

FROM BATCH TO STREAM: UNLOCKING DATABRICKS FOR ALL YOUR ANALYTICS NEEDS

GEIR ALSTAD & VITALIJA BARTUSEVICIUTE

Thank you sponsors!

Platinum



Gold









Bronze









About us



Vitalija Bartuseviciute

Databricks MVP, Senior Consultant - Capgemini

Specializes in DogOps

E-mail: vitalija.bartuseviciute@capgemini.com



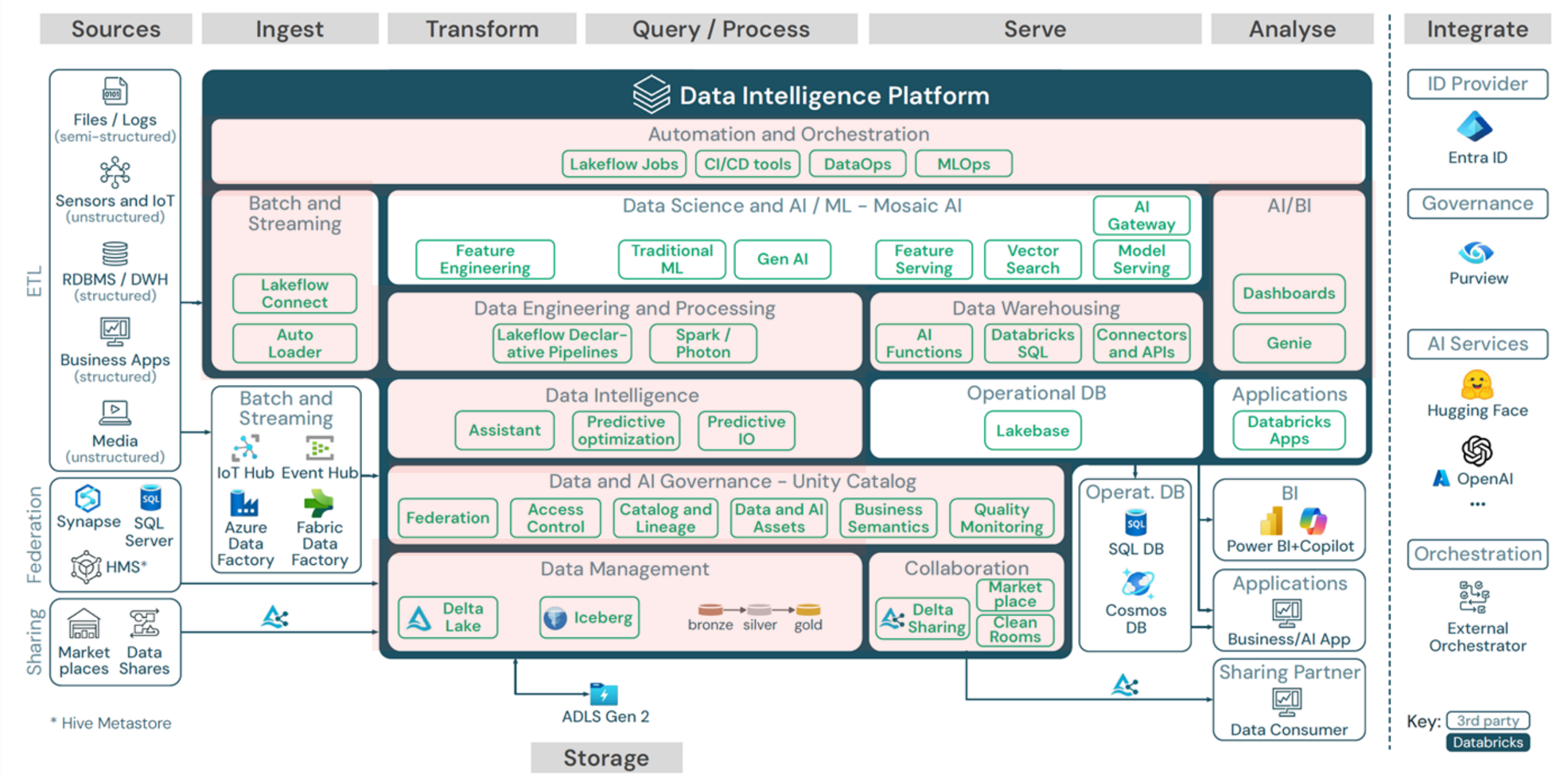
Geir Alstad

Databricks MVP, Chief Data Architect - Gabler

Specializes in DadOps

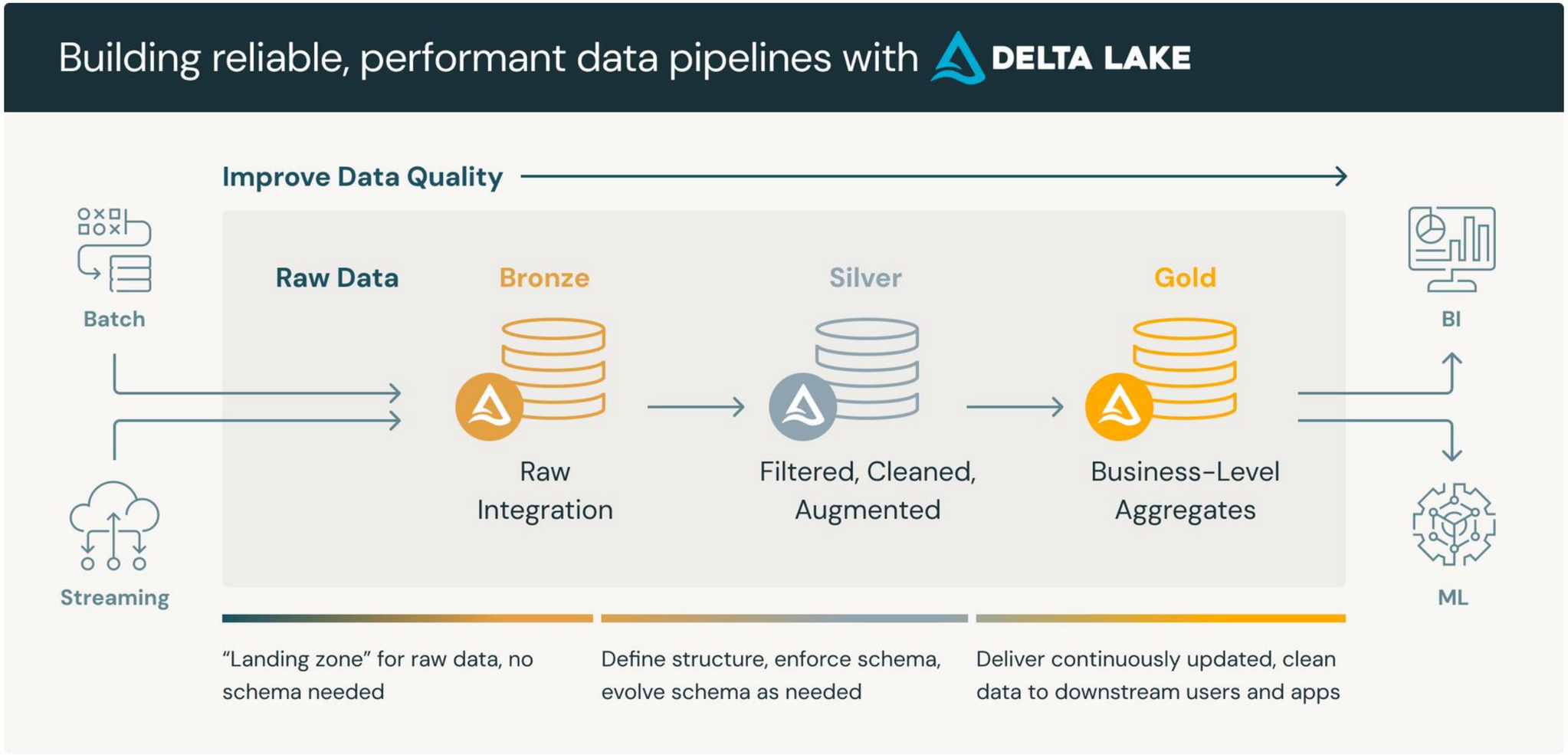
E-mail: geir.alstad@gabler.no

Azure Databricks reference architechture



Data management

Conceptual Foundation - Medallion architechture



Delta Lake Foundation

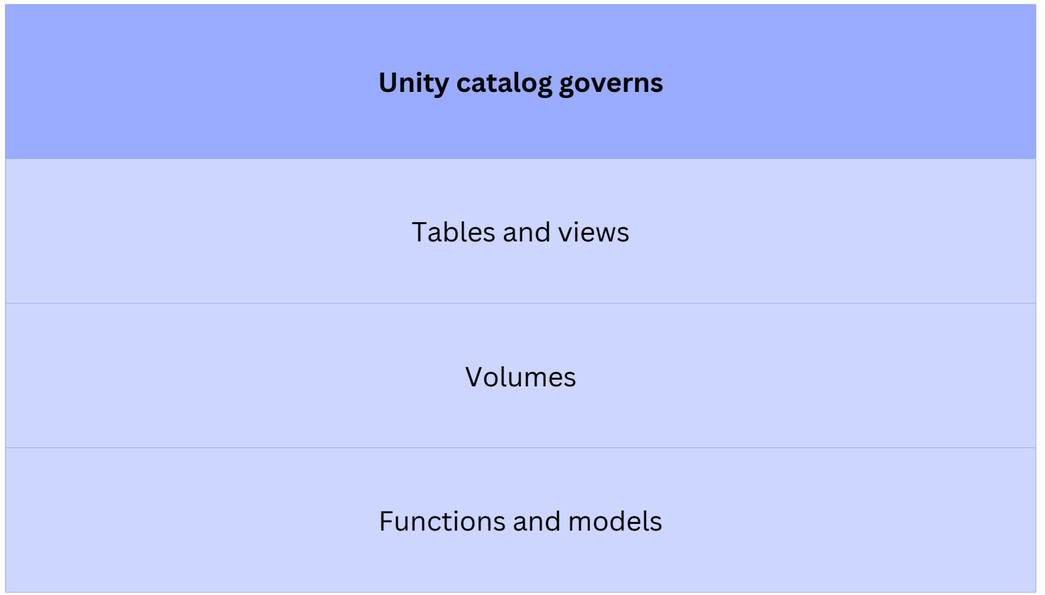


Key features
ACID transactions for reliability
Time travel for audit and recovery
Schema evolution without breaking pipelines
Open format (Parquet-based, no vendor lock-in)

Data governance and intelligence

Unity Catalog - Single Control Plane

```
Catalog --> Schema --> Object
```



Unity Catalog - Key Capabilities

Capabilty	Benefit
Fine-grained access control	Table, column, row-level permissions
Automatic lineage	Source-to-dashboard tracking
Audit logging	Complete access record
Cross-workspace	Share without duplication

Data intelligence

Assistant	Predictive optimization	Predictive I/O
AI coding companion that understands your schema	Automatic table maintenance	Intelligent caching (enabled by default)
Generates SQL/Python from natural language	OPTIMIZE, VACUUM, ANALYZE run automatically	Prefetching data based on predicted access patterns
Context-aware suggestions	Liquid Clustering adapts to query patterns	Use to aid query monitoring

Data Engineering and Processing

Declarative Pipelines Overview

Concept	Purpose
Streaming Tables	Real-time, incremental updates
Materialized Views	Batch-optimized datasets
Flows	Multiple sources --> single target
Sinks	Write to external systems

Example code: Auto Loader + Data Quality

```python
from pyspark import pipelines as dp

@dp.table(name="bronze_events")
def bronze_events():
    return (
        spark.readStream.format("cloudFiles")
        .option("cloudFiles.format", "json")
        .option("cloudFiles.schemaEvolutionMode", "addNewColumns")
        .load("/Volumes/catalog/schema/landing/events")
    )

@dp.materialized_view(name="silver_events")
@dp.expect_or_drop("valid_id", "event_id IS NOT NULL")
def silver_events():
    return spark.read.table("bronze_events")
```

CI/CD with Databricks Asset Bundles

```
bundle:
  name: event_processing

variables:
  catalog: { default: dev_catalog }

resources:
  pipelines:
    ingestion:
      target: "${var.catalog}.events"

targets:
  dev:  { variables: { catalog: dev_catalog } }
  prod: { variables: { catalog: prod_catalog } }
```

Parameterized notebooks

```
# Define widgets
dbutils.widgets.text("catalog", "dev_catalog")
dbutils.widgets.dropdown("mode", "append", ["append",
"overwrite"])

# Use parameters
catalog = dbutils.widgets.get("catalog")
target = f"{catalog}.events.processed"
df.write.mode(mode).saveAsTable(target)
```

Datawarehousing and Serving

Databricks SQL

	Best For	Scaling
Serverless	Variable workloads, cost optimization	Automatic, pay-per-use
Pro	Predictable workloads, advanced features	Manual/Auto
Classic	Legacy compatibility	Manual

AI Functions

Function	Purpose	Example use case
ai_query	Apply any LLM model to data	Everything, all of it
ai_classify	Categorize text into labels	Classify support tickets by urgency
ai_extract	Extract structured data from text	Pull entities from unstructured logs
ai_mask	Redact sensitive information	Anonymize PII in text fields
ai_forecast	Forecast data up to a specified horizon	TVF to extrapolate time series data into the future.

AI Functions: Example Use

```sql
SELECT
    event_id,
    ai_classify(
        event_description,
        ARRAY('security', 'performance', 'user_error')
    ) AS category
FROM silver.events
LIMIT 100;
```

Time travel and restore

```sql
-- Query data as it was yesterday
SELECT * FROM gold.events_summary
TIMESTAMP AS OF '2026-01-10 14:00:00';

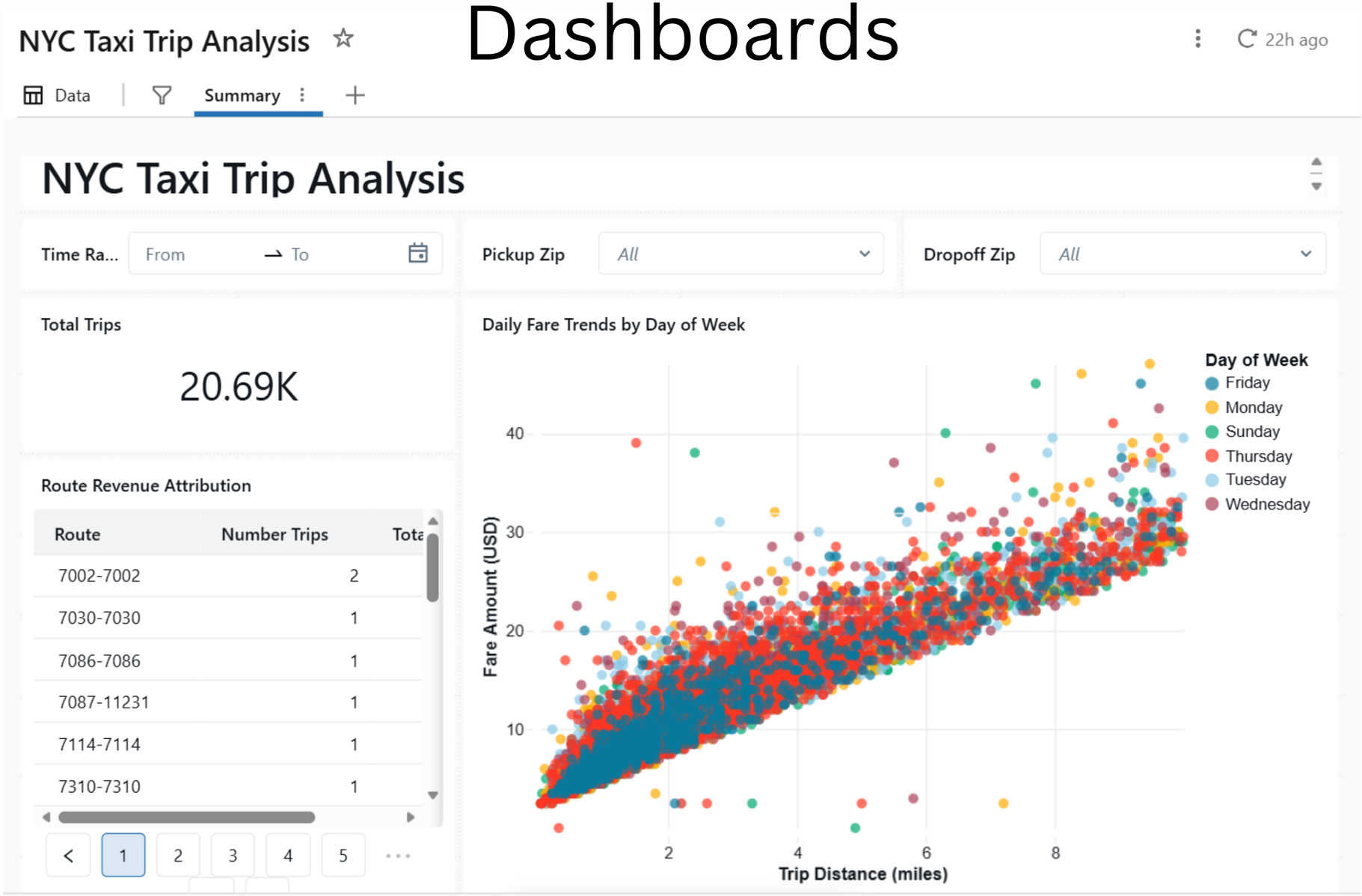
-- View modification history
DESCRIBE HISTORY gold.events_summary;

-- Restore if needed
RESTORE TABLE gold.events_summary TO VERSION AS OF 42;
```

AI/BI Dashboards

Feature	Description
Visual editor	Drag-and-drop chart creation
SQL-based	Build visualizations from query results
Scheduling	Automated refresh and email delivery
Embedding	Integrate dashboards in external applications
Sharing	Collaborate with team members securely

AI/BI Dashboards



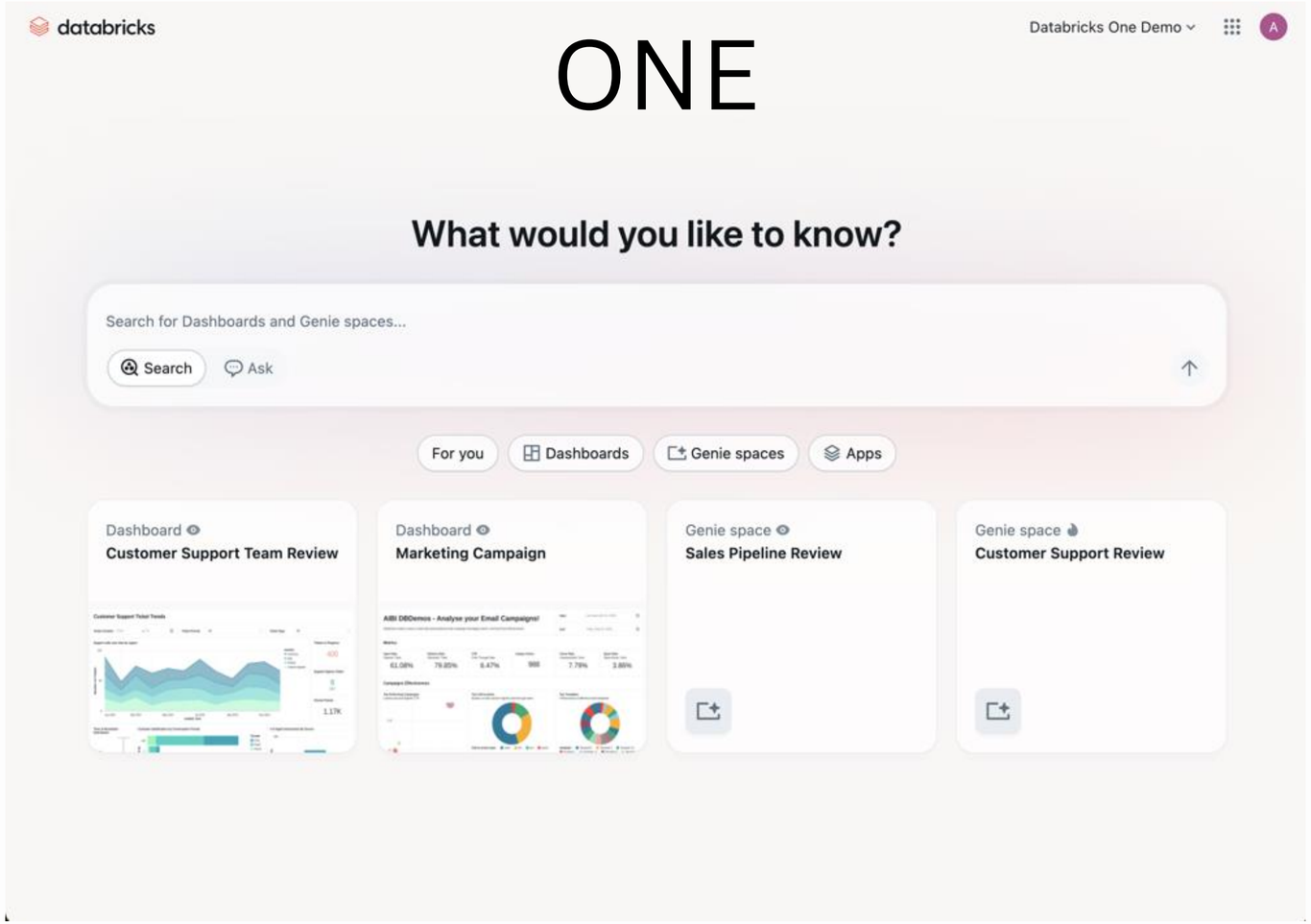
Genie

...Natural language to SQL...

User: "Show me top 10 event types by volume this week"

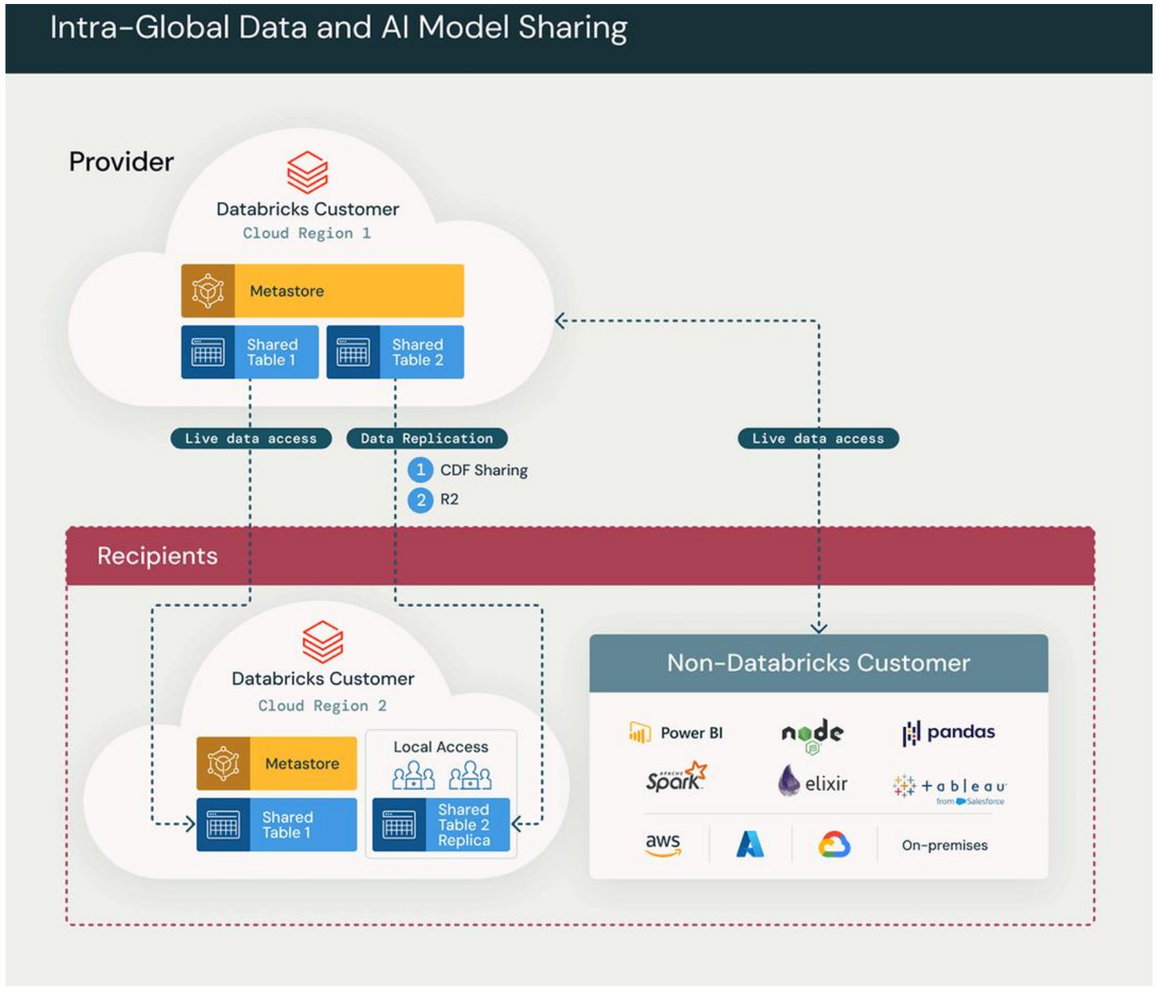
```sql
SELECT event_type, COUNT(*) as count
FROM gold.events_summary
WHERE event_date >= DATE_SUB(CURRENT_DATE(), 7)
GROUP BY event_type
ORDER BY count DESC
LIMIT 10;
```

Databricks ONE



Collaboration

Delta Sharing



Other collaboration capabilities

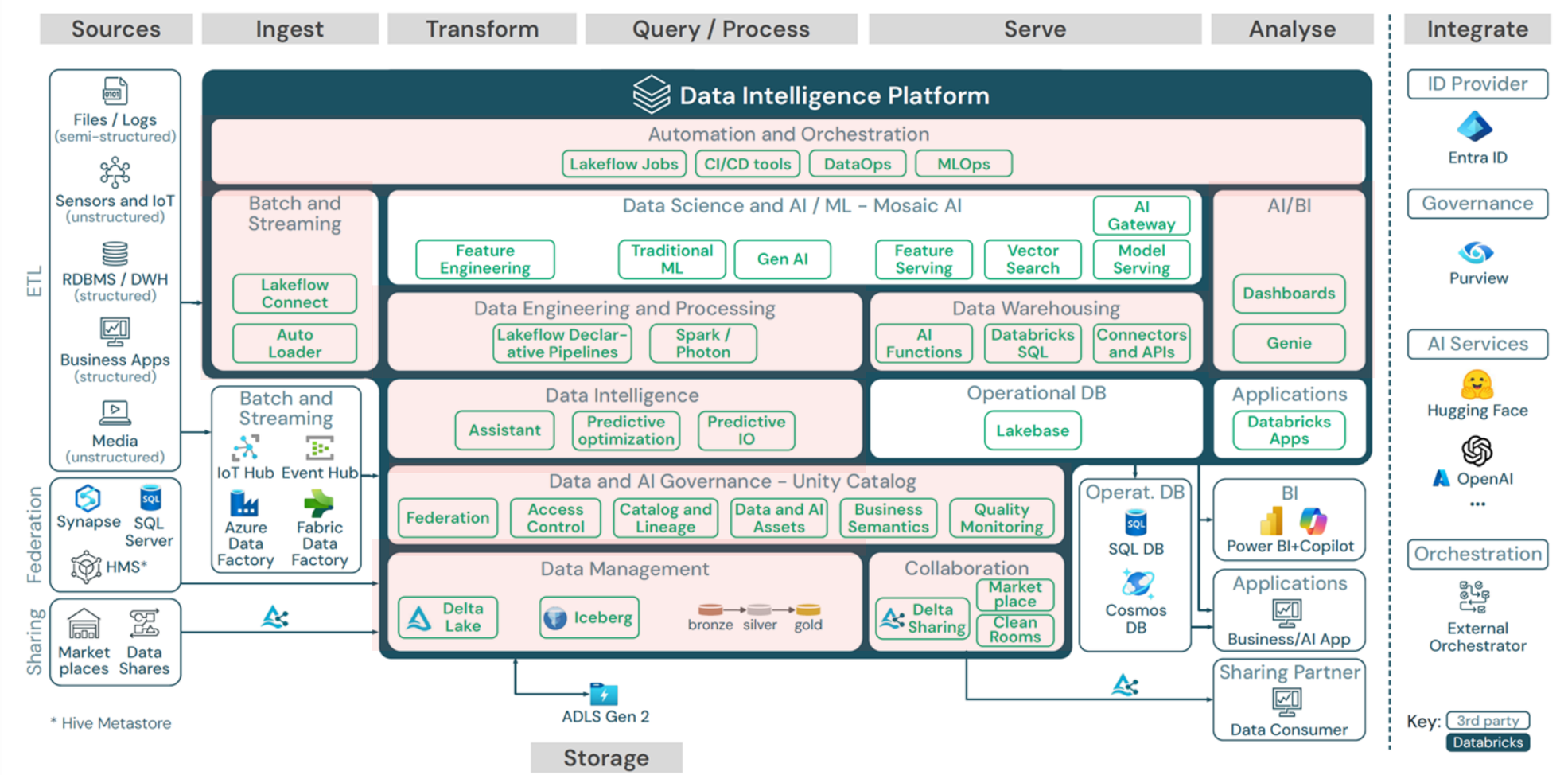
Marketplace	Clean Rooms
Discover and consume data products	Privacy-preserving collaboration
One-click integration	Raw data never leaves your control
Appears as Unity Catalog objects	Only approved analyses, aggregated results
	GDPR/DORA compliant

Takeaways

Key capabilities by persona

Data Engineers	Data Analysts	Decision Makers
Unified Platform	Familiar SQL	Open Standards
Lakeflow	AI Functions	Centralized Governance
DABs	Genie	Secure Collaboration

Summary


Sources
Ingest
Transform
Query / Process
Serve
Analyse
Integrate
Data Intelligence Platform
Automation and Orchestration
Lakeflow Jobs
CI/CD tools
DataOps
MLOps
ETL
Files / Logs (semi-structured)
Sensors and IoT (unstructured)
RDBMS / DWH (structured)
Business Apps (structured)
Media (unstructured)
Batch and Streaming
Lakeflow Connect
Auto Loader
Data Science and AI / ML - Mosaic AI
Feature Engineering
Traditional ML
Gen AI
AI Gateway
Feature Serving
Vector Search
Model Serving
AI/BI
Dashboards
Genie
Data Engineering and Processing
Lakeflow Declar-ative Pipelines
Spark / Photon
Data Warehousing
AI Functions
Databricks SQL
Connectors and APIs
Federation
Synapse
SQL Server
HMS*
Batch and Streaming
IoT Hub
Event Hub
Azure Data Factory
Fabric Data Factory
Data Intelligence
Assistant
Predictive optimization
Predictive IO
Operational DB
Lakebase
Applications
Databricks Apps
Data and AI Governance - Unity Catalog
Federation
Access Control
Catalog and Lineage
Data and AI Assets
Business Semantics
Quality Monitoring
Operat. DB
SQL DB
Cosmos DB
BI
Power BI+Copilot
Sharing
Market places
Data Shares
Data Management
Delta Lake
Iceberg
bronze
silver
gold
Collaboration
Delta Sharing
Market place
Clean Rooms
Applications
Business/AI App
Sharing Partner
Data Consumer
* Hive Metastore
ADLS Gen 2
Storage
ID Provider
Entra ID
Governance
Purview
AI Services
Hugging Face
OpenAI
...
Orchestration
External Orchestrator
Key:
3rd party
Databricks

THANK YOU!

Feedback

Vitalija Bartuseviciute

vitalija.bartuseviciute@capgemini.com

Geir Alstad

geir.alstad@gabler.no