Flox
Data Platform

Designed for future

- Real time first
- Decoupled by design
- Data as product

Data Platform – Use Case

Use Case	User	Challenge	Platform Advantage	Success Criteria
Early Disease Detection Model Development	ML Engineer	Slow data/model workflow	Unified data, context, rapid train/deploy, monitoring	Faster deployment, more models, better performance
Self-Service Data Discovery & Pipeline	Data Engineer/Scientist	Siloed data, slow pipeline creation	Central catalogue, templates, automation, quality checks	More pipelines, faster onboarding, higher satisfaction
	Data Consumers	Disconnected datasets	Data product approach	Well defined datasets with versioning, unified access and metadata.
Model Decision Transparency	Product/CS/DS	Lack of explainability for predictions	Explainability reports, context, dashboards, feedback	All products explainable, faster answers, fewer escalations

- ***POC adoption***
  - *Self-Service Data Discovery: Data Cataloging*
  - *Self-Service Data Discovery: NLP*
  - *Pipeline Development: Spark declarative framework*
  - *Early disease detection: Stream based data processing*

Data Platform – Data Intelligence and Model

Characteristic	Description	Example
Schema definition and Evolution	Maintains a registry of all data schemas (e.g., camera, sensor, operational logs) with versioning	**Initial schema**: farm_no, timestamp, temperature **Evolved schema**: farm_no, timestamp, temperature, humidity
	Supports backward-compatible changes (e.g., adding new columns) and tracks schema lineage. When a source schema changes, Genie automatically updates downstream schemas and notifies affected users and pipelines.	
Mapping Heterogeneous Data Sources	Integrate and Ingests data from CSVs, IoT sensors, camera metadata, and operational databases	**Sensor feed**: farm_id, temp, hum **Camera feed**: farm_no, image_id, timestamp **Mapping farm_id to farm_no and aligns timestamps.**
	Uses mapping rules and transformation logic to align fields from different sources to a unified canonical model	
Semantic Enrichment	Maintains a glossary of terms (e.g., “bird_activity”, “sound_level”) and their relationships	• Adds a location_type tag to data from a specific shed. • Converts temperature from Fahrenheit to Celsius if needed.
	Automatically tags data with semantic labels (e.g., “environmental”, “operational”), units, and context (e.g., farm location, shed number).	
Downstream Intelligence Products	Exposes curated, semantically-enriched tables and views for ML models, dashboards, and analytics	Dashboard queries for “average bird activity by farm and shed”, automatically getting harmonised, context-rich data
	Consumers, receive not just raw values, but also context (e.g., schema version, data source, semantic tags).	

- ***POC adoption***
  - *Schema Definition and evolution: Autoloader*
  - *Downstream Intelligent Products: Aggregates and joined data table.*

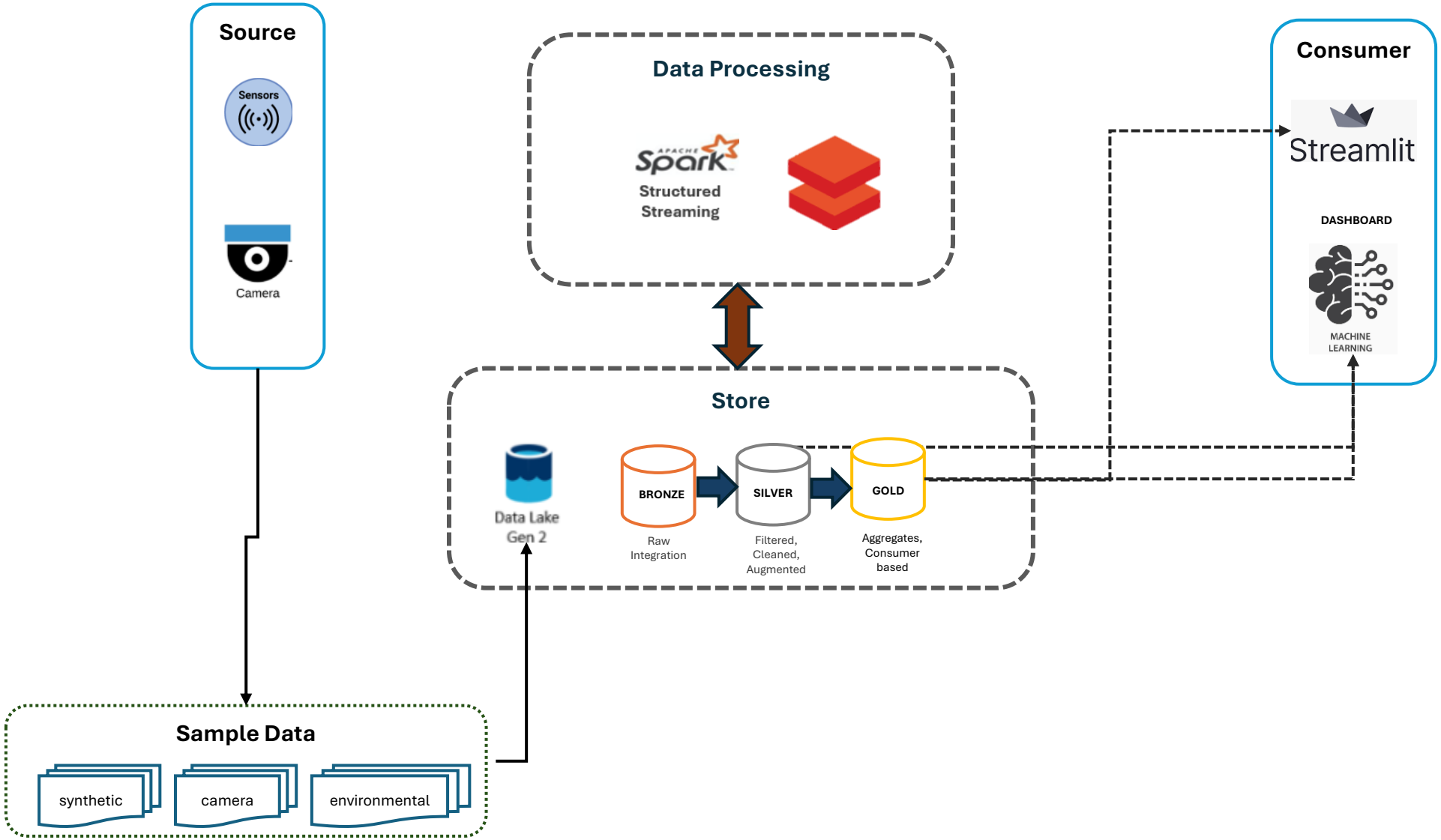
Data Solution – Scope & Trade Offs

Characteristic	Assumption	Constraint
Scope	Data Integration	Sample files provided
	Data Quality	Unavailability of data quality rules and governance policy.
	ML and AI	Time & Effort
Trade offs	Databricks vs Spark	• Spark require manual setup, coding and tunning • Databricks better for enterprise level workflows providing management, governance and monitoring. • Databricks integrates and provides single platform for Data engineering, ML and AI
	Data Zones	• Incremental processing • Cater to Analytics, ML and AI use cases
	Cloud based	• Flexible cost-based optimisation • Support of data volume and security • BC & DR

Data Platform – Assumptions & Constraints

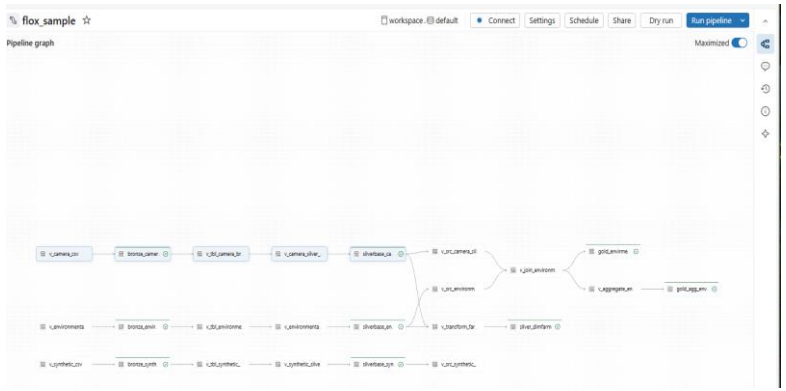
Characteristic	Assumption	Constraint	Implementation
Data Scale and Latency	Handle data from hundreds of farms, each generating multi-modal data (camera, sensor, operational) at minute-level or higher frequency.	The platform is optimised for batch and micro-batch processing (latency of minutes), not real-time sub-second analytics	Storage and compute resources are scalable, and initial deployments will be cloud based.
User Personas and Skill Levels	Primary users are data scientists, ML/AI engineers, data engineers, and product managers	User skill levels vary—some are proficient in SQL and Python, others prefer no-code/low-code interfaces	Provide both code-based APIs and user-friendly UIs for data discovery, pipeline creation, and model deployment.
Operational or Regulatory Considerations	Data may include sensitive operational or animal welfare information subject to privacy, security, and agricultural regulations.	Platform must be able to support compliance with relevant standards (e.g., GDPR, food safety, animal welfare reporting).	Support role-based access control, data lineage, and audit trail

POC – Solution Architecture



POC– Attributes

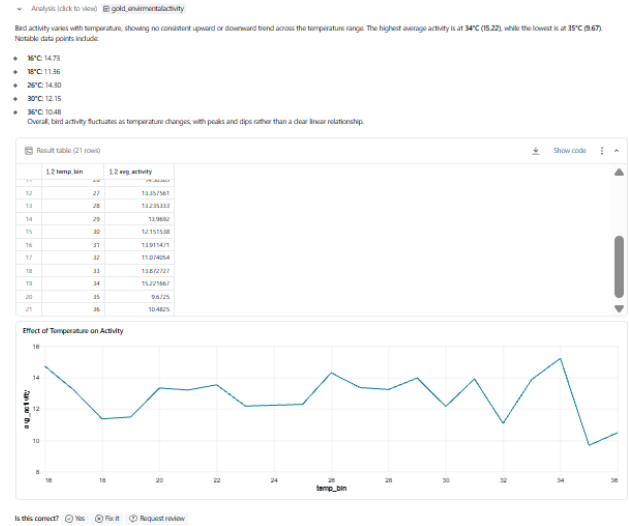
DAG



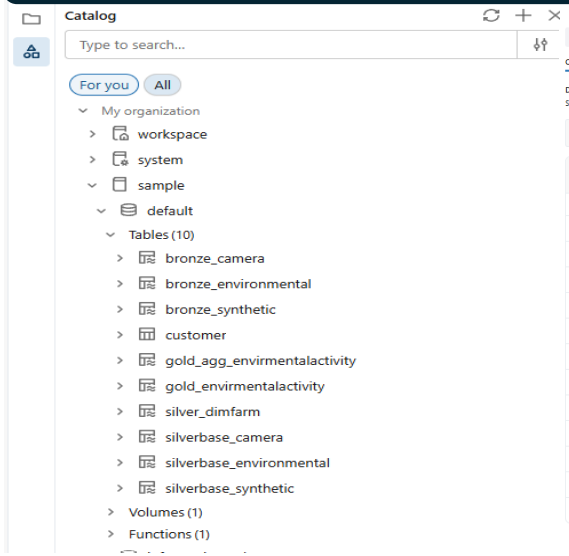
PERFORMANCE

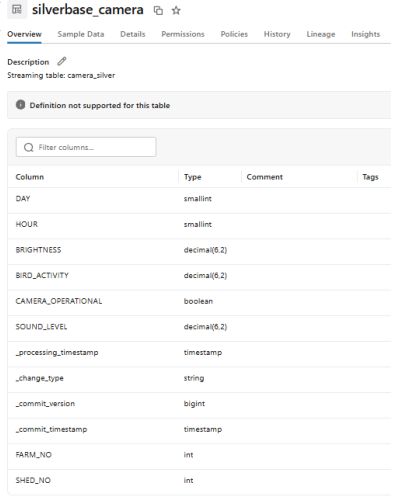
Statement	Started At	Duration	Rows read	Bytes read	Bytes written
REFRESH STREAMING TABLE sample.default.silver_dimfarm /* FLOW sample.default.silver_dimfarm */	Feb 13, 2026, 11:43 AM	7 s 470 ms	4	38.13 KB	1.25 KB
REFRESH STREAMING TABLE sample.default.gold_envirmentalactivity /* FLOW sample.default.gold_envirm...	Feb 13, 2026, 11:43 AM	7 s 632 ms	4,608	38.13 KB	20.64 KB
REFRESH STREAMING TABLE sample.default.gold_agg_envirmentalactivity /* FLOW sample.default.gold_ag...	Feb 13, 2026, 11:43 AM	7 s 684 ms	2,400	38.13 KB	3.30 KB
REFRESH STREAMING TABLE sample.default.silverbase_environmental /* FLOW sample.default.silverbase_...	Feb 13, 2026, 11:43 AM	16 s 229 ms	3,458	0 B	19.27 KB
REFRESH STREAMING TABLE sample.default.silverbase_synthetic /* FLOW sample.default.silverbase_synt...	Feb 13, 2026, 11:43 AM	17 s 273 ms	38	0 B	4.87 KB
REFRESH STREAMING TABLE sample.default.silverbase_camera /* FLOW sample.default.silverbase_camera ...	Feb 13, 2026, 11:43 AM	17 s 973 ms	3,458	0 B	18.92 KB
REFRESH STREAMING TABLE sample.default.bronze_environmental /* FLOW workspace.default.tbl_environm...	Feb 13, 2026, 11:43 AM	8 s 730 ms	1,152	20.50 KB	12.19 KB
REFRESH STREAMING TABLE sample.default.bronze_camera /* FLOW workspace.default.tbl_camera_bronze */	Feb 13, 2026, 11:43 AM	9 s 463 ms	1,152	31.39 KB	17.96 KB
REFRESH STREAMING TABLE sample.default.bronze_synthetic /* FLOW workspace.default.tbl_synthetic_br...	Feb 13, 2026, 11:43 AM	10 s 347 ms	12	94 B	1.72 KB

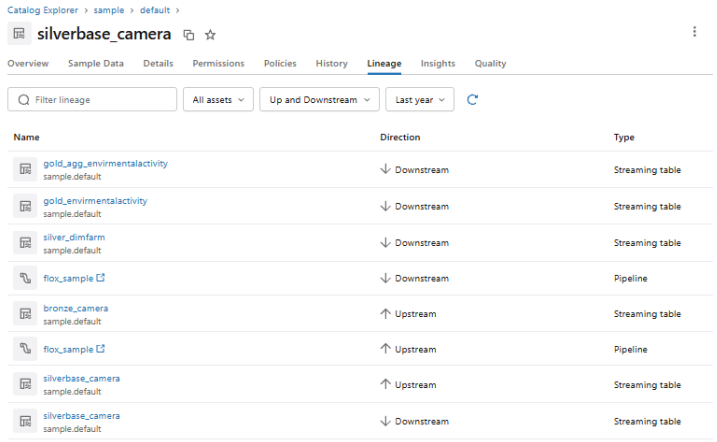
NLP



Governance







POC– Features & Opportunities

Features

Incremental Data Processing
Ensure data governance, historical data auditing, and support diverse data workloads like BI, Machine Learning.

Real time data processing
Enable immediate action and insights. Batch and micro batch-based processing support.

Data Lineage
Automated data lineage and data dependency management. DAG based data entity processing.

Data observation and monitoring
Data processing statistics, insights and metadata management.

Layered and modular approach
Enables rapid adoption of new technologies based on business needs supporting parallel development and experimentation.

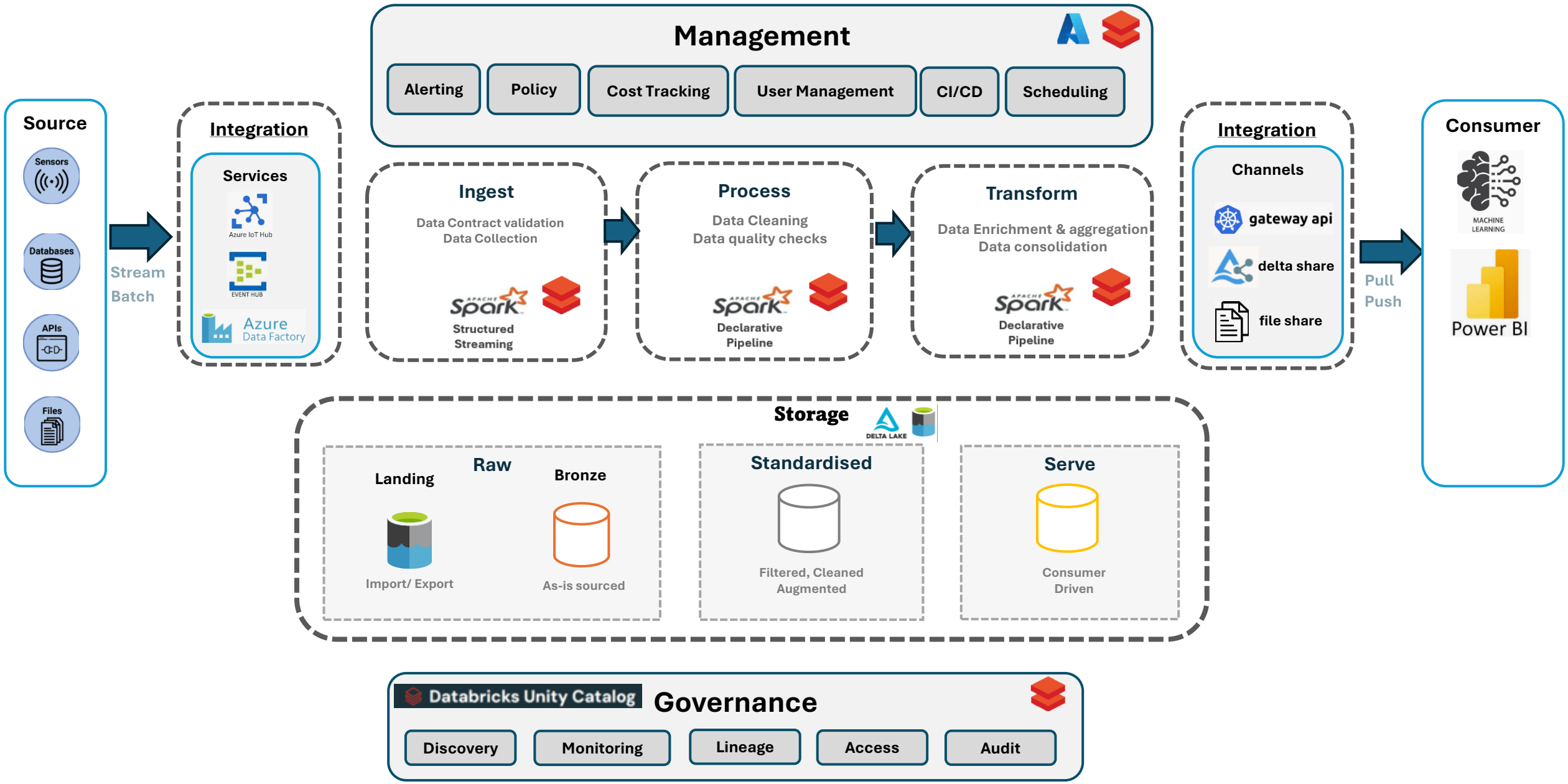
Opportunities

Integration Layer
Facilitate source data integration services for IOT, API, Sensor, and files. Make data available through multiple channels for diverse consumers.

MLOPS
Combine data and Machine learning as well as AI into one streamline artifacts and deployment management

Data Expectations
Implement declarative data quality conditions for improved data quality

Data Platform – Architecture



Data Platform – Roadmap