Microsoft Azure

From Metadata to Lakehouse

Andreas Bergstedt
Data and AI Global Black Belt

Introduction / Agenda

The Evolution of ETL

Why Lakehouse

Metadata Driven Engineering

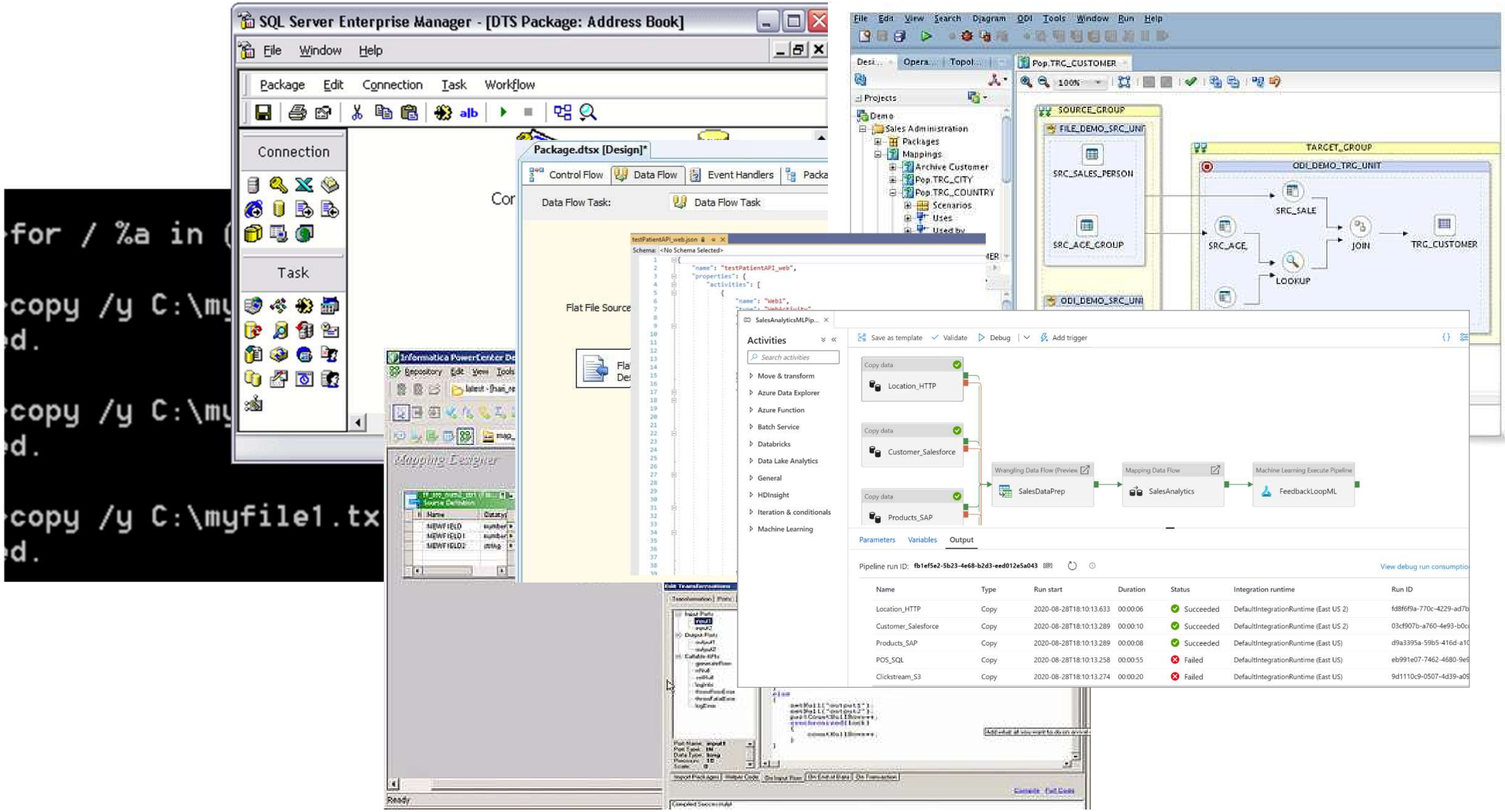
Driving at Scale

Key take-aways

Introduction / Agenda

The Evolution of ETL

Microsoft Azure



```
for / %a in (

copy /y C:\my
d.

copy /y C:\my
d.

copy /y C:\myfile1.tx
d.
```

testPatientAPI_web.json

Schema: <No Schema Selected>

```
{
    "name": "testPatientAPI_web",
    "properties": {
        "activities": [
            {
                "name": "Web1",
                "type": "WebActivity",
```

Pipeline run ID: fb1ef5e2-5b23-4e68-b2d3-eed012e5a043

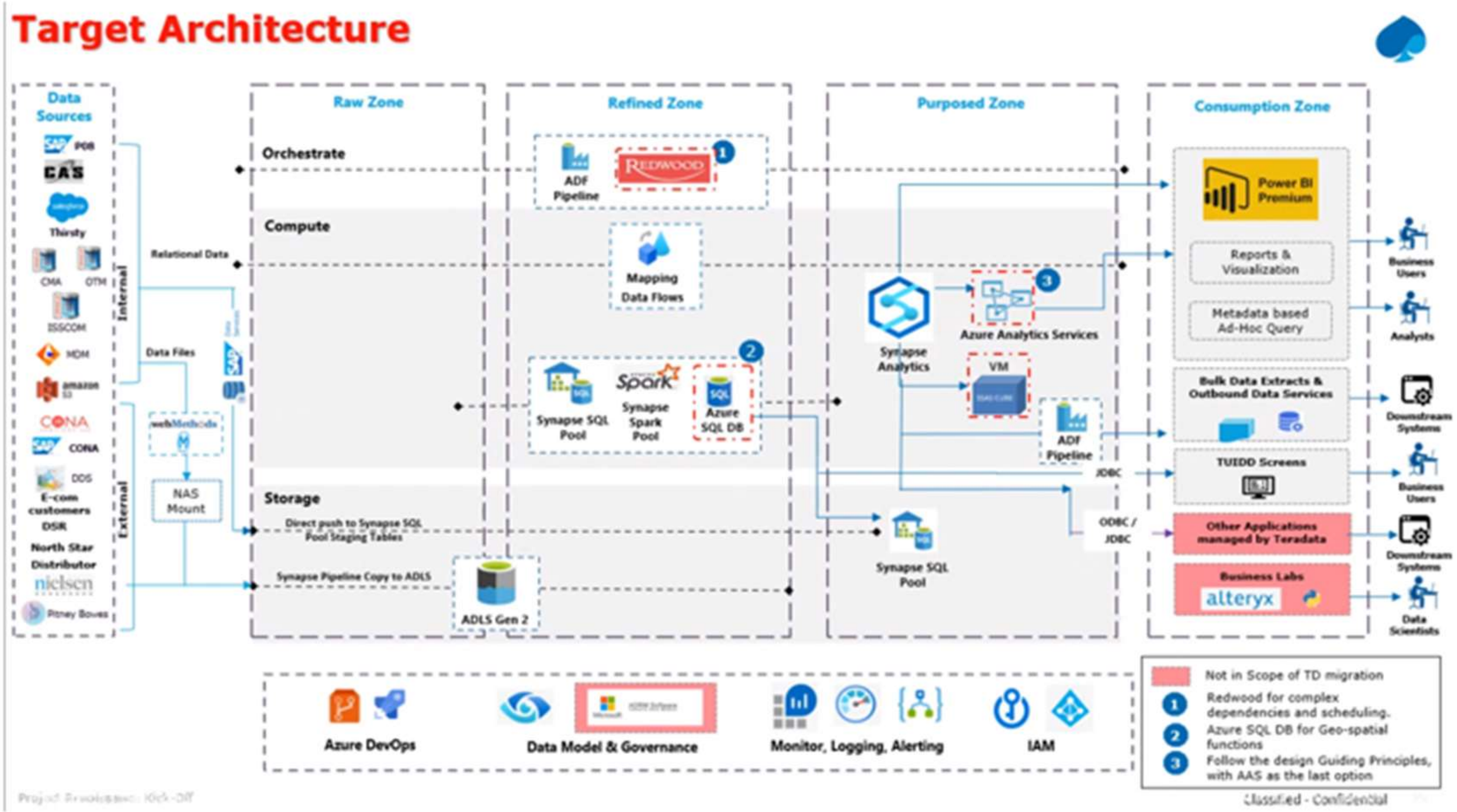
Name	Type	Run start	Duration	Status	Integration runtime	Run ID
Location_HTTP	Copy	2020-08-28T18:10:13.633	00:00:06	✓ Succeeded	DefaultIntegrationRuntime (East US 2)	fd8f6f9a-770c-4229-ad7b
Customer_Salesforce	Copy	2020-08-28T18:10:13.289	00:00:10	✓ Succeeded	DefaultIntegrationRuntime (East US 2)	03cf907b-a760-4e93-b0c
Products_SAP	Copy	2020-08-28T18:10:13.289	00:00:08	✓ Succeeded	DefaultIntegrationRuntime (East US)	d9a3395a-59b5-416d-a1
POS_SQL	Copy	2020-08-28T18:10:13.258	00:00:55	☒ Failed	DefaultIntegrationRuntime (East US)	eb991e07-7462-4680-9e
Clickstream_S3	Copy	2020-08-28T18:10:13.274	00:00:20	☒ Failed	DefaultIntegrationRuntime (East US)	9d1110c9-0507-4d39-a0

APACHE
Spark

DELTA LAKE

databricks

Why Lakehouse



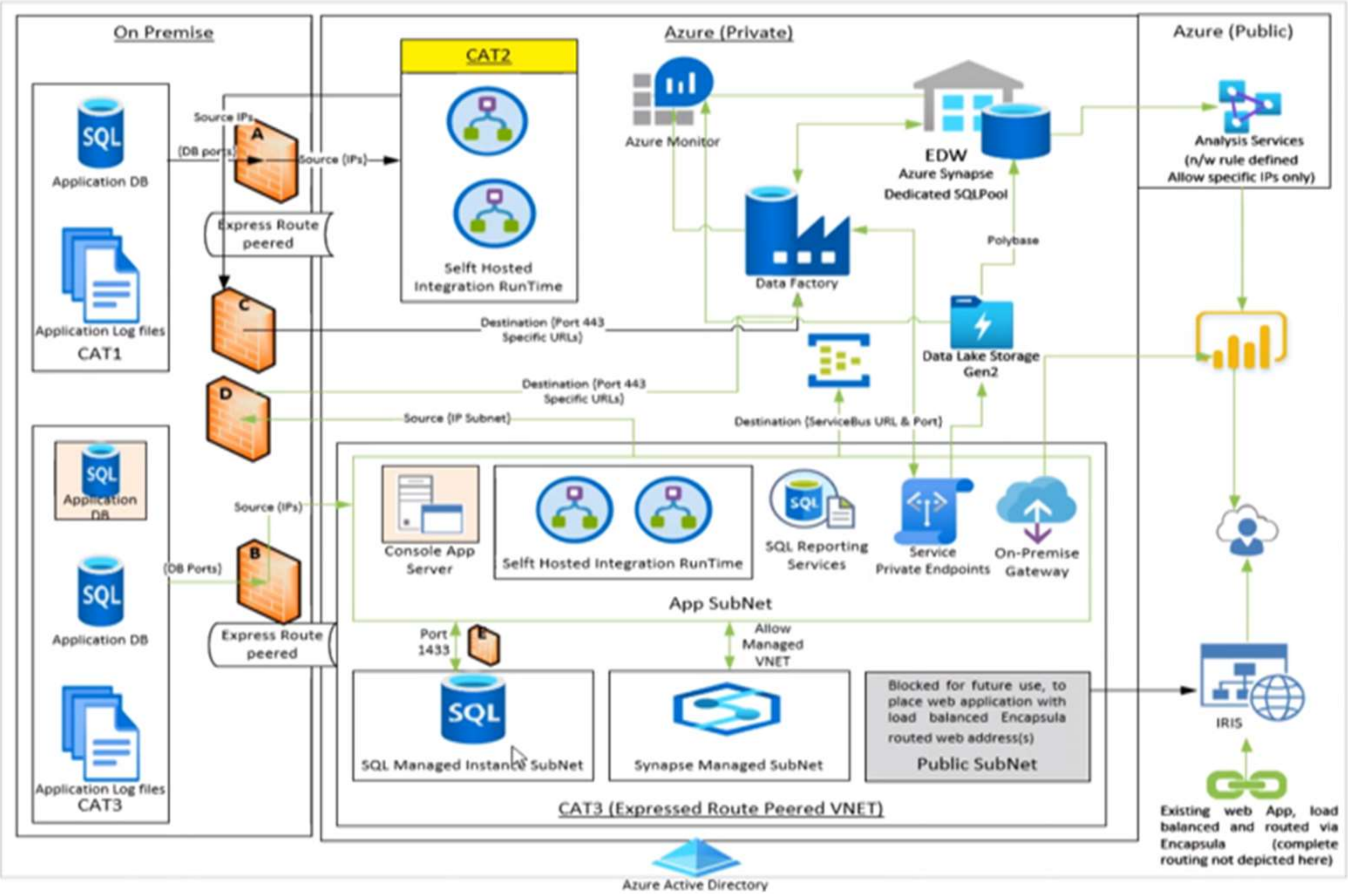
Target Architecture

🟥 Not in Scope of TD migration

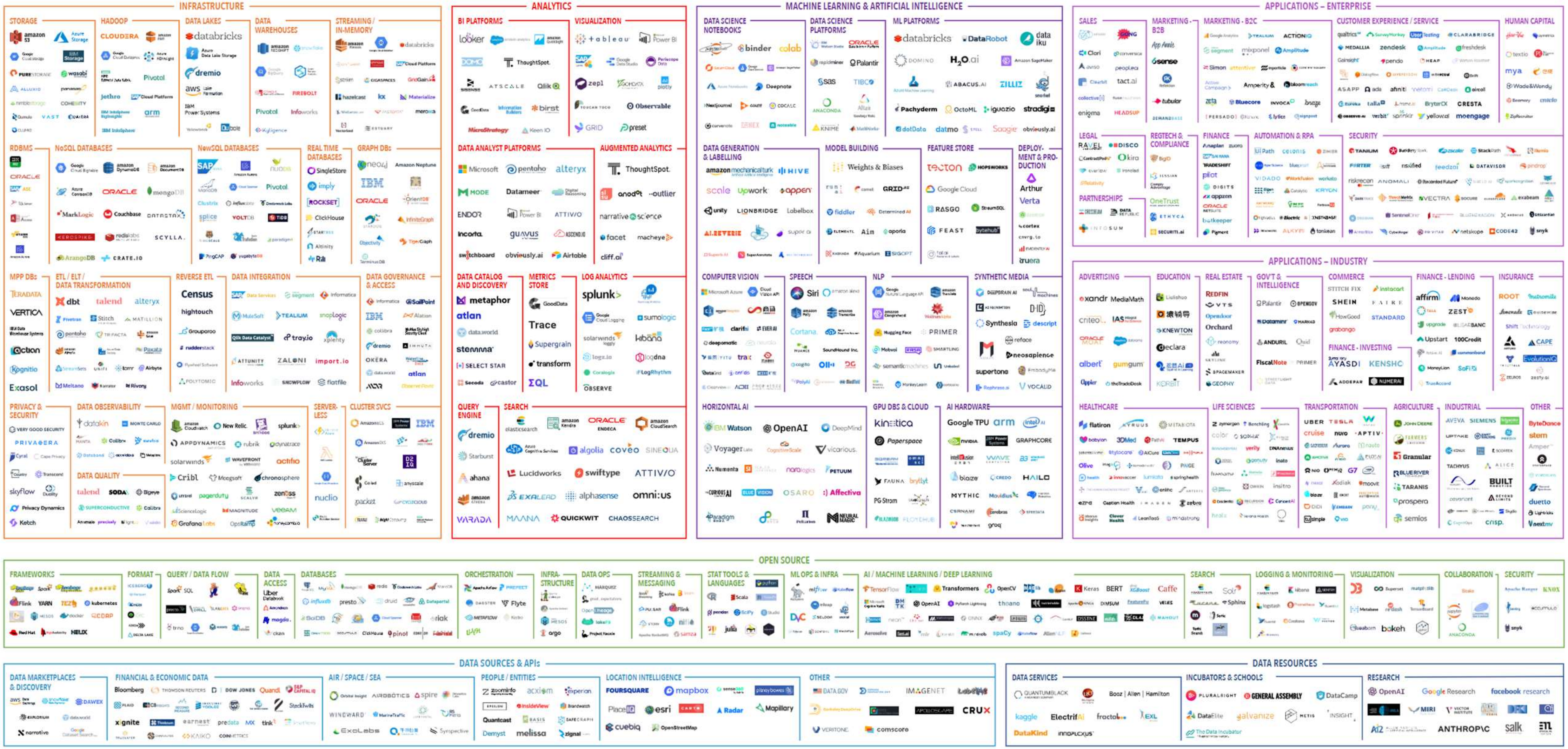
1. Redwood for complex dependencies and scheduling.
2. Azure SQL DB for Geo-spatial functions
3. Follow the design Guiding Principles, with AAS as the last option

Project Renaissance: Kick-Off

Classified - Confidential


On Premise
Application DB
Application Log files
CAT1
Application DB
Application DB
Application Log files
CAT3
Source IPs
(DB ports)
A
Source (IPs)
Express Route peered
C
D
B
Source (IPs)
(DB Ports)
Express Route peered
Azure (Private)
CAT2
Selft Hosted Integration RunTime
Azure Monitor
EDW
Azure Synapse Dedicated SQLPool
Data Factory
Polybase
Data Lake Storage Gen2
Destination (Port 443 Specific URLs)
Destination (Port 443 Specific URLs)
Source (IP Subnet)
Destination (ServiceBus URL & Port)
Console App Server
Selft Hosted Integration RunTime
SQL Reporting Services
Service Private Endpoints
On-Premise Gateway
App SubNet
Port 1433
E
Allow Managed VNET
SQL
SQL Managed Instance SubNet
Synapse Managed SubNet
Blocked for future use, to place web application with load balanced Encapsula routed web address(s)
Public SubNet
CAT3 (Expressed Route Peered VNET)
Azure (Public)
Analysis Services (n/w rule defined Allow specific IPs only)
IRIS
Existing web App, load balanced and routed via Encapsula (complete routing not depicted here)
Azure Active Directory

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



Version 2.0 - October 2021

© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

mattturck.com/data2021

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

https://mattturck.com/data2021/

Microsoft Azure

APACHE Spark

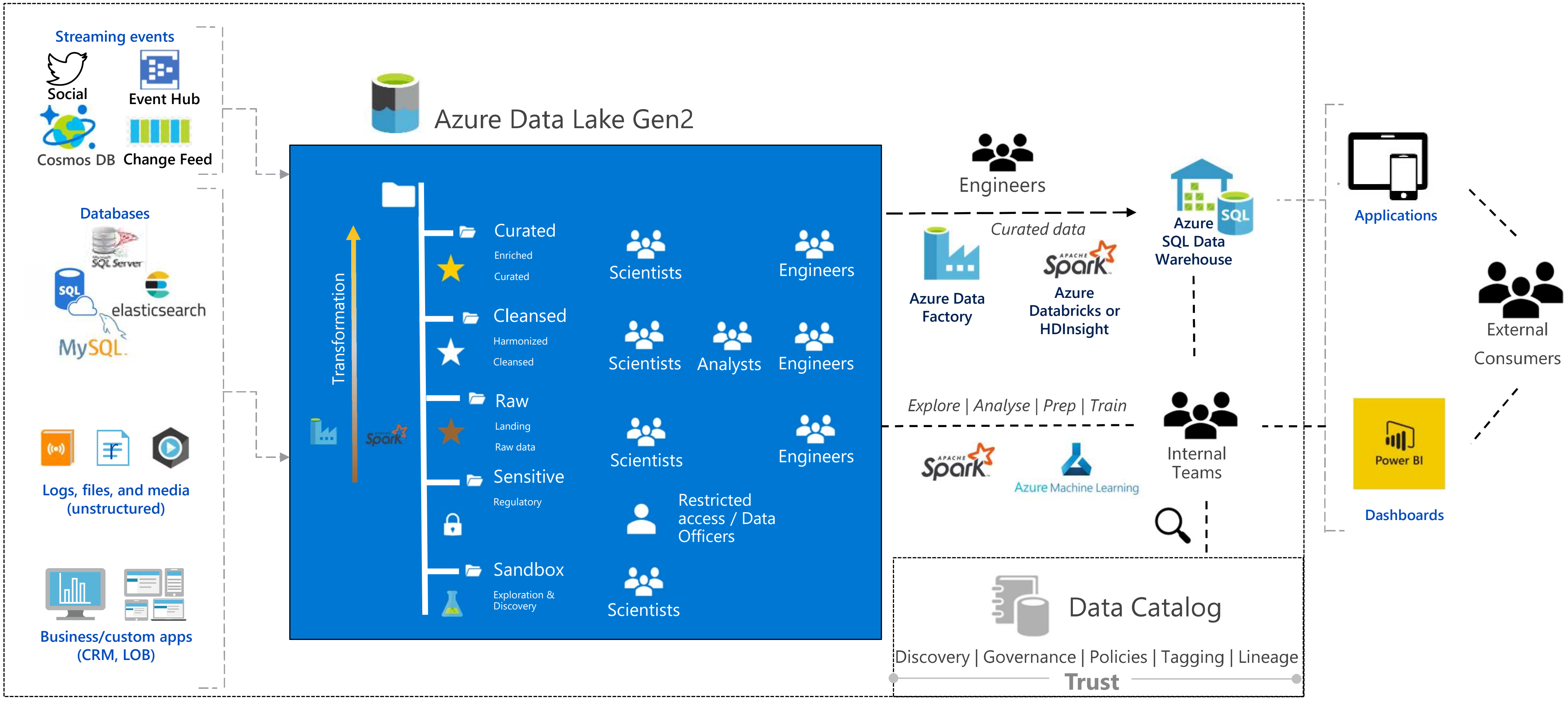
Azure Synapse Analytics

DELTA LAKE

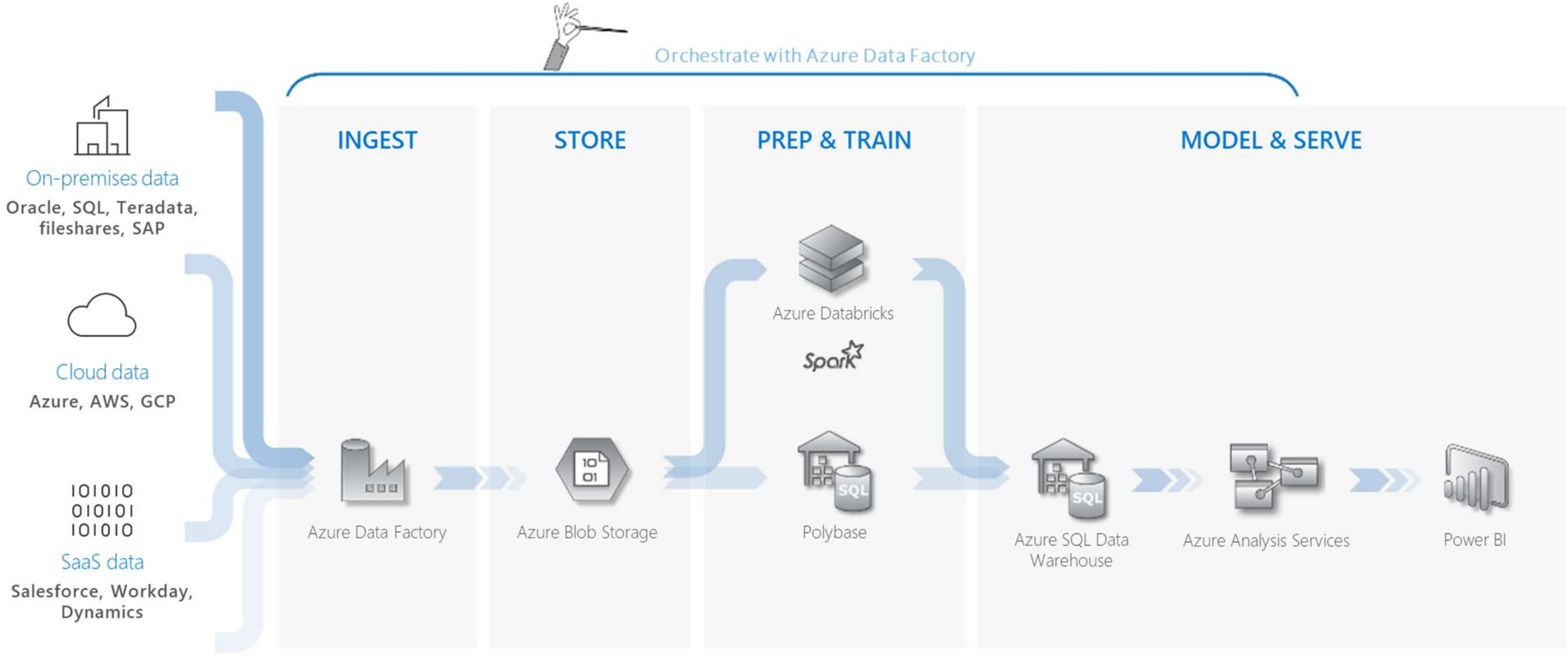
databricks

ICEBERG

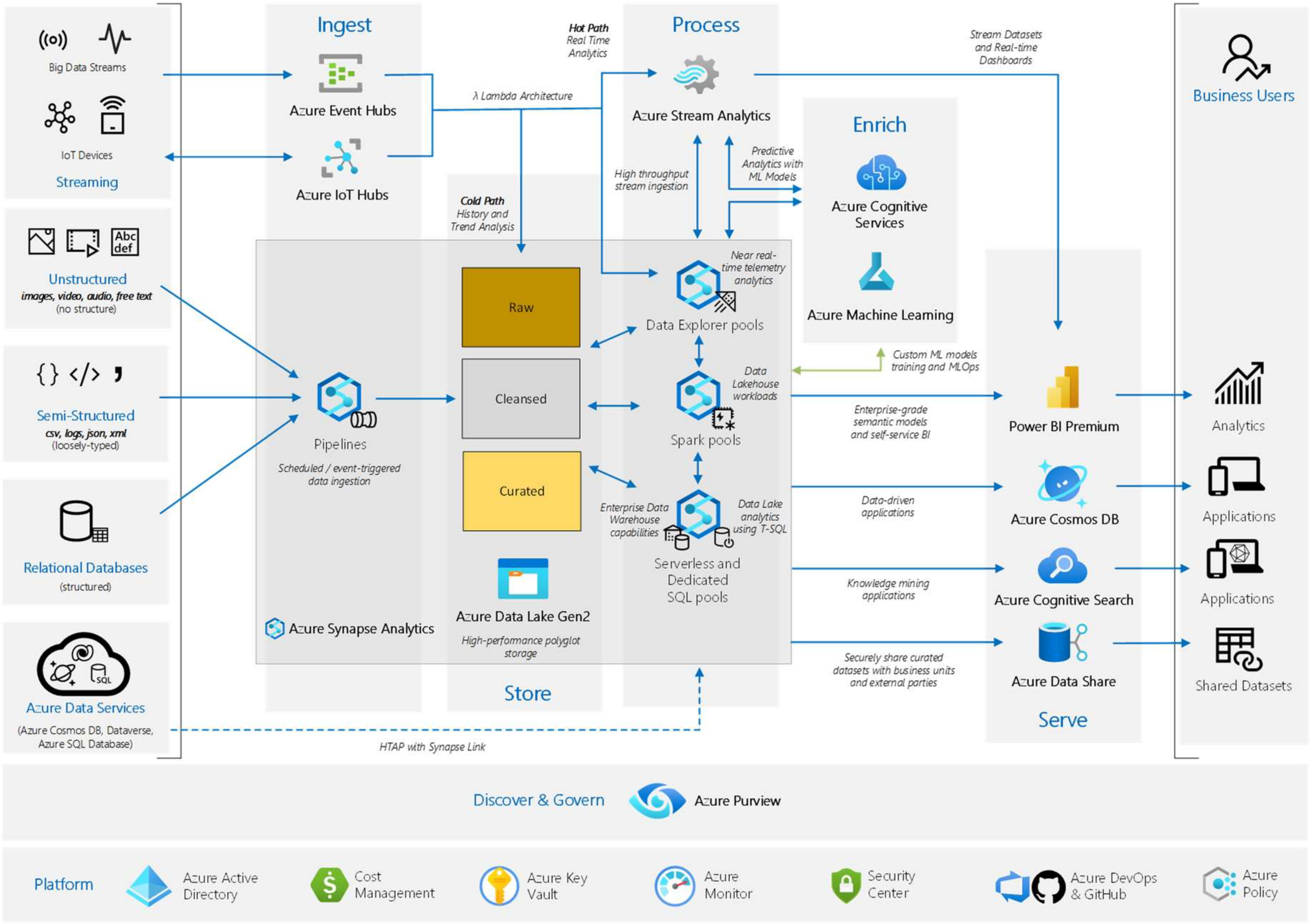
Data Lake Architecture – Concepts, Tools & Process



Modernize your enterprise data warehouse at scale



Microsoft Azure also supports other **Big Data** services like **Azure HDInsight**, **Azure SQL Database** and **Azure Data Lake** to allow customers to tailor the above architecture to meet their unique needs.


Big Data Streams
IoT Devices
Streaming
Unstructured
images, video, audio, free text
(no structure)
Semi-Structured
csv, logs, json, xml
(loosely-typed)
Relational Databases
(structured)
Azure Data Services
(Azure Cosmos DB, Dataverse, Azure SQL Database)
Ingest
Azure Event Hubs
Azure IoT Hubs
Pipelines
Scheduled / event-triggered data ingestion
Azure Synapse Analytics
Hot Path
Real Time Analytics
λ Lambda Architecture
Cold Path
History and Trend Analysis
Raw
Cleansed
Curated
Azure Data Lake Gen2
High-performance polyglot storage
Store
HTAP with Synapse Link
Process
Azure Stream Analytics
High throughput stream ingestion
Predictive Analytics with ML Models
Near real-time telemetry analytics
Data Explorer pools
Data Lakehouse workloads
Spark pools
Enterprise Data Warehouse capabilities
Data Lake analytics using T-SQL
Serverless and Dedicated SQL pools
Enrich
Azure Cognitive Services
Azure Machine Learning
Custom ML models training and MLOps
Stream Datasets and Real-time Dashboards
Enterprise-grade semantic models and self-service BI
Data-driven applications
Knowledge mining applications
Securely share curated datasets with business units and external parties
Power BI Premium
Azure Cosmos DB
Azure Cognitive Search
Azure Data Share
Serve
Business Users
Analytics
Applications
Applications
Shared Datasets
Discover & Govern
Azure Purview
Platform
Azure Active Directory
Cost Management
Azure Key Vault
Azure Monitor
Security Center
Azure DevOps & GitHub
Azure Policy

Medallion architecture

Although the 3-layered design is common and well-known, there are many discussions on the scope, purpose, and best practices on each of these layers.

Bronze layer

Typically raw, "as-is"

- Maintains the raw state in the structure “as-is”
- Data is immutable (read-only)
- Delivery-based partitioned tables, i.e., YYYYMMDD
- Mostly Parquet. Sometimes other formats
- Can be any combination of streaming and batch transactions
- May include extra metadata (schema)
- May be fed from a “mediation layer”
- Used for debugging, testing

Silver layer

Matched and conformed

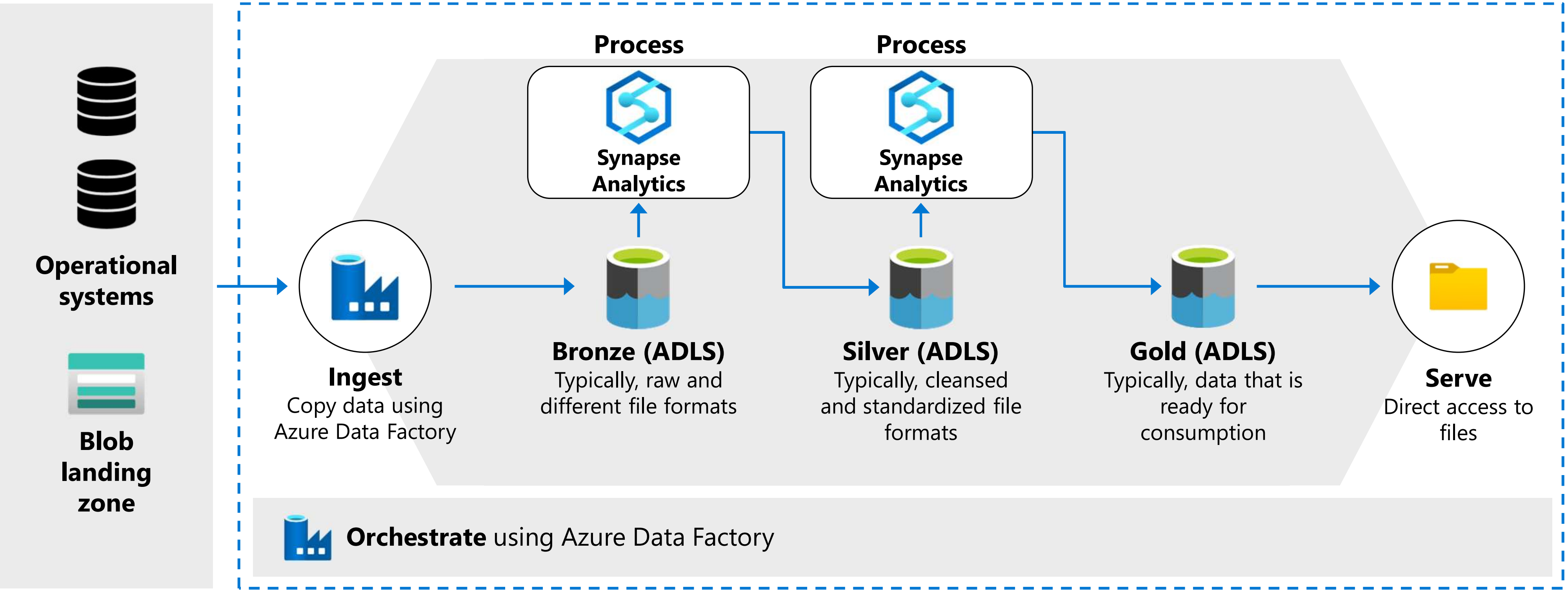
- Uses data quality rules for validation
- Usually only functional data
- Historization is merged (SCD2)
- Efficient storage format; Delta
- Versioning for rolling back
- Handles missing or incorrect data
- Usually enriched with reference data
- Source-oriented, although queryable and cluttered around subject areas
- Usually used by operational analytical teams

Gold layer

Refined business-level

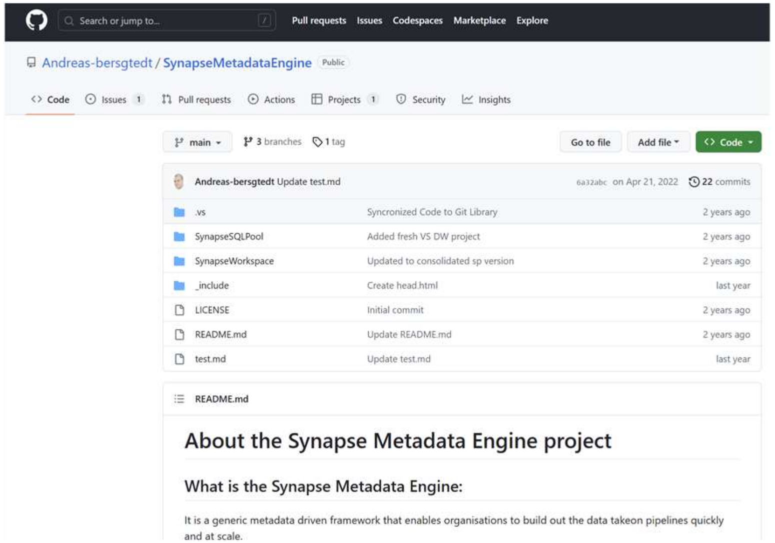
- What enterprises call data products: consumer-ready / user-friendly data
- Data is highly governed and well-documented
- Historization is applied only for the set of use cases or consumers
- Contains complex business rules, such as calculations and enrichments
- Efficient storage format; Delta
- Versioning for rolling back
- Might contain additional sub layers for sharing or distributing data

Basic Lakehouse Architecture


Operational systems
Blob landing zone
Ingest
Copy data using Azure Data Factory
Process
Synapse Analytics
Process
Synapse Analytics
Bronze (ADLS)
Typically, raw and different file formats
Silver (ADLS)
Typically, cleansed and standardized file formats
Gold (ADLS)
Typically, data that is ready for consumption
Serve
Direct access to files
Orchestrate using Azure Data Factory

Metadata Driven Engineering

Available Frameworks



Andreas-bersgtedt / SynapseMetadataEngine Public

Andreas-bersgtedt Update test.md 6a32abc on Apr 21, 2022 22 commits

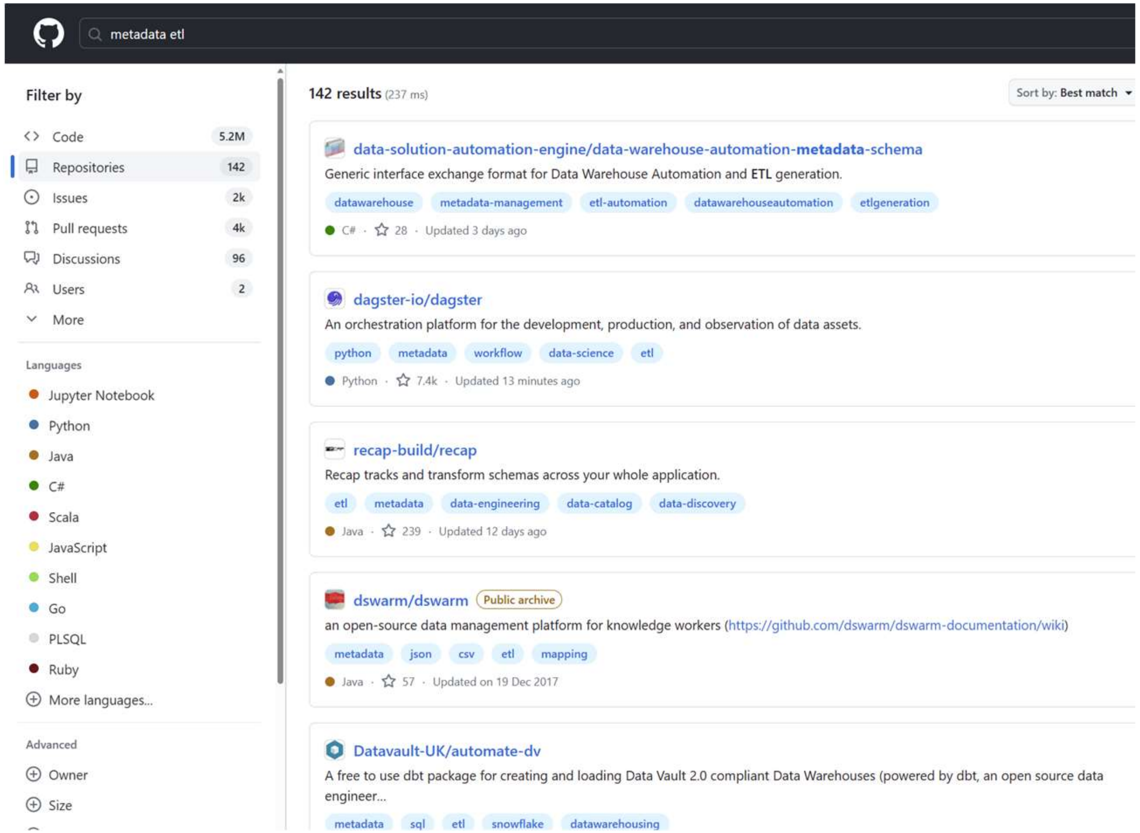
.vs	Syncronized Code to Git Library	2 years ago
SynapseSQLPool	Added fresh VS DW project	2 years ago
SynapseWorkspace	Updated to consolidated sp version	2 years ago
_include	Create head.html	last year
LICENSE	Initial commit	2 years ago
README.md	Update README.md	2 years ago
test.md	Update test.md	last year

README.md

About the Synapse Metadata Engine project

What is the Synapse Metadata Engine:

It is a generic metadata driven framework that enables organisations to build out the data takeon pipelines quickly and at scale.



142 results (237 ms)

data-solution-automation-engine/data-warehouse-automation-metadata-schema
Generic interface exchange format for Data Warehouse Automation and ETL generation.
datawarehouse · metadata-management · etl-automation · datawarehouseautomation · etlgeneration
C# · 28 · Updated 3 days ago

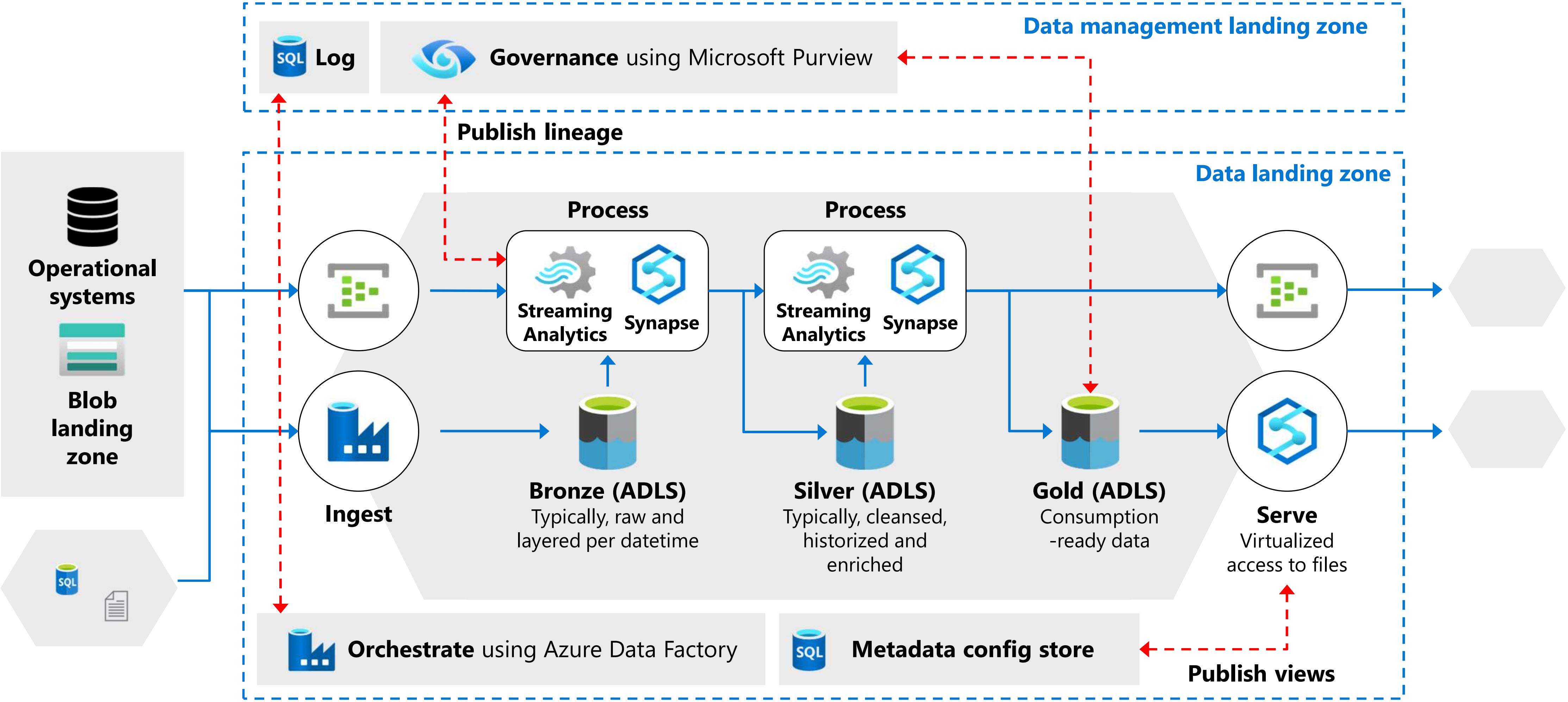
dagster-io/dagster
An orchestration platform for the development, production, and observation of data assets.
python · metadata · workflow · data-science · etl
Python · 7.4k · Updated 13 minutes ago

recap-build/recap
Recap tracks and transform schemas across your whole application.
etl · metadata · data-engineering · data-catalog · data-discovery
Java · 239 · Updated 12 days ago

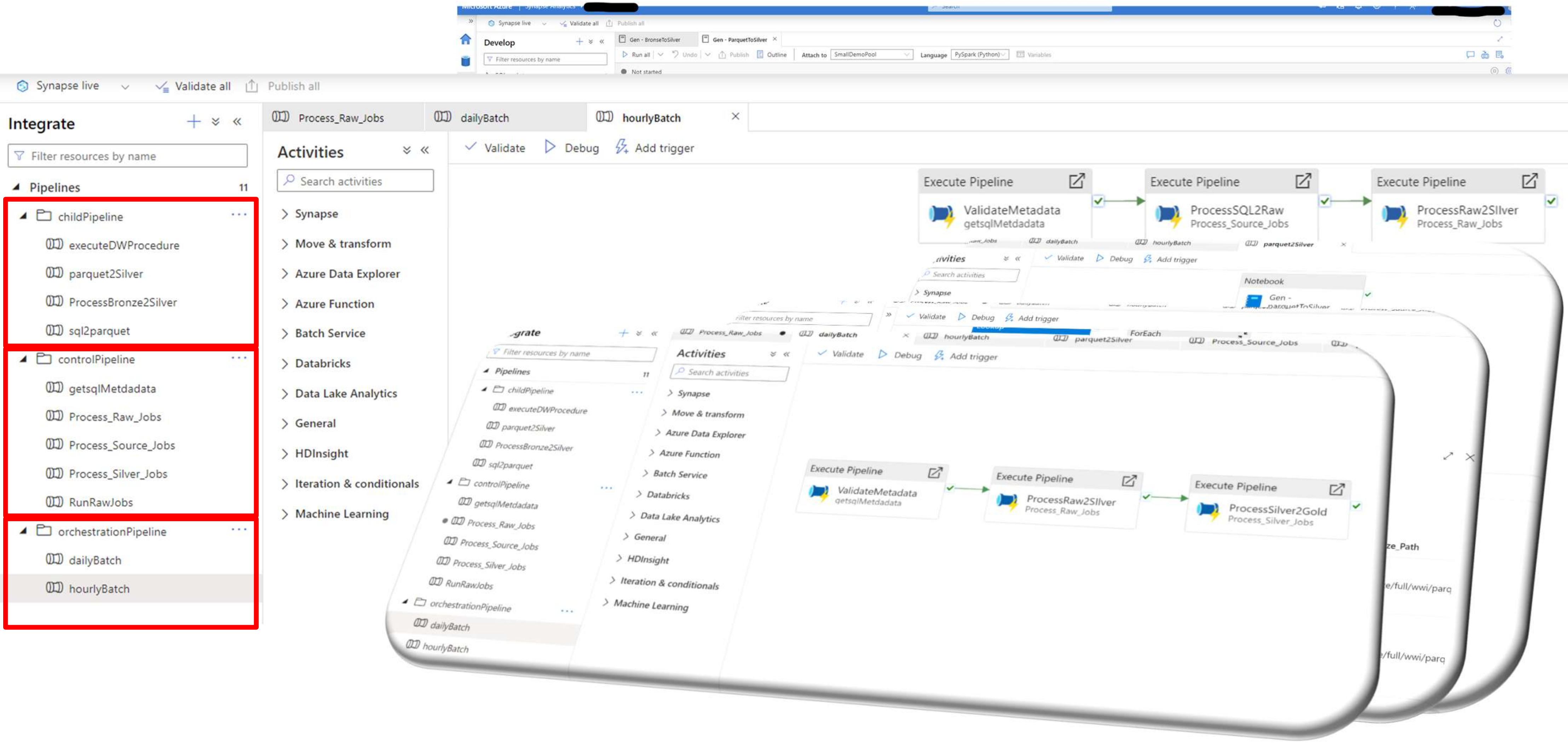
dswarm/dswarm Public archive
an open-source data management platform for knowledge workers (https://github.com/dswarm/dswarm-documentation/wiki)
metadata · json · csv · etl · mapping
Java · 57 · Updated on 19 Dec 2017

Datavault-UK/automate-dv
A free to use dbt package for creating and loading Data Vault 2.0 compliant Data Warehouses (powered by dbt, an open source data engineer...
metadata · sql · etl · snowflake · datawarehousing

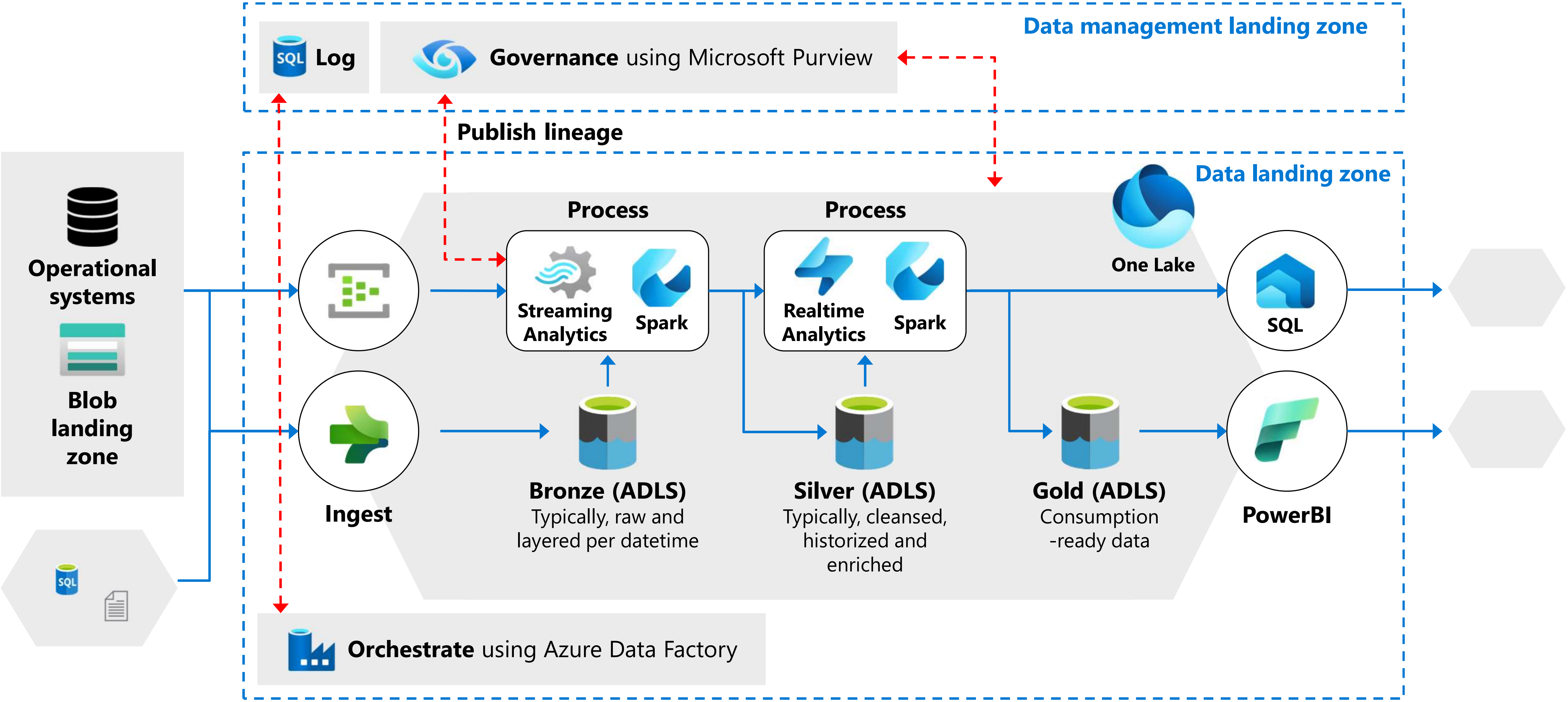
The Lakehouse Pattern in Azure



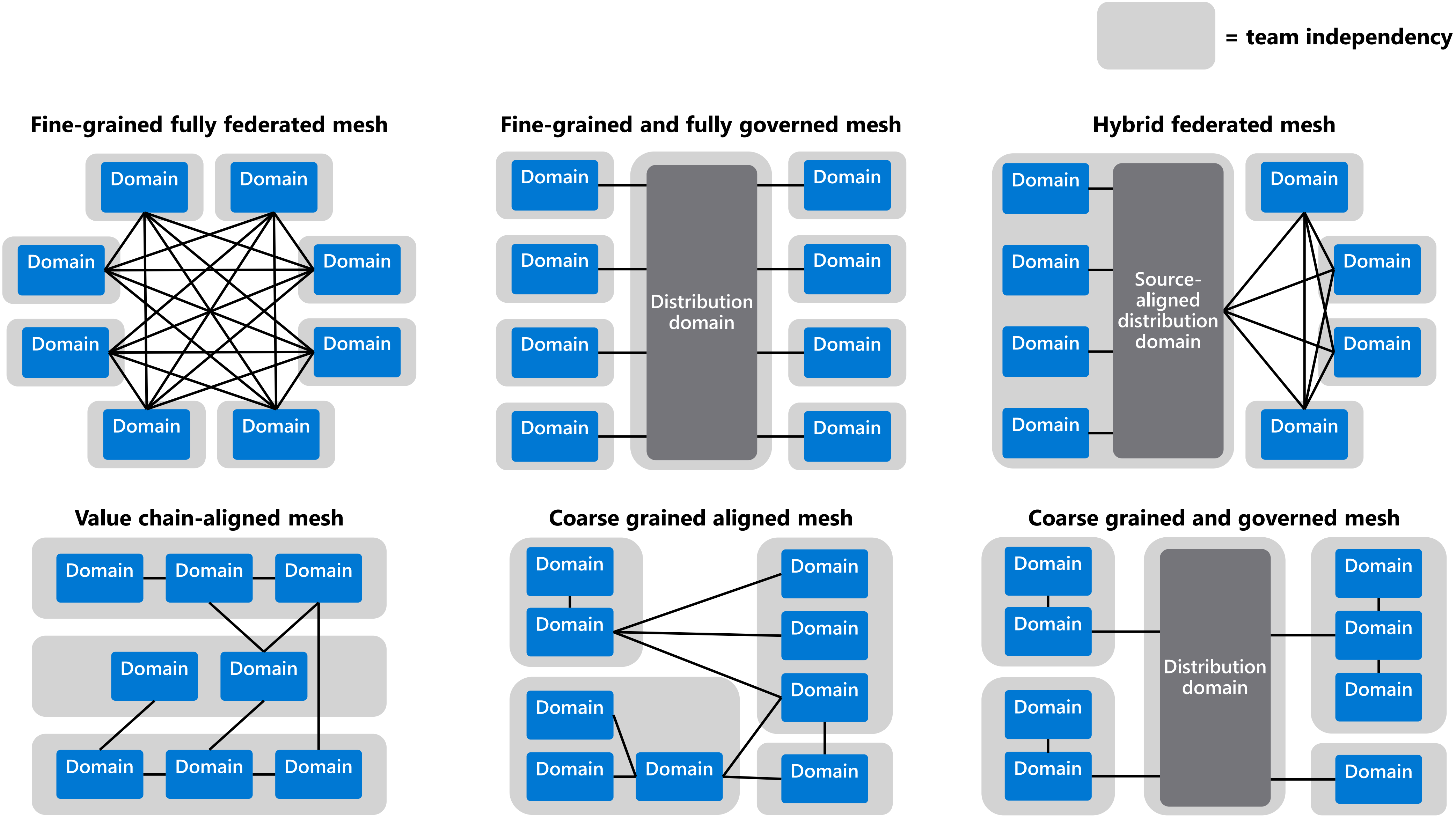
Metadata Driven Engineering Examples



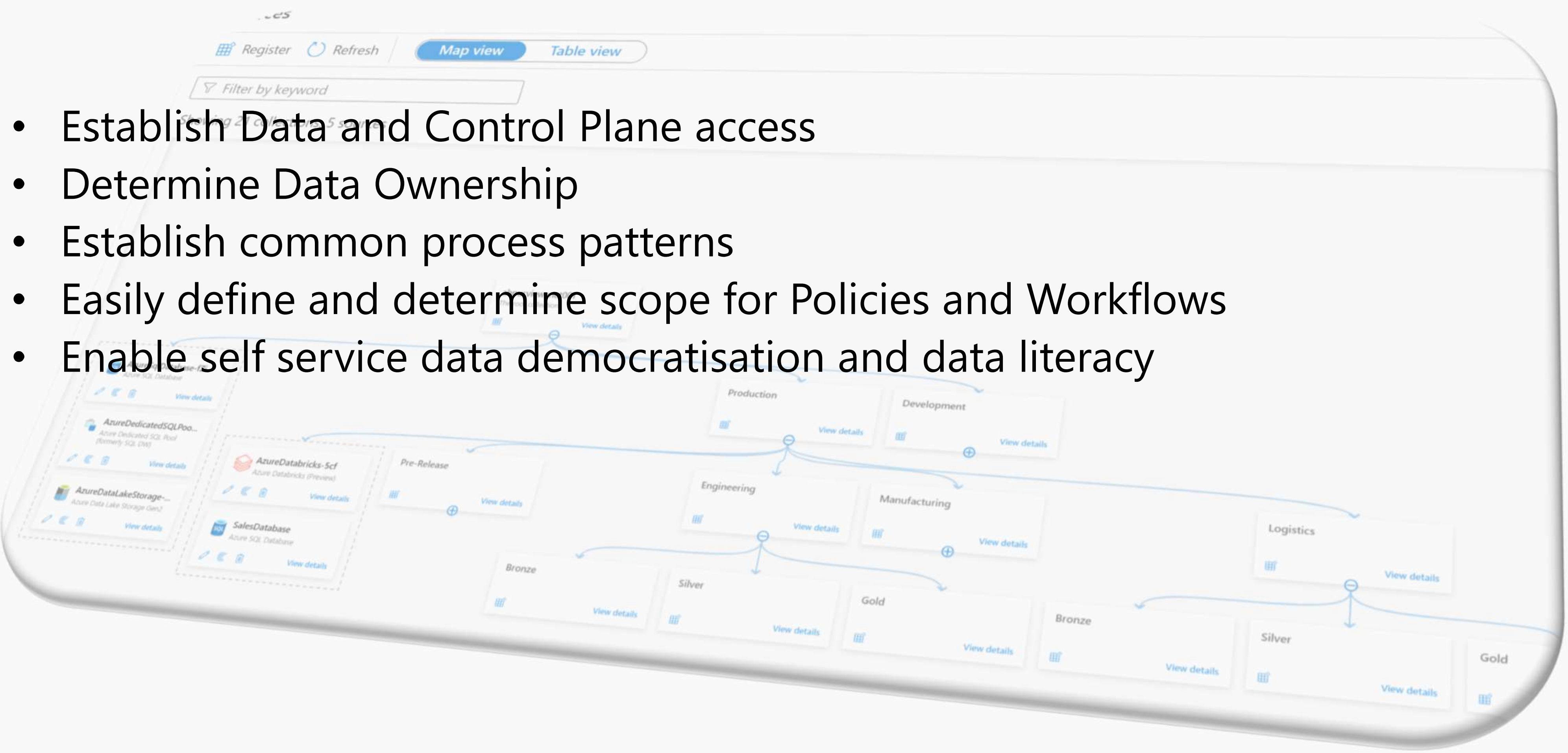
The Lakehouse Pattern in Microsoft Fabric



Driving at Scale

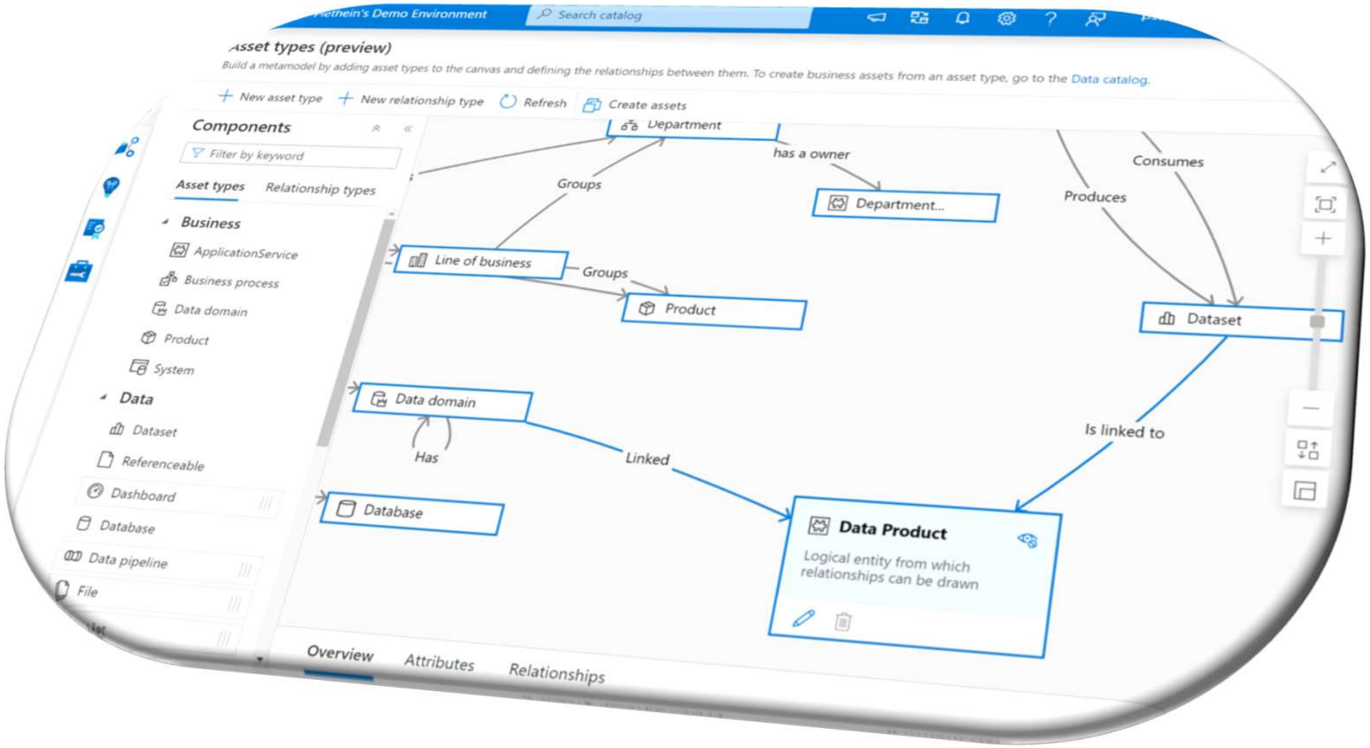

= team independency
Fine-grained fully federated mesh
Domain
Domain
Domain
Domain
Domain
Domain
Domain
Domain
Fine-grained and fully governed mesh
Domain
Domain
Domain
Domain
Distribution domain
Domain
Domain
Domain
Domain
Hybrid federated mesh
Domain
Domain
Domain
Domain
Source-aligned distribution domain
Domain
Domain
Domain
Domain
Value chain-aligned mesh
Domain
Domain
Domain
Domain
Domain
Domain
Domain
Domain
Coarse grained aligned mesh
Domain
Domain
Domain
Domain
Domain
Domain
Domain
Domain
Domain
Coarse grained and governed mesh
Domain
Domain
Domain
Domain
Distribution domain
Domain
Domain
Domain
Domain

Define the Lakehouse in Purview



- Establish Data and Control Plane access
- Determine Data Ownership
- Establish common process patterns
- Easily define and determine scope for Policies and Workflows
- Enable self service data democratisation and data literacy

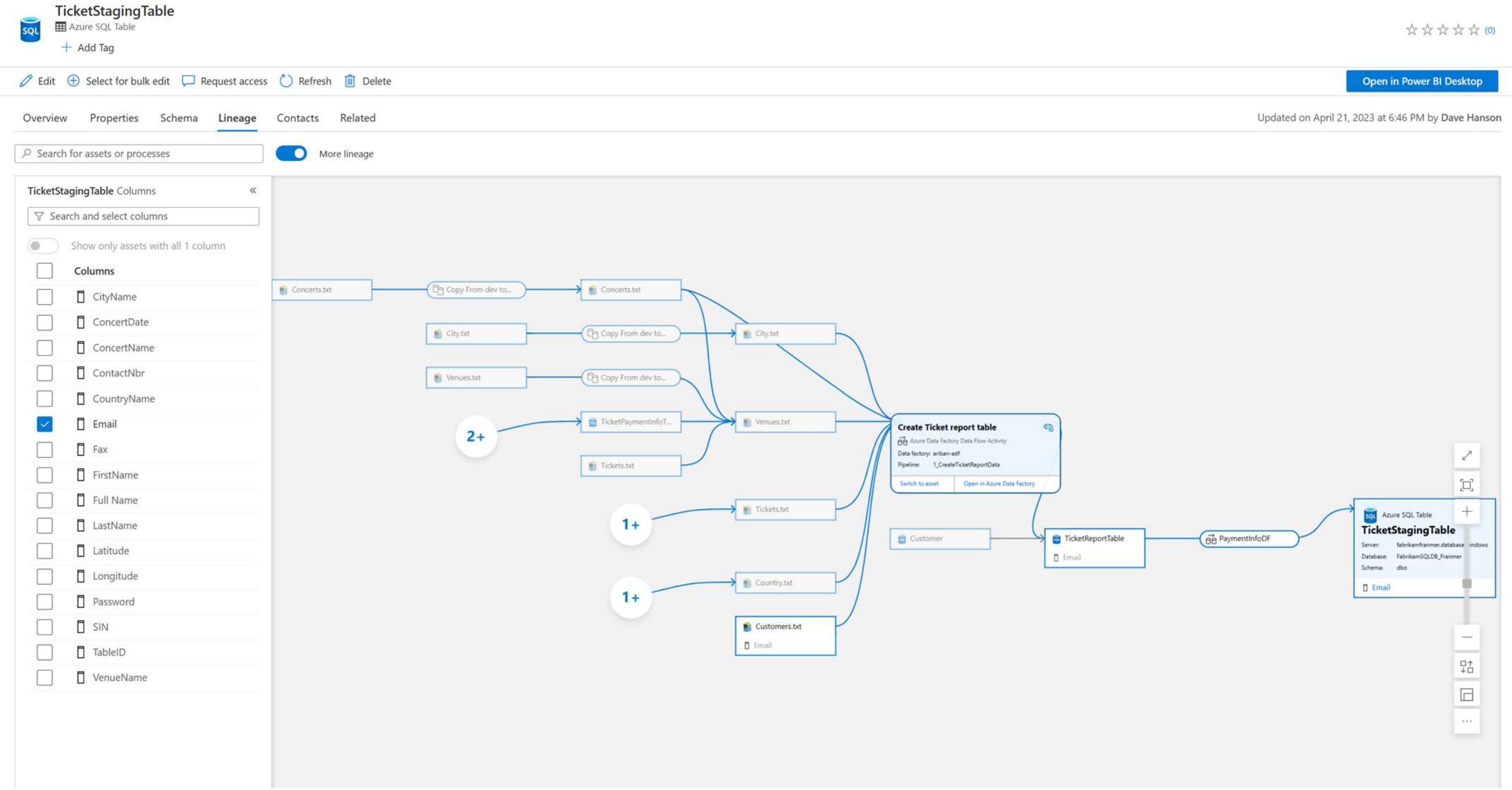
Encourage and practice good stewardship



Asset types (preview)

Build a metamodel by adding asset types to the canvas and defining the relationships between them. To create business assets from an asset type, go to the Data catalog.

Use native capabilities to enable visibility



TicketStagingTable

Azure SQL Table

+ Add Tag

☆☆☆☆☆ (0)

Updated on April 21, 2023 at 6:46 PM by Dave Hanson

More lineage: ●

TicketStagingTable Columns

○ Show only assets with all 1 column

□ Columns
- □ CityName
- □ ConcertDate
- □ ConcertName
- □ ContactNbr
- □ CountryName
- ☑ Email
- □ Fax
- □ FirstName
- □ Full Name
- □ LastName
- □ Latitude
- □ Longitude
- □ Password
- □ SIN
- □ TableID
- □ VenueName

Create Ticket report table

Azure Data Factory Data Flow Activity

Data factory: ariban-adf

Pipeline: 1_CreateTicketReportData

Azure SQL Table

TicketStagingTable

Server: fabrikamfranmer.database.windows

Database: FabrikamSQLDB_Franmer

Schema: dbo

Key take-aways

- The Lakehouse pattern is a scalable long term supportable concept

- Metadata driven data engineering provides clear data boundaries at scale

- Data engineering goes hand in hand with data governance

https://github.com/Andreas-bersgtedt/Metdata2Lakehouse