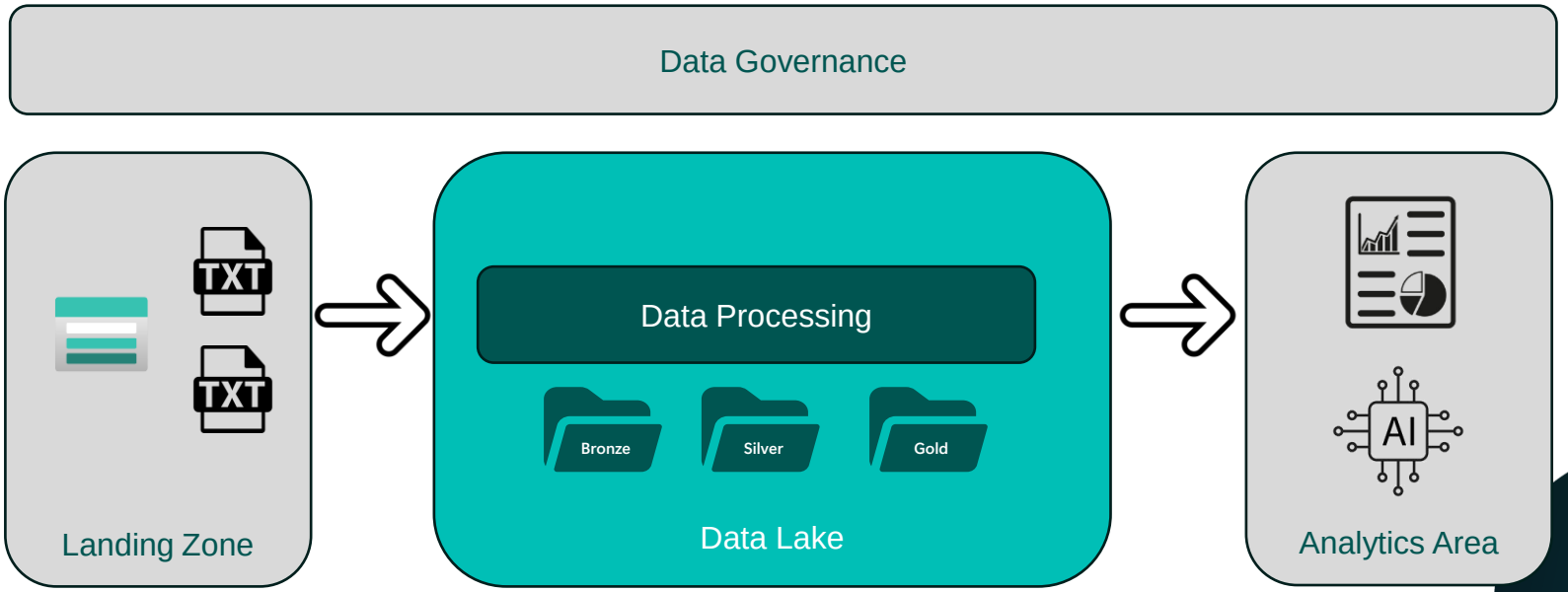
Feature Matrix

	Fabric	Databricks
Storage	OneLake	ADLS Gen2
Storage Format	Delta Lake	Delta Lake
Compute Engines	Spark, T-SQL, Dataflows, …	Spark
Data Catalog	Lakehouse	Unity Catalog
Data Ingestion	Pipelines, Dataflows	LakeFlow Connect (Preview)
Scheduling & Orchestration	Pipelines	Workflows
Reporting	Power BI	Dashboards

Typical Scenarios

Real-life Examples

Reference Demo Scenario

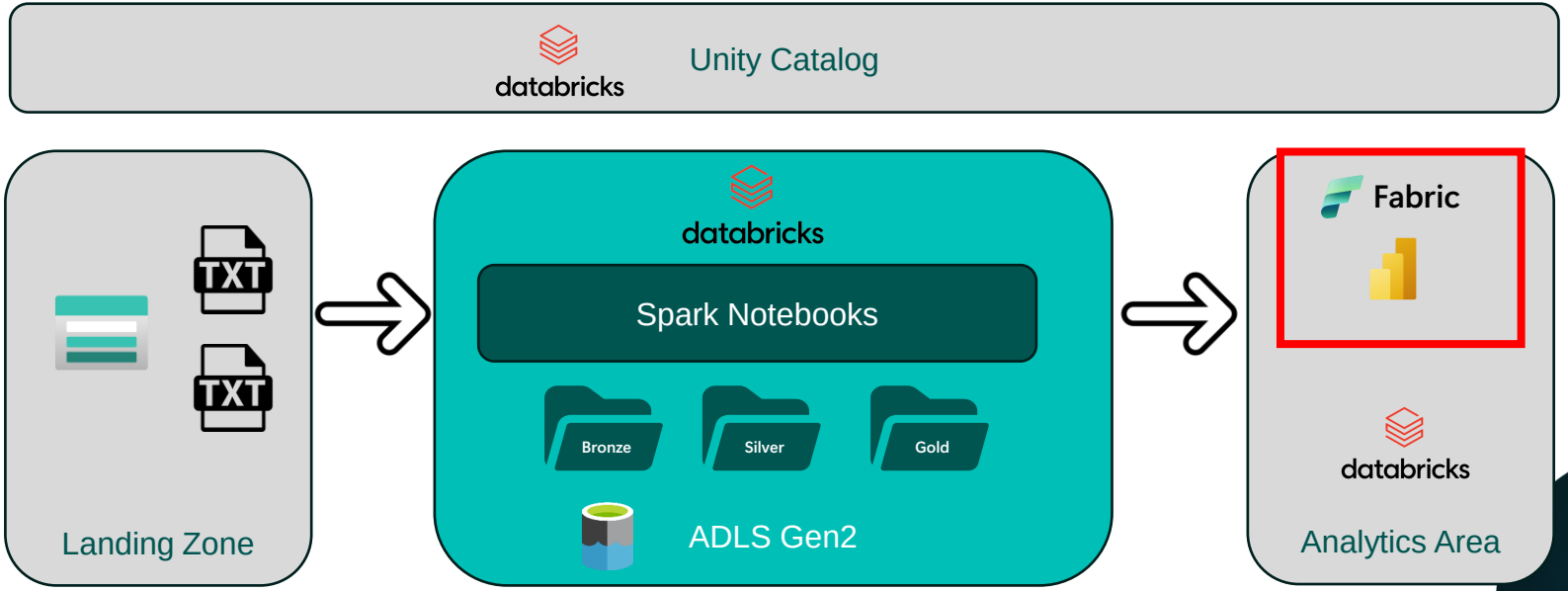


Data Source

Data Ingestion, Cleansing & Preparation

Querying, Reporting, AI & ML

Scenario 1: Databricks + Reporting in Fabric (Power BI)



Data Source

Data Ingestion, Cleansing & Preparation

Querying, Reporting, AI & ML

SQL KONFERENZ 2024

Considerations

Which Power BI mode do you want to use?

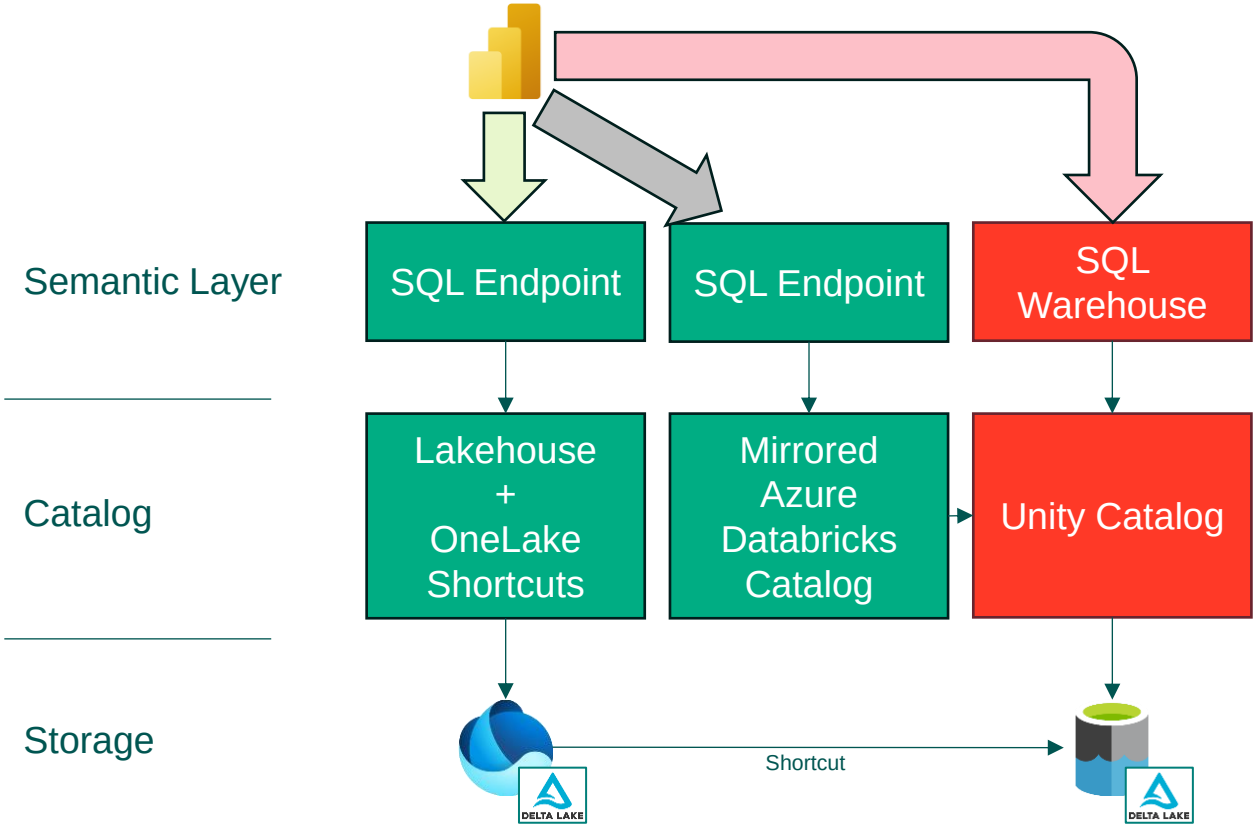
- Import vs. Direct Query vs. Direct Lake

Where to implement the Semantic Layer?

Where to maintain Security? Data Governance?

Integrations - Reporting

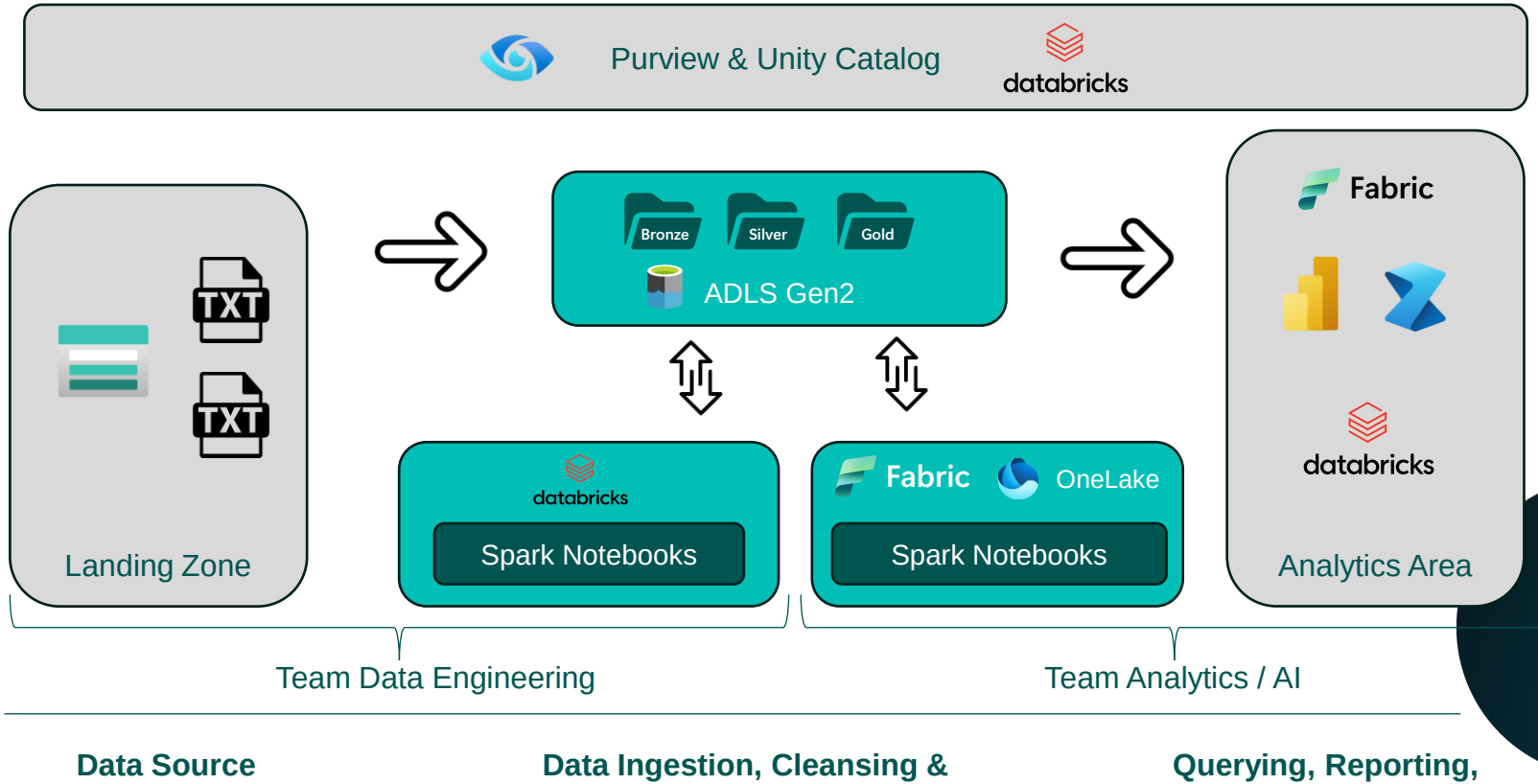
Import	Direct Query	Direct Lake
☑	☑	☐
☑	☑	☑
☑	☑	☑



Take-Aways

	Databricks SQL Warehouse	Lakehouse + OneLake Shortcuts	Mirrored Azure Databricks Catalog
Pros	Simple setup, no additional components	Direct Lake support (Low latency, no processing)	Direct Lake support (Low latency, no processing)
	Robust and proven approach	Semantic Layer close to Power BI	Semantic Layer close to Power BI
	Single security layer (UC)		Automated sync of tables
Cons	No Direct Lake support	Manual sync of tables	
	[Additional costs]	Might not support all Delta Lake features used in Databricks (!)	Might not support all Delta Lake features used in Databricks (!)
		Two security layers (UC + Fabric)	Two security layers (UC + Fabric)

Scenario 2: Mixed Teams and Responsibility (Central Lake)



Considerations

Why use different tools/technologies at all?

Which tool/technology for which layer?

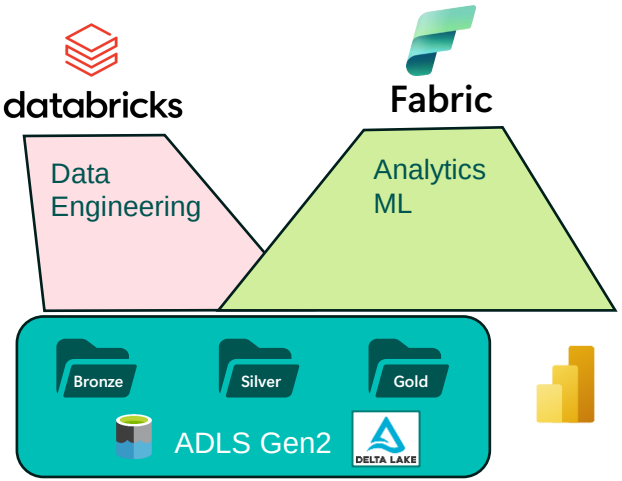
Which storage technology to use?

- ADLS Gen2 or OneLake?

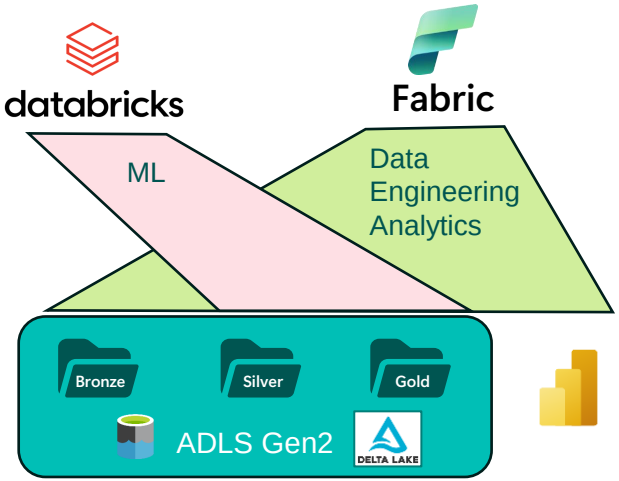
Data Governance? Security? Linage?

Integrations – Mixed Teams

Databricks for lower layers
Fabric for analytical layers



Fabric end-to-end
Databricks for specific use-case

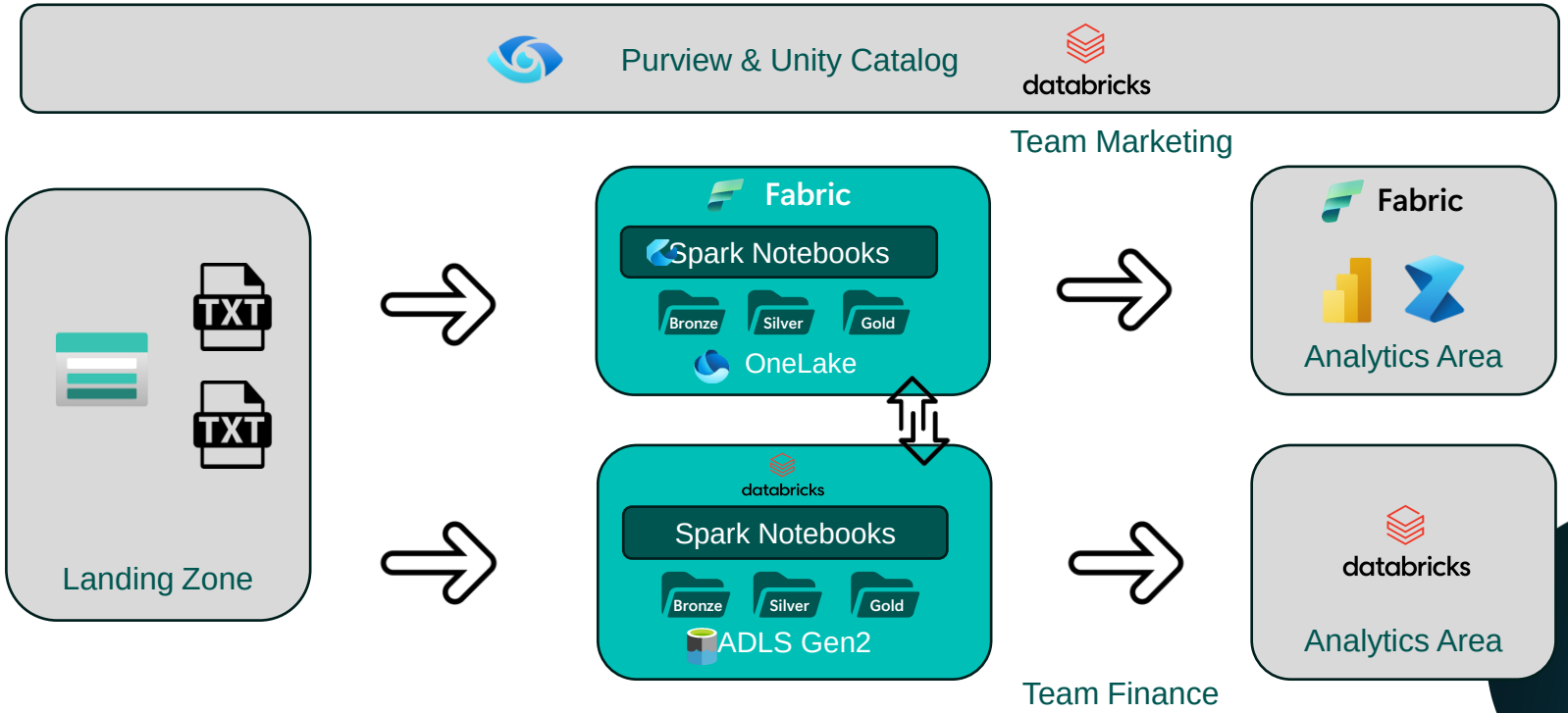


SQL KONFERENZ 2024

Take-Aways

	Databricks	Fabric
Why?	Easier Raw Data Processing with Auto Loader + Delta Live Tables	Better Source Data Integration due various connectivity options
	Latest Spark + Delta Lake features	Integration with other Fabric components
	Higher Maturity	Better Reporting Performance for Power BI (V-Order / Direct Lake)
	[Machine Learning / Data Science]	End-to-End implementation in one Platform
	Professional Development	Business-User friendly (No-Code tools)
	Collaboration	
	Different teams/responsibilities/skillsets	

Scenario 3: Mixed Teams and Responsibility (Data Mesh)



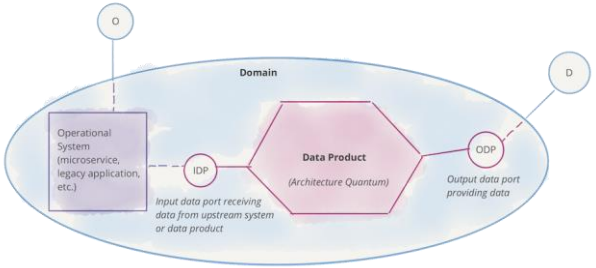
Data Source

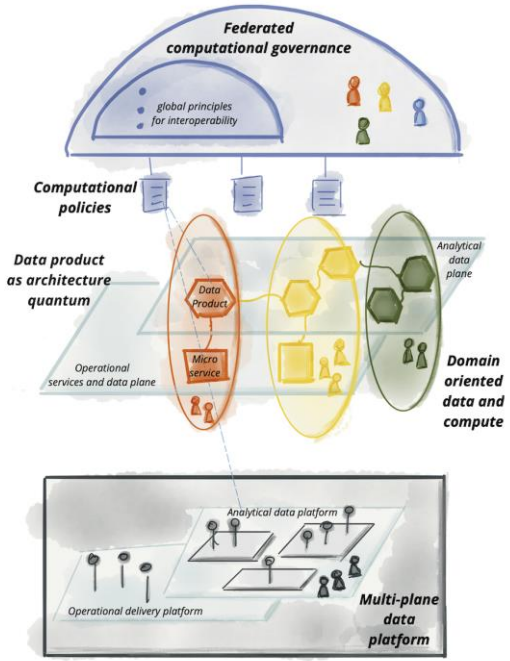
Data Ingestion, Cleansing & Preparation

Querying, Reporting, AI & ML

Principles of Data Mesh

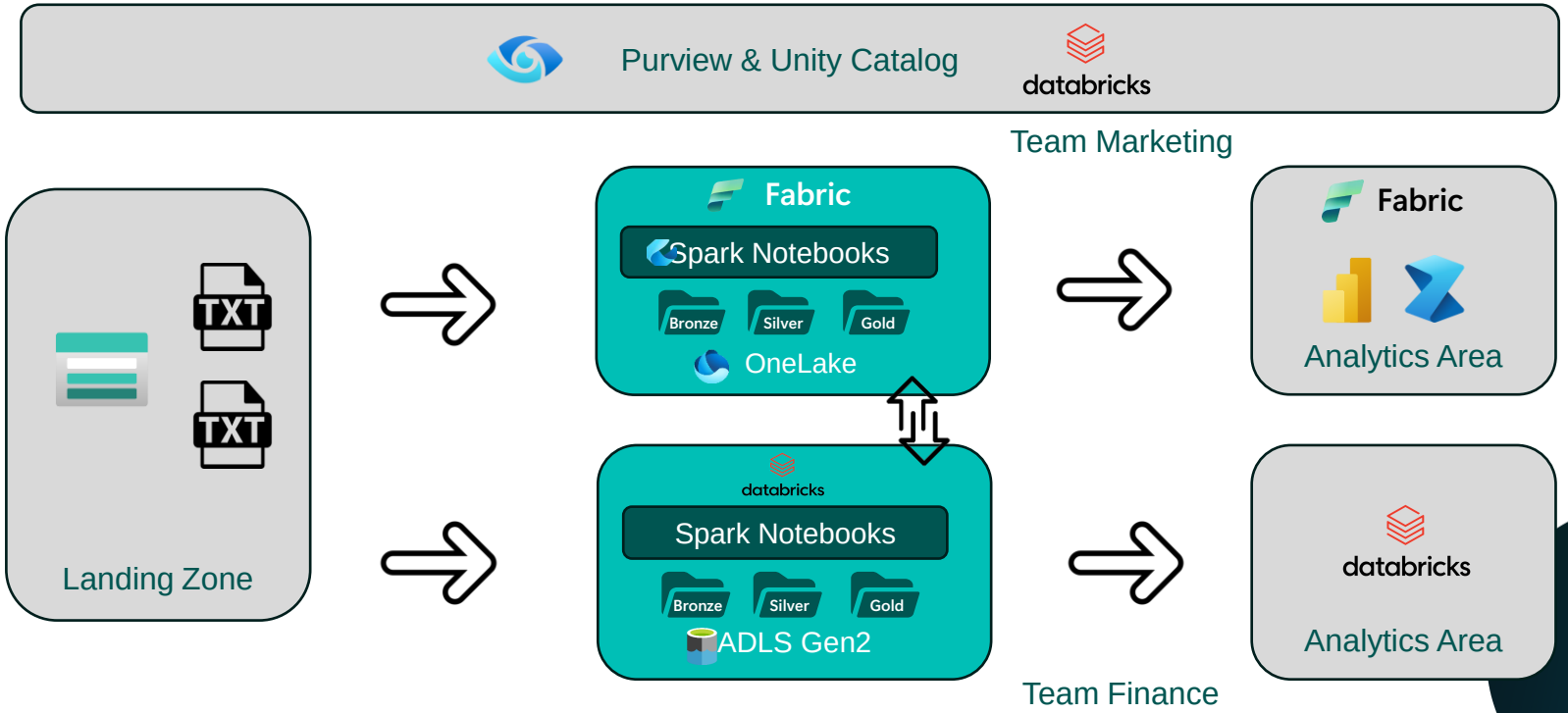
- Domains & Domain Ownership
- **Data as a Product**
- Self-serve data platform
- Federated governance





[Data Mesh Principles and Logical Architecture (martinfowler.com)]

Scenario 3: Mixed Teams and Responsibility (Data Mesh)



Data Source

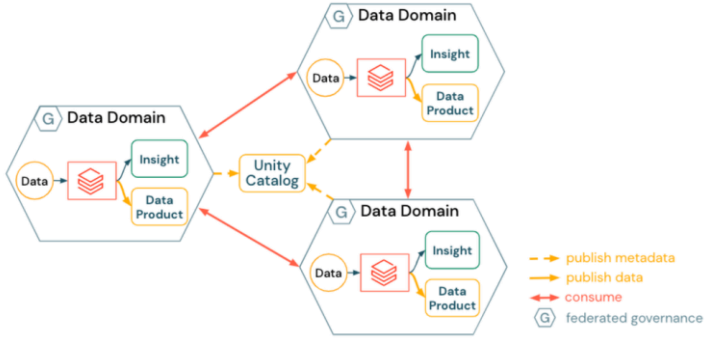
Data Ingestion, Cleansing & Preparation

Querying, Reporting, AI & ML

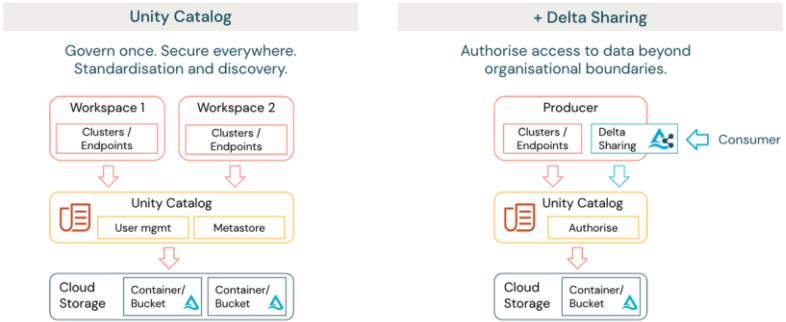
Considerations

- Are Data Mesh concepts supported?
- How do I design and publish Data Products?
- How can I exchange Data Products?
- How do I establish federated Data Governance?

Data Mesh support in Databricks



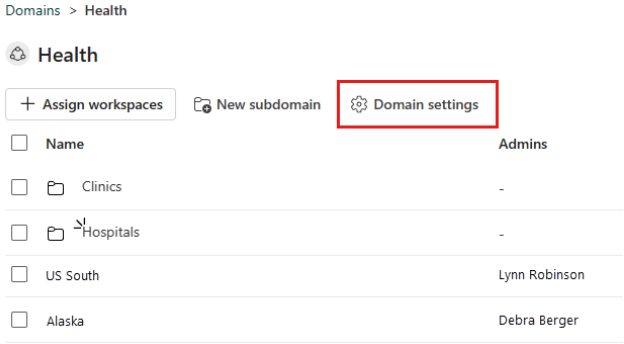
- Domains are represented logically
- Data Products can be created and consumed via Unity Catalog or Delta Sharing
- Cross-Region & Cross-Tenant support
- Federated Data Governance via Unity Catalog (one per region per tenant)

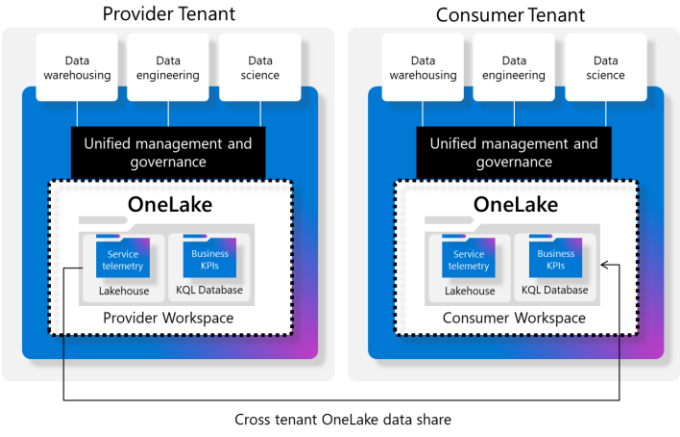


SQL KONFERENZ 2024

Data Mesh support in Fabric

- Domains are a key concept in the Product
- Data Products can be created and shared via Data Sharing
- Further Data Product functionality in Purview
- Cross-Region & Cross-Tenant support
- Data Products can be consumed via Fabric Data Sharing
- Central Data Governance via Purview





2024

Take-Aways

- **Data Domains & Domain Ownership** can be organized in both technologies but no further integration
- **Data Products** can be created in both technologies
- **Data Sharing**
  - Delta Shares can be consumed in Fabric (via Power BI, Dataflows & Spark)
  - Fabric Data Shares can not be consumed in Databricks
- **Data Governance & Data Discovery**
  - Unity Catalog does not integrate with Fabric
  - Purview can be extended with Unity Catalog objects via API

Summary

- Both are very capable data platforms on their own
- Both platforms have their strengths and weaknesses
- You can mix and match to get best of both worlds

What's next?

Databricks and Microsoft are working on even better compatibility:

- Better Integration with Unity Catalog in Fabric (Public Preview, read-only)
- Unity Catalog support for External Locations to OneLake (Private Preview Databricks)

Session Feedback