
AI
AI
Databricks
AI
NIKHIL SHARMA

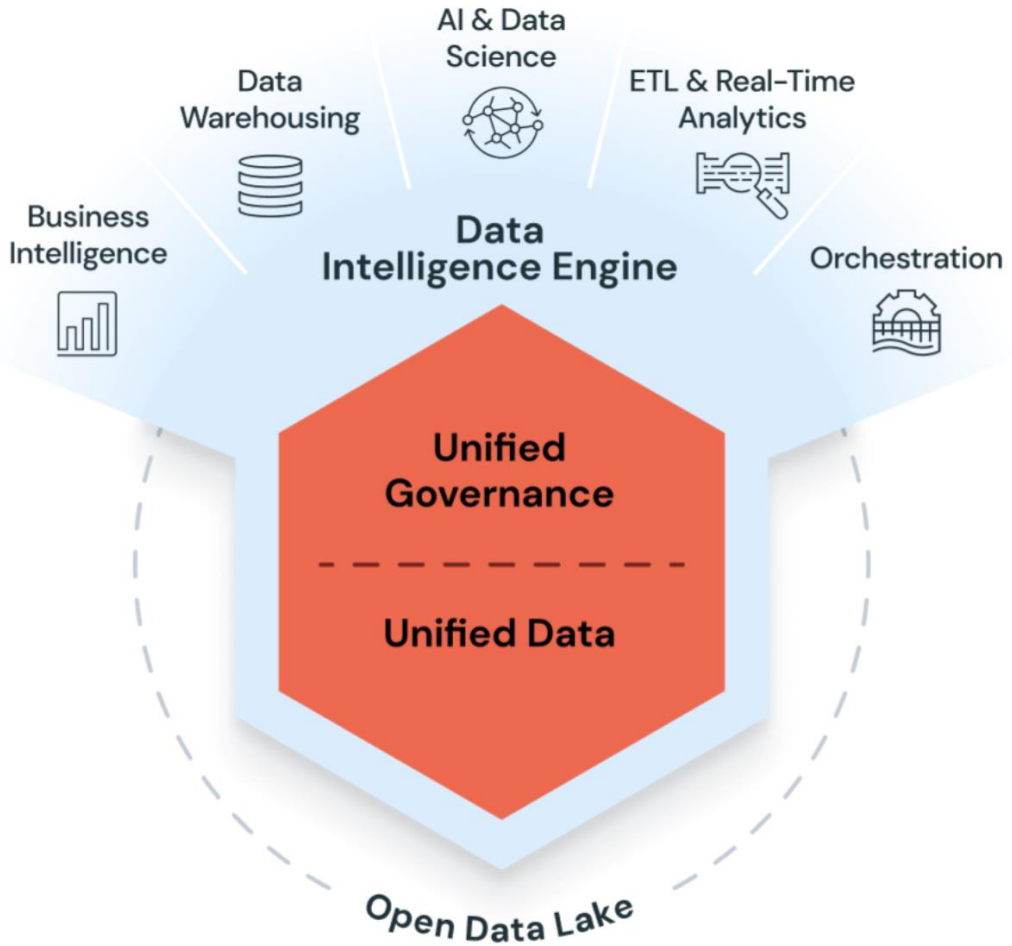
- Fully-managed data lake analytics platform
- Based on open-source technologies
- Integrated with Azure for resource management and security

Azure Databricks is a cloud-based data analytics platform that provides a unified environment for data engineering, machine learning, and analytics

Databricks

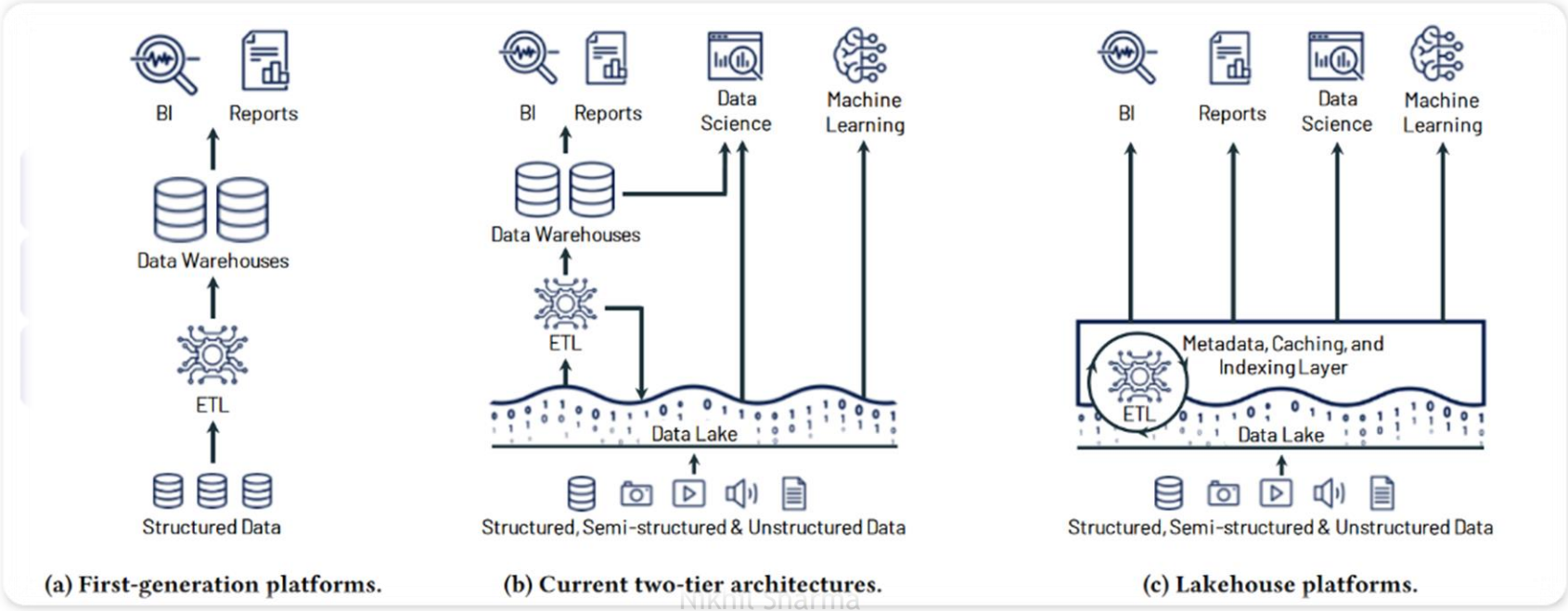
Databricks is a cloud-based platform that serves as a one-stop shop for all data needs, such as storage and analysis. It was created by the people behind Apache Spark. Databricks can generate insights with SparkSQL, link to visualization tools like Power BI, Qlikview, and Tableau, and develop predictive models with SparkML. You can also use Databricks to generate tangible interactive displays, text, and code. One could say it's a MapReduce system alternative.

Databricks is a unified, open analytics platform for building, deploying, sharing, and maintaining enterprise-grade data, analytics, and AI solutions at scale. The Databricks Data Intelligence Platform integrates with cloud storage and security in your cloud account and manages and deploys cloud infrastructure for you.

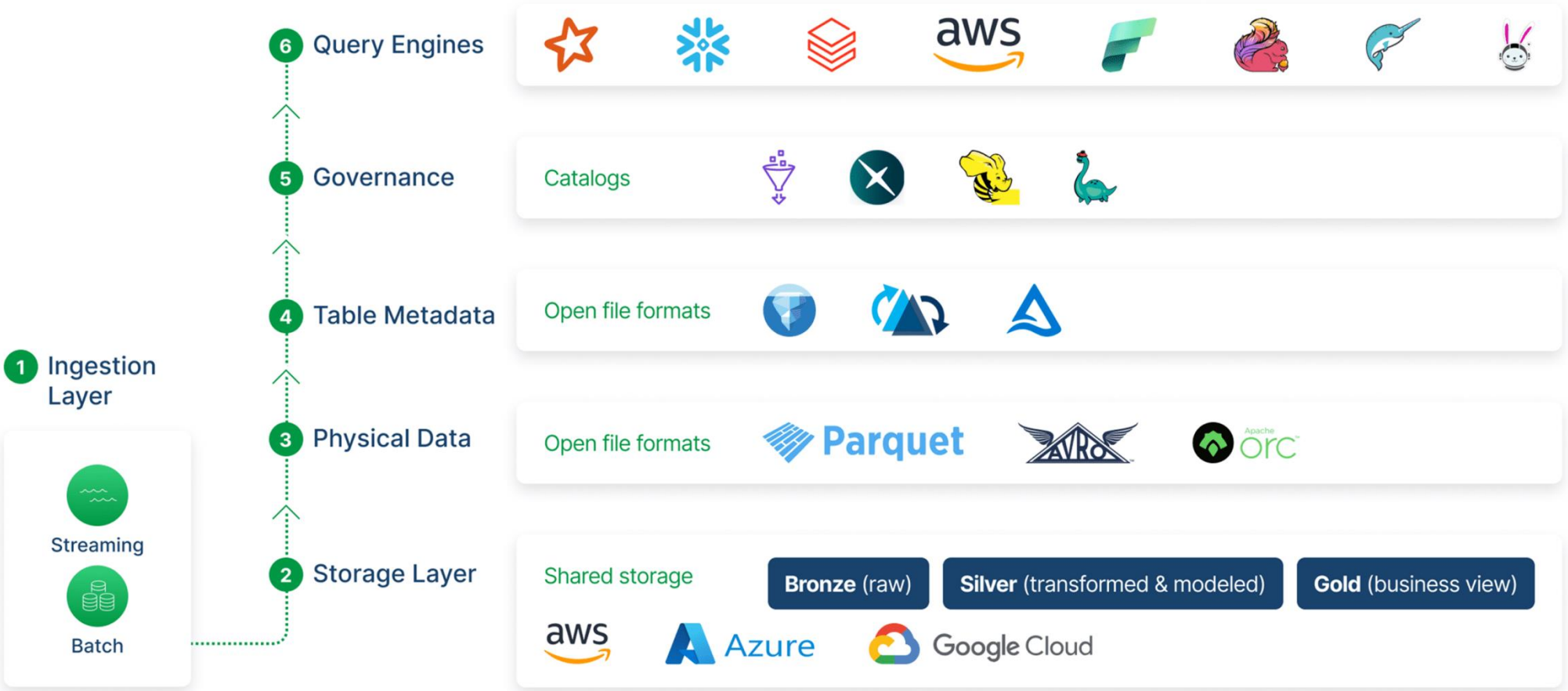


Lakehouse

A data lakehouse is a data management architecture that combines key capabilities of data lakes and data warehouses into a unified platform. It brings the benefits of a data lake, such as low-cost storage and broad data access, and the benefits of a data warehouse, such as data structure, performance, and management features. Lakehouses are increasingly built utilizing open data and open table formats such as Apache Iceberg, Hudi and Delta tables to provide flexibility and interoperability.

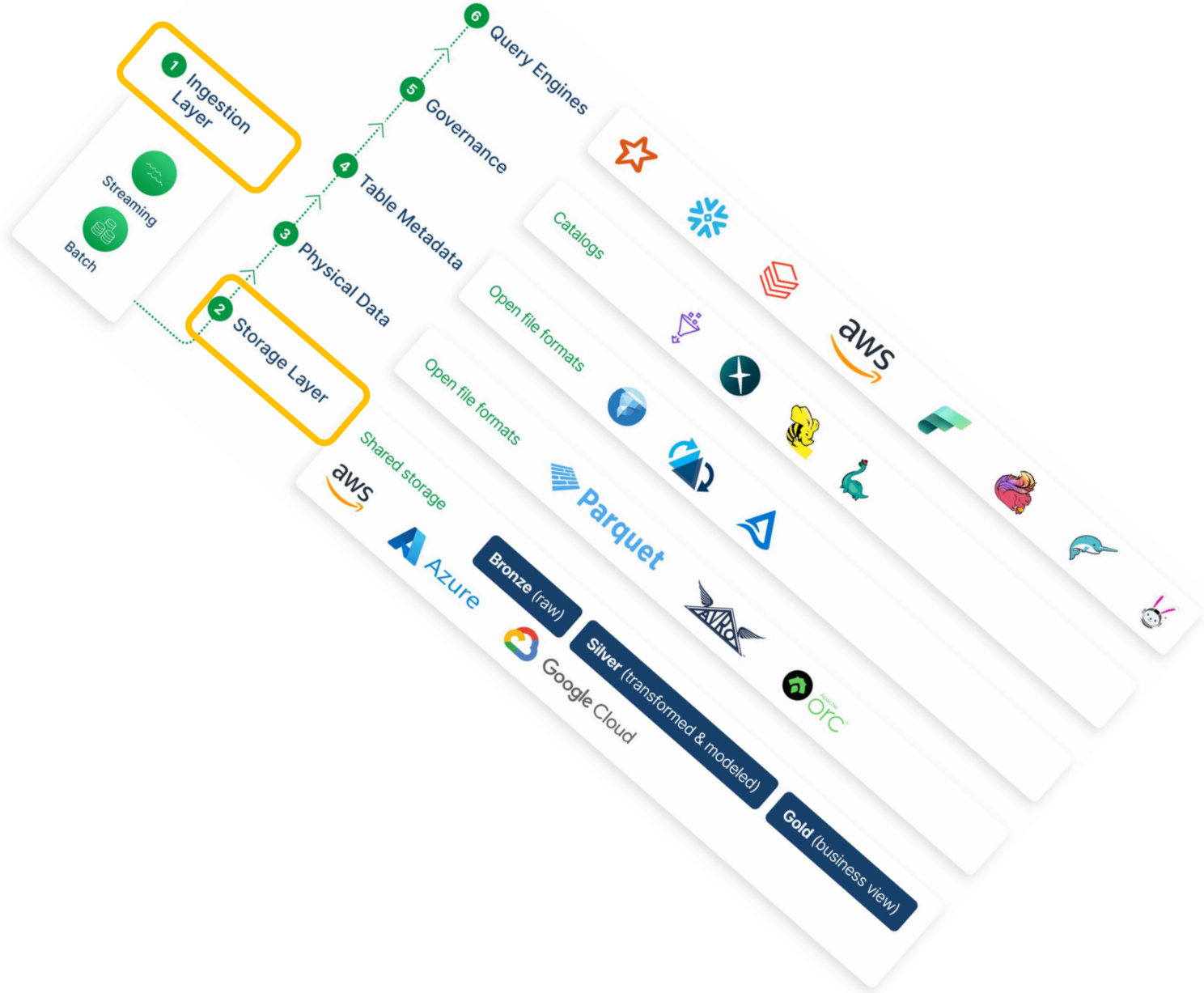


Data Lakehouse Architecture



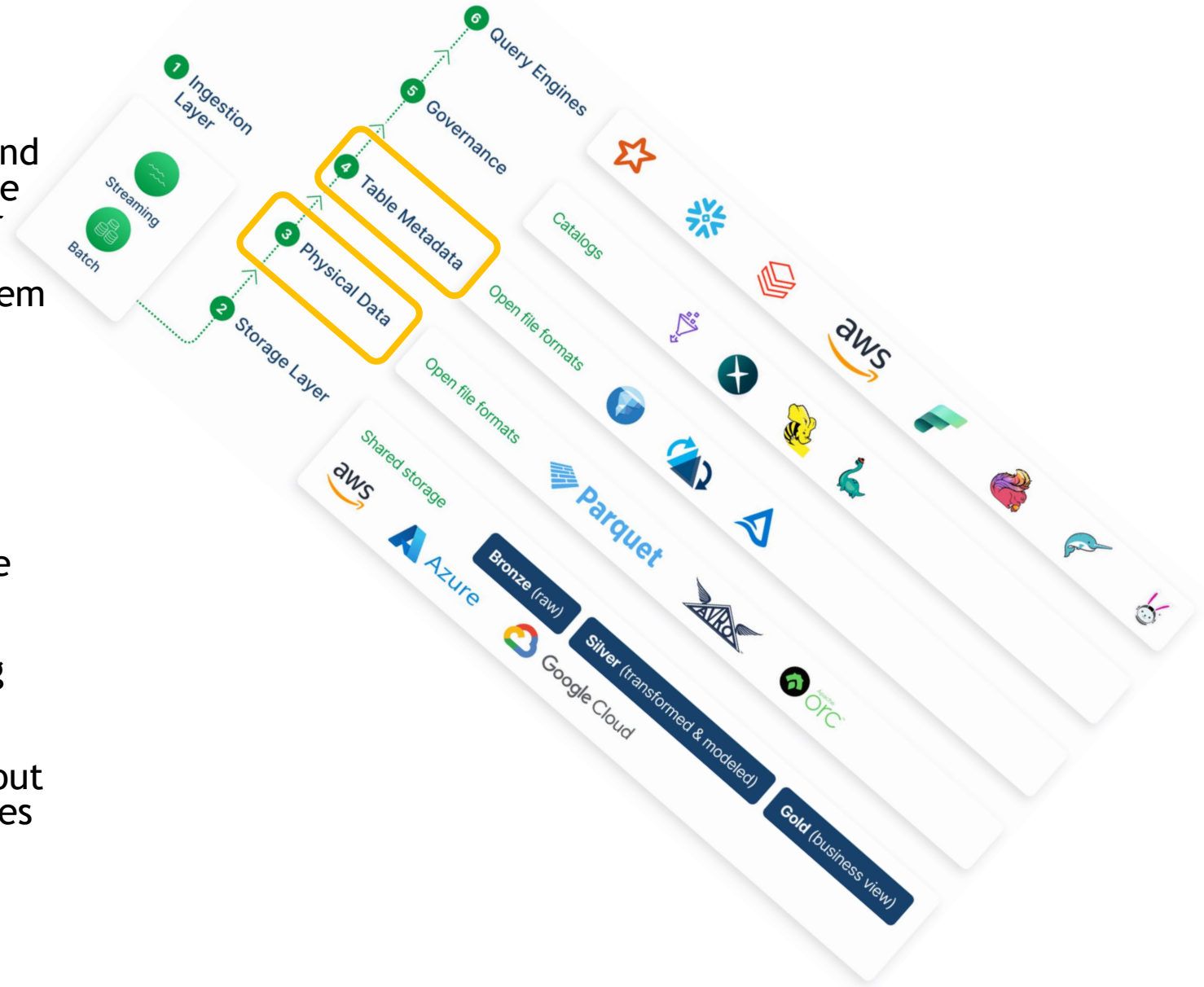
1. **Ingestion Layer:** Offers capabilities to ingest data from various sources into the lakehouse, including batch and real-time data pipelines using change data capture (CDC) or streaming. Should offer capabilities to easily ingest and load high volumes of data in real time to the lakehouse with just a few clicks.

2. Storage Layer: Stores all types of data (structured, semi-structured, unstructured) in a single unified platform, often using cloud-based object stores like AWS S3, Azure Blob Storage, or Google Cloud Storage. Data can be stored in a raw, transformed or clean business ready buckets with necessary transformation and cleansing.



3. Physical Data Layer: Open file formats define how a lakehouse writes and reads data. Open file formats focus on efficient storage and compression of data and significantly impacts speed and performance They define how the raw bytes representin records and columns are organized and encoded on disk or in a distributed file system such as Amazon S3. Some of the more common open file formats for lakehouses include **Apache Parquet**, Apache Avro and ORC.

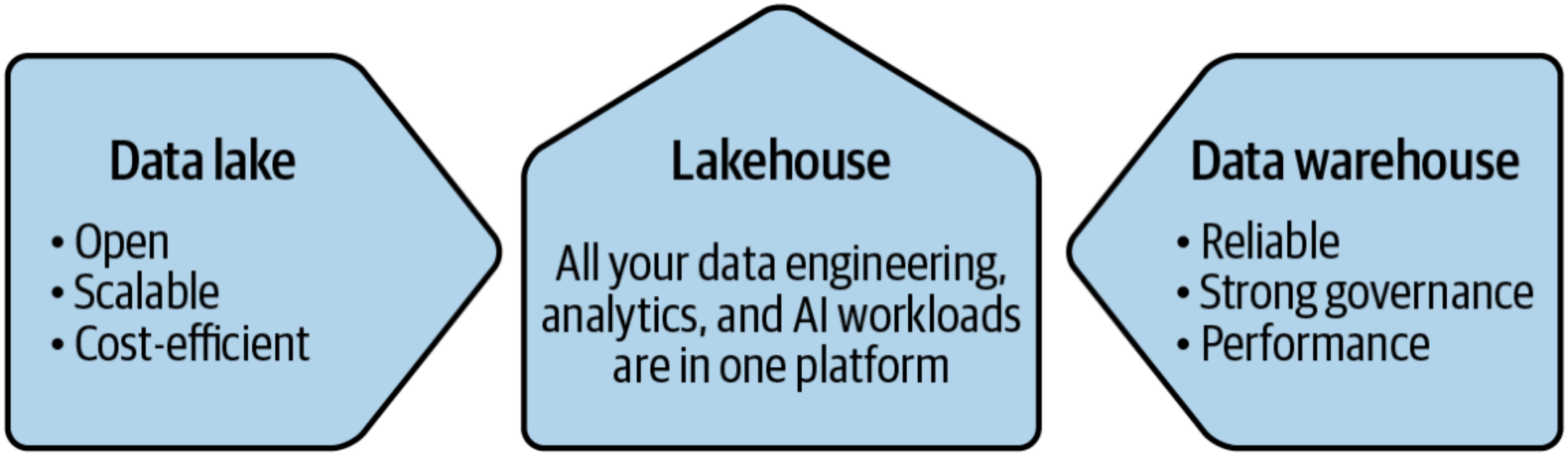
4. Table Formats/Metadata Layer: The differentiating factor between a Data Lake and a Lakehouse is a table format or a table metadata layer. It provides an abstraction layer on top of the physical data layer to facilitate organizing, querying and updating data. Common open table formats include **Apache Iceberg**, Apache Hudi and Delta Tables, that store the information about which objects are part of a table and enables SQL engines to see a collection of files as a table with rows and columns that can be queried and updated transactionally.



5. **Catalog layer:** A catalog refers to a central registry within the lakehouse framework that tracks and manages the metadata of the tables underneath It essentially acts as a source of truth for where t find the current state of a table, including its schema, partitions, and data locations, allowing different compute engines to access and manipulate lakehouse tables consistently. Examples include AWS Glue catalog, Snowflake open catalog, Polaris, Unity Catalog, Hive Catalog, Project Nessie, and REST catalogs.

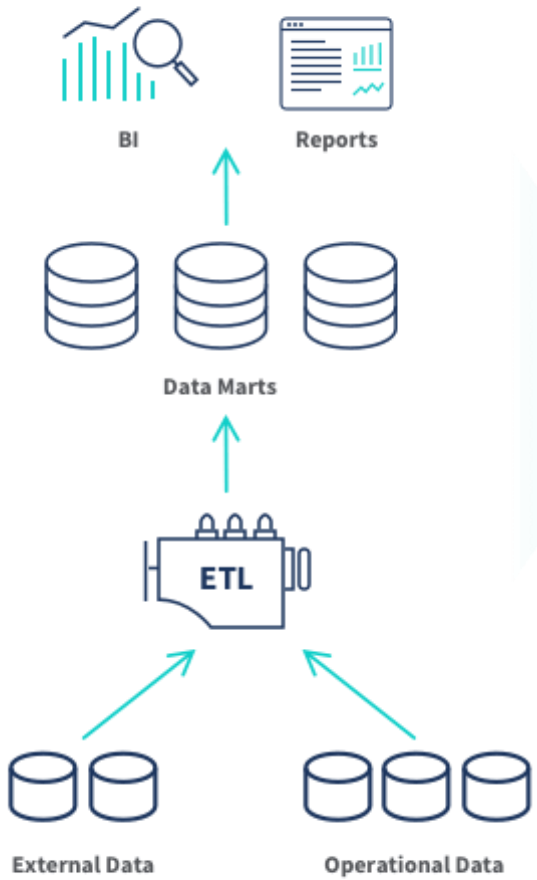
6. Query/ Compute layer: Provides processing power to analyze and query data stored in the storage layer. It may also utilize distributed processing engines like Apache Spark, Presto, or Hive or other Cloud data engines to handle large datasets efficiently. This layer enables users to access and analyze data from the lakehouse using diverse tools and applications like query engines, BI dashboards, data science platforms, and SQL clients.

Convergence of data lakes and data warehouses
Into
a unified data
Lakehouse Platform



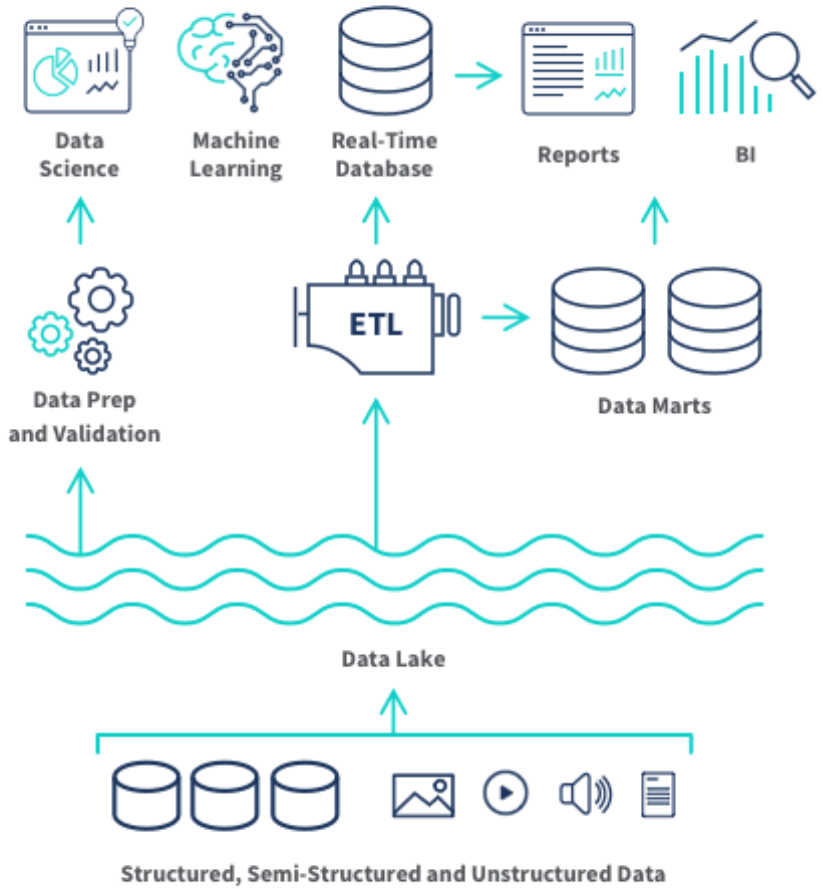
LATE 1980'S

Data Warehouse



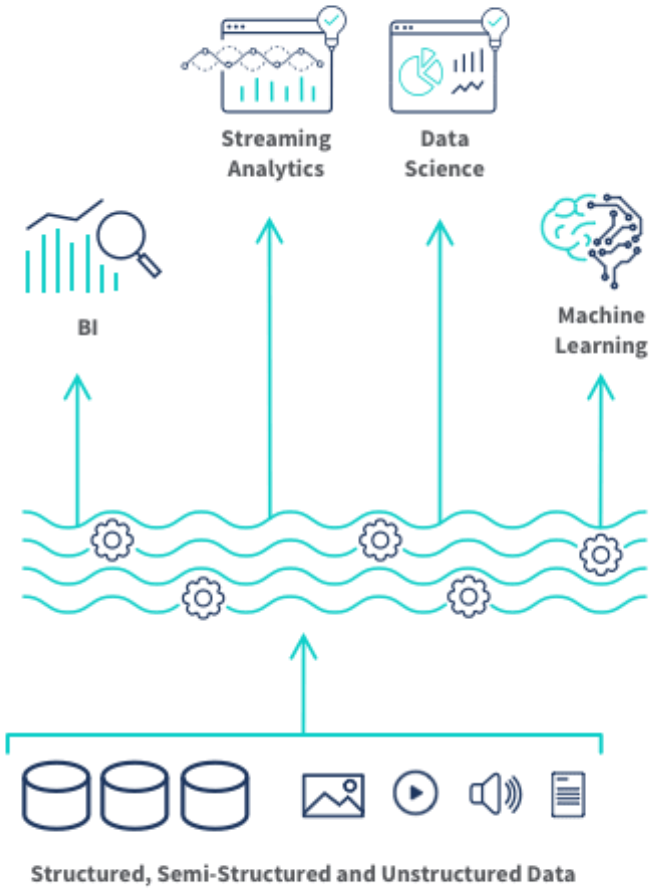
2011

Data Lake



2020

Lakehouse



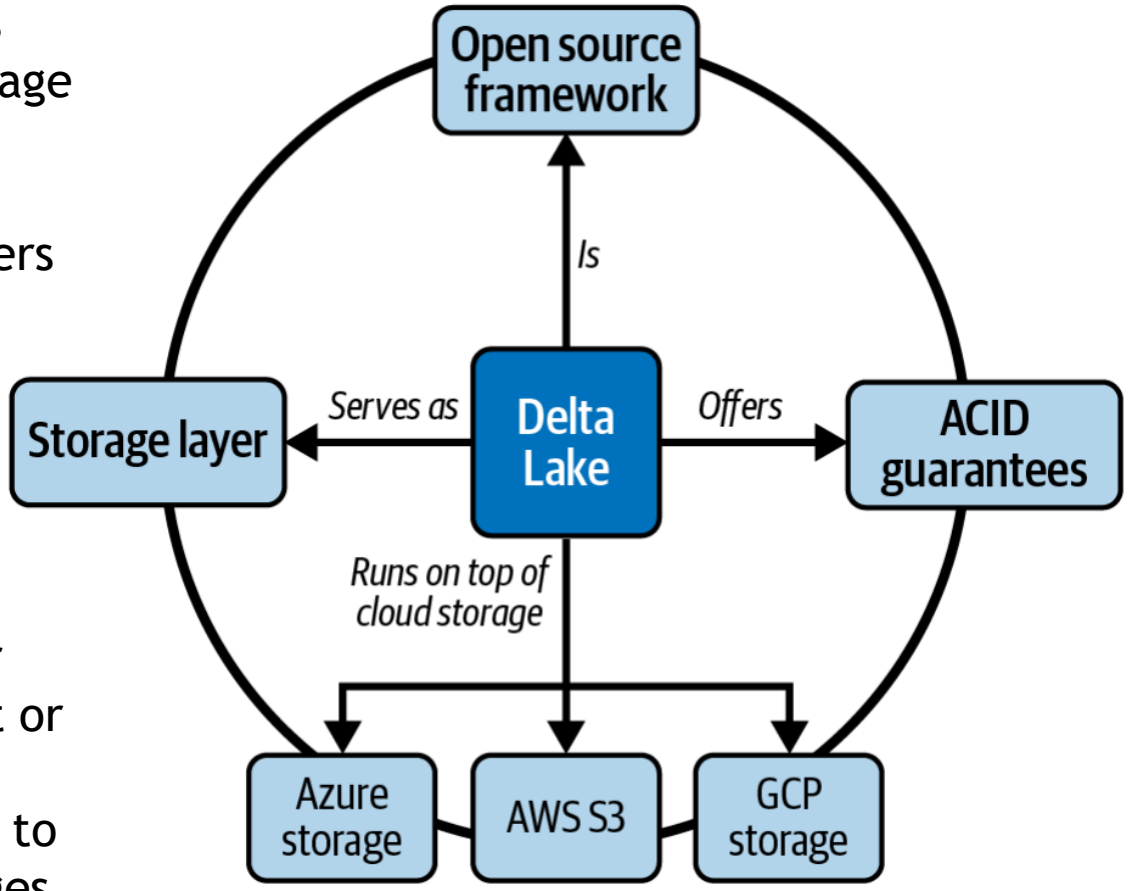
Attributes	Data Warehouse	Data Lake	Data Lakehouse
Overview	Data warehouses ingest and hold highly structured and unified data to support specific business intelligence and analytics needs. The data has been transformed and fit into a defined schema.	Data lakes ingest and hold raw data in a wide variety of formats to directly support data science, AI and machine learning. Massive volumes of structured and unstructured data like ERP transactions and call logs can be stored cost effectively. Data teams can build data pipelines and schema-on-read transformations to make data stored in a data lake available for BI and analytics tools.	Data Lakehouse combines the best of data warehouse and data lakes and can eliminate data redundancies, improving data quality while offering lower costs. Utilizing open table formats enable data to be stored cost-efficiently in cloud object stores while being able to be queried or processed with multiple engines.
Data Format	Closed proprietary format	Open format	Open format
Type of Data	Structured data, with limited support for semi-structured	All types: structured, semi-structured data, textual data, unstructured (raw)data	All types: Structured data, semi-structured data, textual data, unstructured (raw) data
Data Access	SQL only; no direct access to files	Open APIs for direct access with SQL, R, Python and other languages	SQL, along with API extensions to access tables and data
Reliability	High quality - reliable data with ACID transactions	Low quality - becomes a data swamp without data catalogs and the right governance	High quality - reliable data with ACID transactions
Governance and Security	Fine-grained security and governance at the row/column level for tables	Fine-grained security and governance at the row/column level for tables	Fine-grained security and governance at the row/column level for tables
Scalability	Scaling becomes exponentially more expensive	Scales to hold any amount of data at low cost, regardless of type	Scales to hold any amount of data at low cost, regardless of type
Streaming	Partial; limited scale	Yes	Yes
Query Engine Lock-In	Yes	No	No
Resources	Learn more about data warehouses Learn More about cloud data warehouses	Learn more about data lakes Take a deeper look at data lake vs data warehouse	Learn more about lakehouses Guide to Iceberg lakehouses

Delta Lake

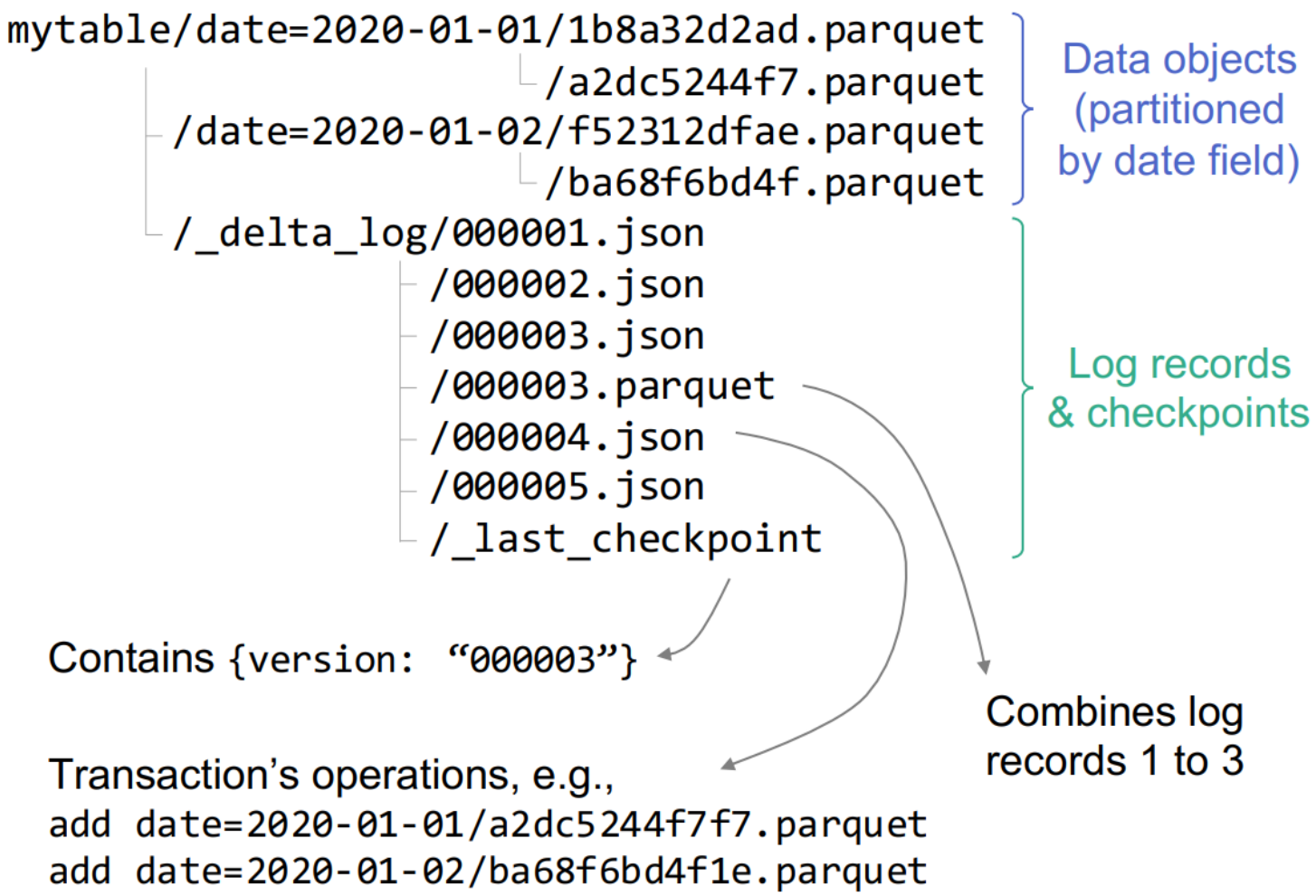
Delta Lake is an open-source storage layer that brings reliability to data lakes by adding a transactional storage layer on top of data stored in cloud storage.

In the context of data lakehouses, a storage layer refers to the framework responsible for managing and organizing data
stored within the data lake. It serves as an Intermediary platform through which data is ingested, queried, and processed.

In other words, Delta Lake is not a storage medium or storage format. Common storage formats like Parquet or JSON define how data is physically stored in the lake. However, Delta Lake runs on top of such data formats to provide a robust solution that overcomes the challenges of data lakes.



Delta Lake Table



Iceberg

Apache Iceberg is an open table format designed to manage large-scale data lakehouses and enable high-performance analytics on open data formats. It allows files to be treated as logical table entities, making it well-suited for lakehouse architectures.

With Iceberg, users can store data in cloud object stores and process/query it utilizing multiple different engines, offering flexibility and interoperability across platforms.

Iceberg supports some key features such as ACID compliance, dynamic partitioning, time travel, and schema evolution, ensuring high performance and data integrity.

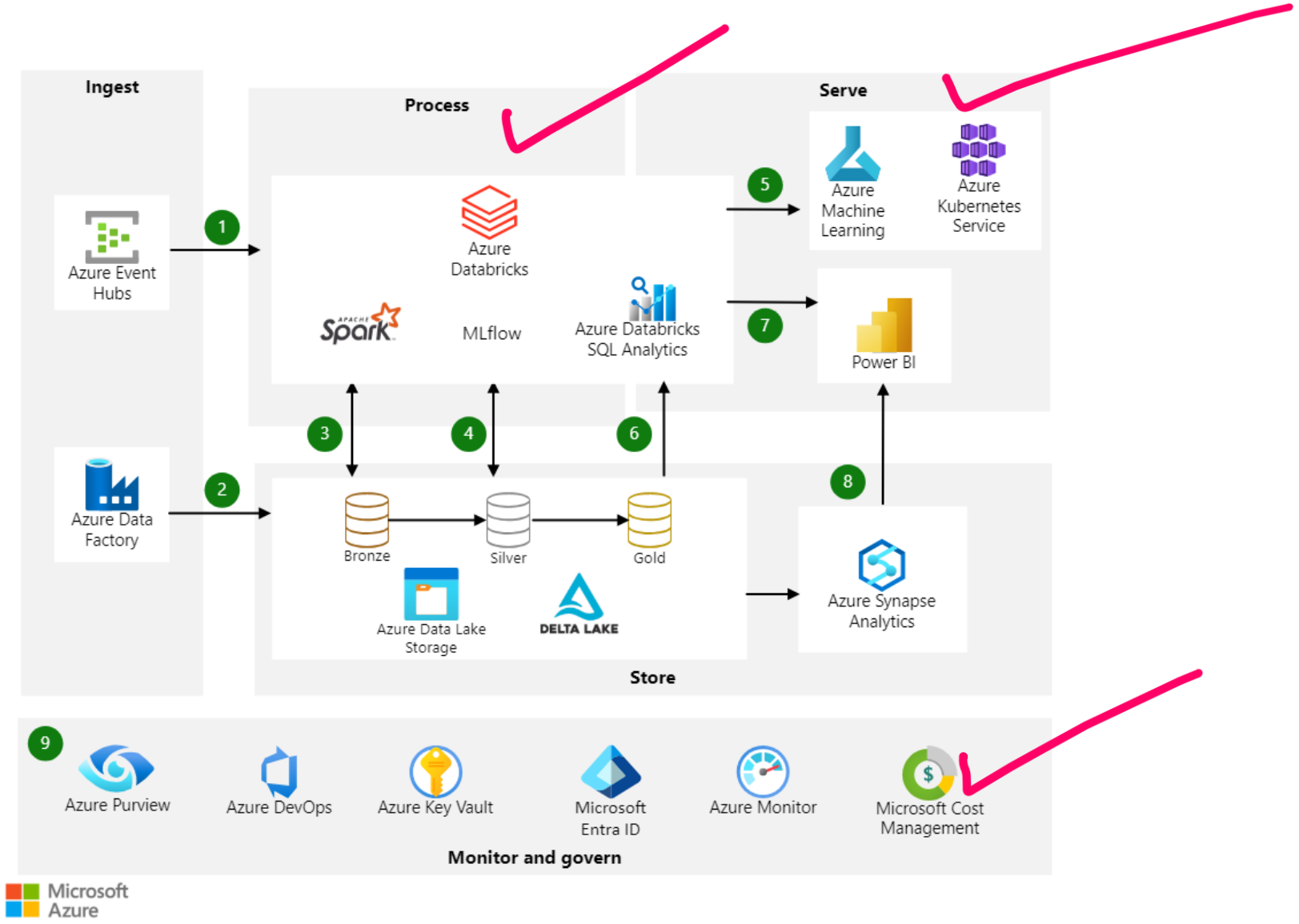
Additionally, Apache Iceberg fosters a strong open-source community, making it a reliable, versatile and open solution for modern data management needs.

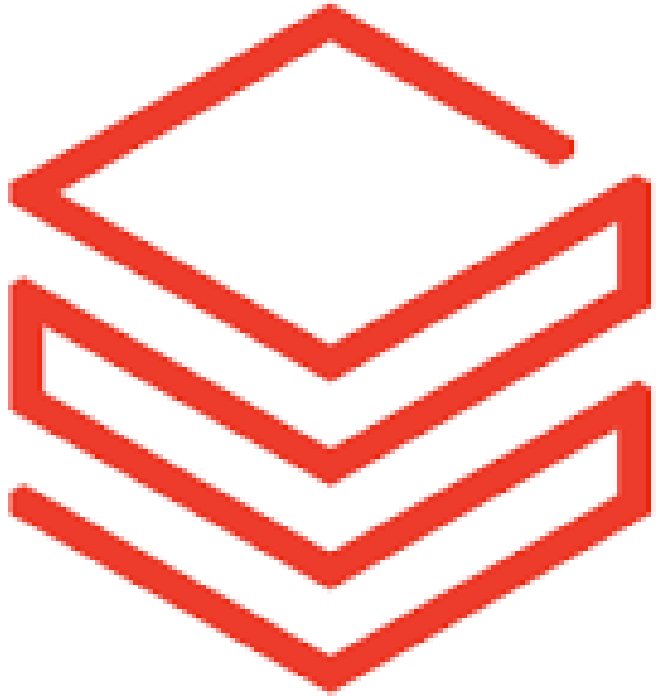
Learn more about Apache Iceberg here

Apache Hudi

Apache Hudi is an open data lakehouse platform, built on a high-performance open table format to bring database functionality to your data lakes.
Hudi reimagines slow old-school batch data processing with a powerful new incremental processing framework for low latency minute-level analytics.

https://hudi.apache.org/





Azure Databricks

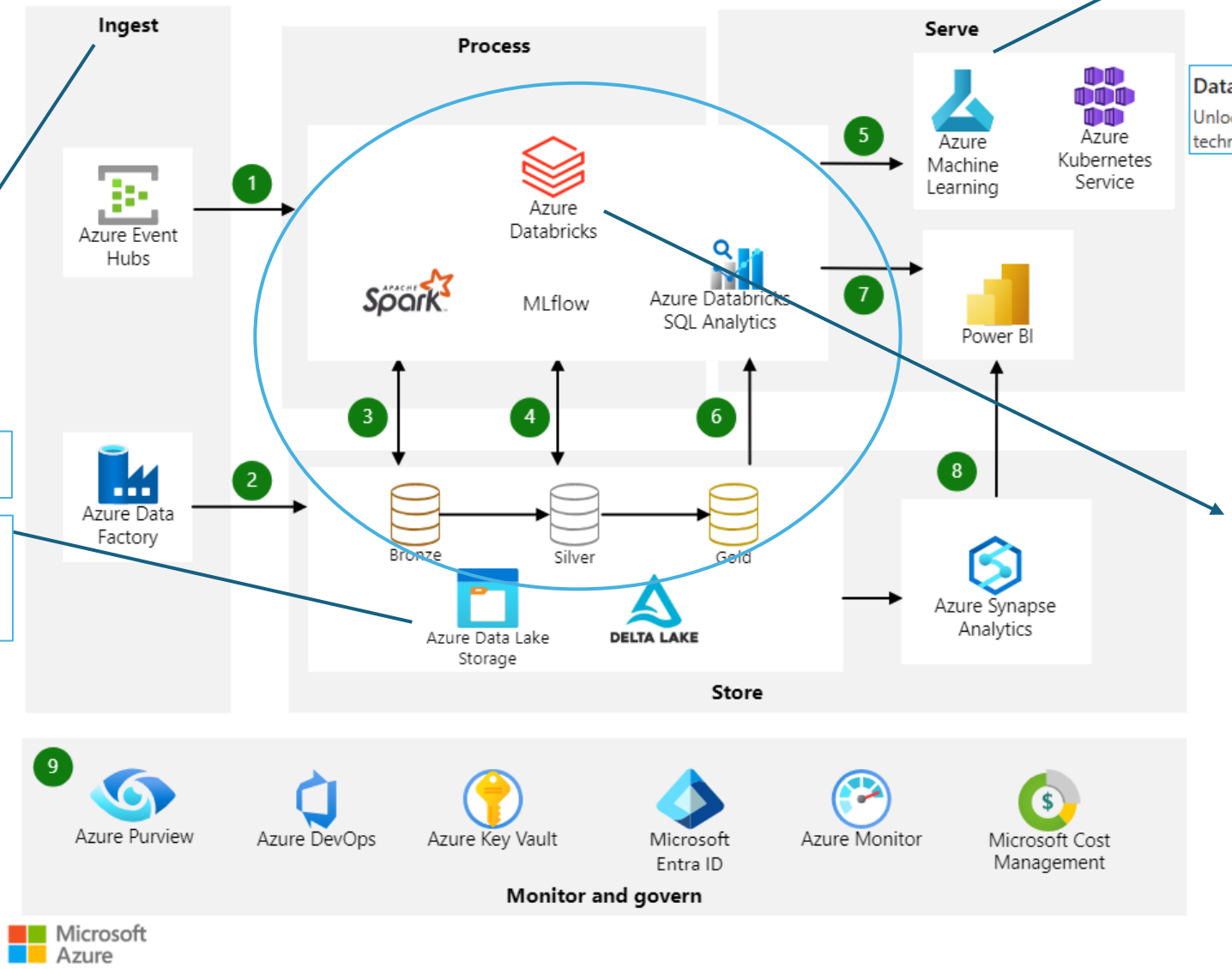
(From a bird’s eye view)

- Azure Databricks is a cloud-based data analytics platform that provides **a unified environment for**
  - **Data engineering,**
  - **Machine learning,**
  - **Analytics**
- Fully-managed data lake analytics platform
- Based on open-source technologies
- Integrated with Azure for resource management and security

Modern analytics architecture with Azure Databricks

Machine Learning Area / Data Scientist

Data Science
Unlock powerful insights using AI and machine learning technology.



Data engineering area

Data Engineering
Create a lakehouse and operationalize your workflow to build, transform, and share your data estate.

Implementing a Data Analytics Solution with Azure Databricks

Modern analytics architecture with Azure Databricks

Event Hubs is a big data streaming platform. (PaaS), this event ingestion service is fully managed.

Data Factory is a hybrid data integration service. You can use this fully managed, serverless solution to create, schedule, and orchestrate data transformation workflows.

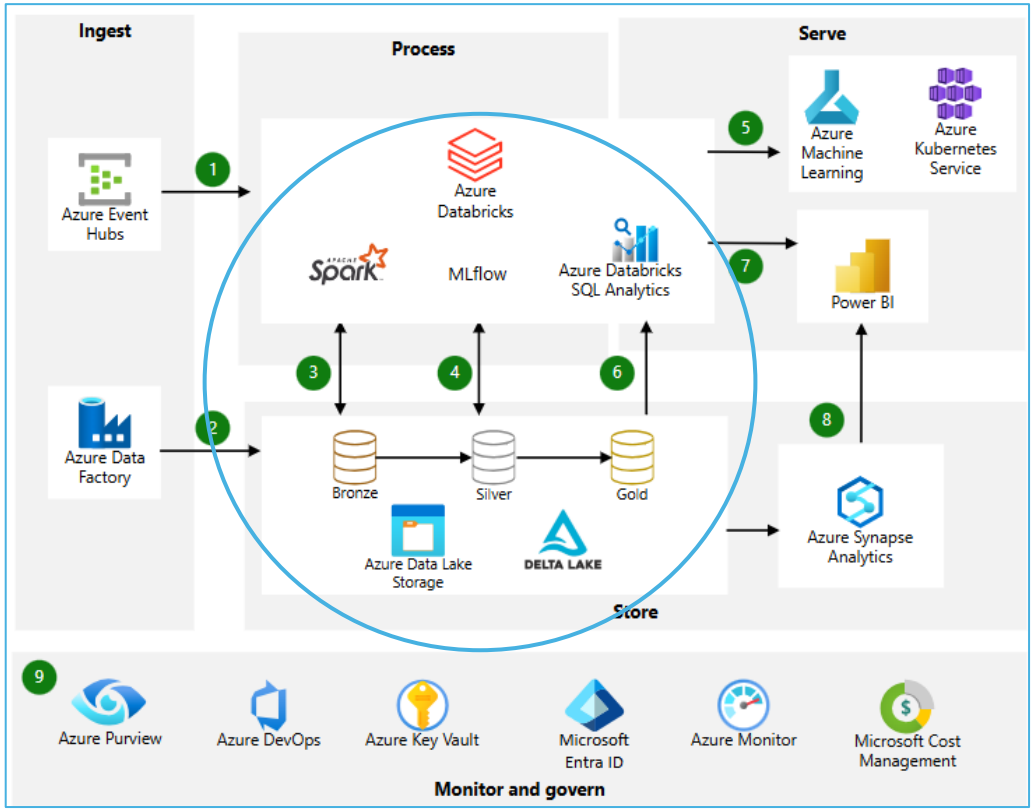
Azure Databricks is a data analytics platform. Its fully **managed Spark clusters** process large streams of data from multiple sources. Azure Databricks cleans and transforms structureless data sets. It combines the processed data with structured data from operational databases or data warehouses. Azure Databricks also trains and deploys scalable machine learning and deep learning models.

MLflow is an open-source platform for the machine learning lifecycle. Its components monitor machine learning models during training and running. MLflow also stores models and loads them in production.

Azure Databricks SQL Analytics runs queries on data lakes. This service also visualizes data in dashboards.

Machine Learning is a cloud-based environment that helps you build, deploy, and manage predictive analytics solutions. With these models, you can forecast behavior, outcomes, and trends.

AKS is a highly available, secure, and fully managed Kubernetes service. AKS makes it easy to deploy and manage containerized applications.



Data Lake Storage Gen2 **is a scalable and secure data lake for high-performance analytics workloads. This service can manage multiple petabytes of information while sustaining hundreds of gigabits of throughput. The data may be structured, semi-structured, or unstructured. It typically comes from multiple, heterogeneous sources like logs, files, and media.**

Delta Lake is a storage layer that uses an open file format. This layer runs on top of cloud storage such as **Data Lake Storage Gen2**. Delta Lake supports data versioning, rollback, and transactions for updating, deleting, and merging data.

Azure Synapse is an analytics service for data warehouses and big data systems. This service integrates with Power BI, Machine Learning, and other Azure services.
Azure Synapse connectors efficiently transfer large volumes of data between Azure Databricks clusters and Azure Synapse instances.
SQL pools provide a data warehousing and compute environment in Azure Synapse. The pools are compatible with Azure Storage and Data Lake Storage Gen2.

Reporting and governing components

•**Power BI** is a collection of software services and apps. These services create and share reports that connect and visualize unrelated sources of data. Together with Azure Databricks, Power BI can provide root cause determination and raw data analysis.

•**Microsoft Purview** manages on-premises, multicloud, and software as a service (SaaS) data. This governance service maintains data landscape maps. Features include automated data discovery, sensitive data classification, and data lineage.

•**Azure DevOps** is a DevOps orchestration platform. This SaaS provides tools and environments for building, deploying, and collaborating on applications.

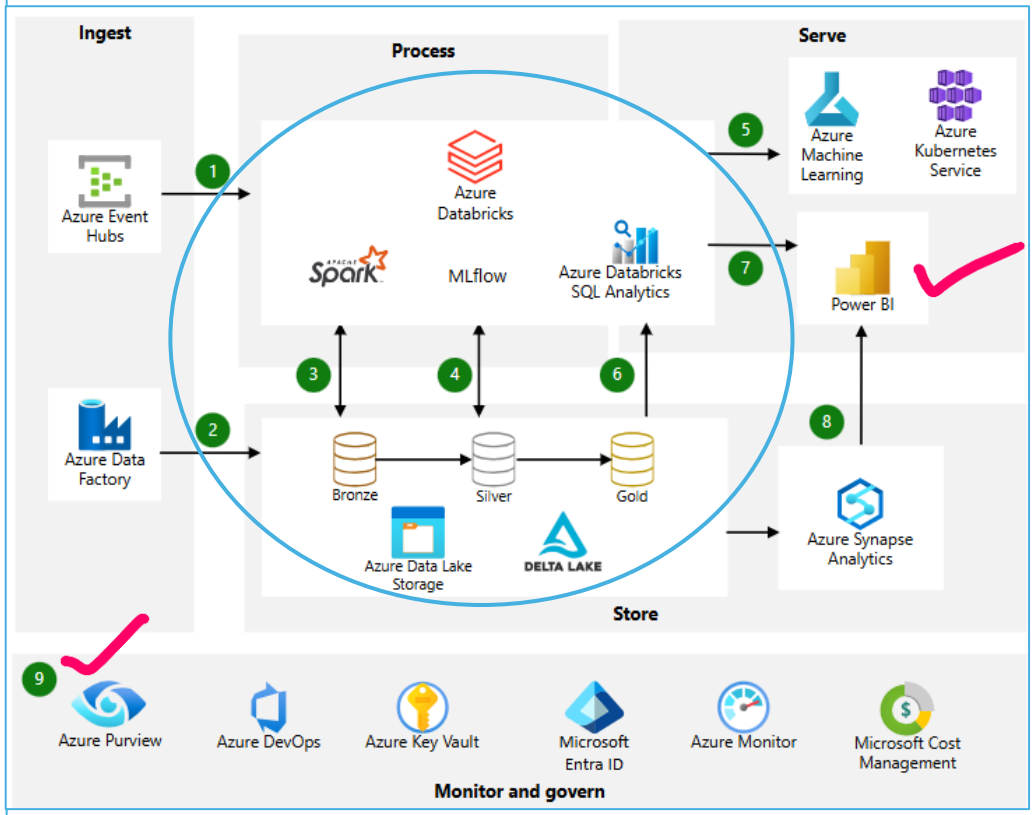
•**Azure Key Vault** stores and controls access to secrets such as tokens, passwords, and API keys. Key Vault also creates and controls encryption keys and manages security certificates.

•**Microsoft Entra ID** offers cloud-based identity and access management services. These features provide a way for users to sign in and access resources.

•**Azure Monitor** collects and analyzes data on environments and Azure resources. This data includes app telemetry, such as performance metrics and activity logs.

•**Microsoft Cost Management** manages cloud spending. By using budgets and recommendations, this service organizes expenses and shows how to reduce costs.

Modern analytics architecture with Azure Databricks



Dataflow of → Modern analytics architecture with Azure Databricks

1-Azure Databricks ingests raw streaming data from **Azure Event Hubs.**

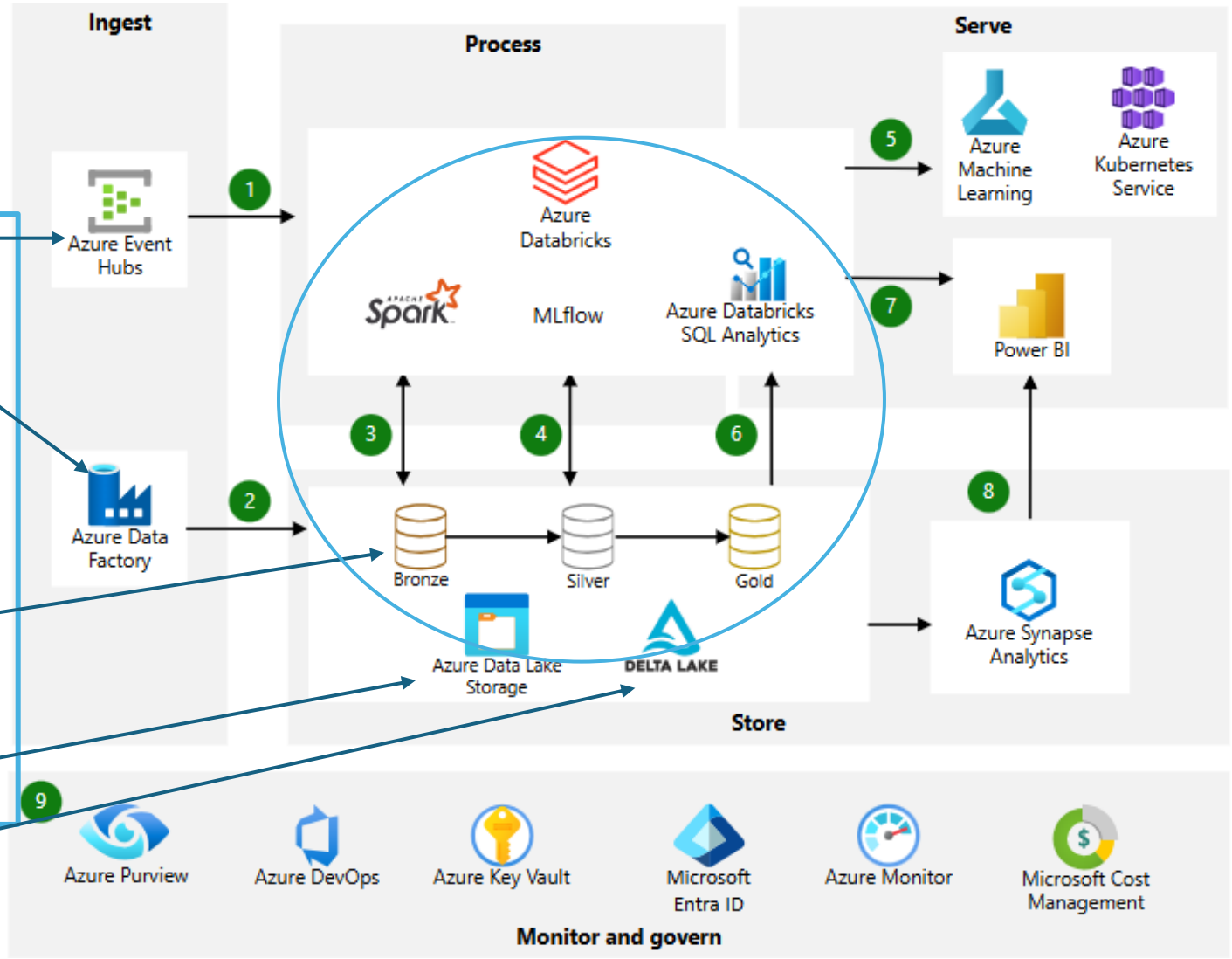
2- Data Factory **loads raw batch data** into Data Lake Storage Gen2.

3. **For data storage:**

a) **Data Lake Storage Gen2** houses data of all types, such as structured, unstructured, and semi-structured. It also stores batch and streaming data.

b) **Delta Lake** forms the **curated layer of the data lake** (processed data that is ready for analysis and consumption.). It stores the **refined data in an open-source format.**

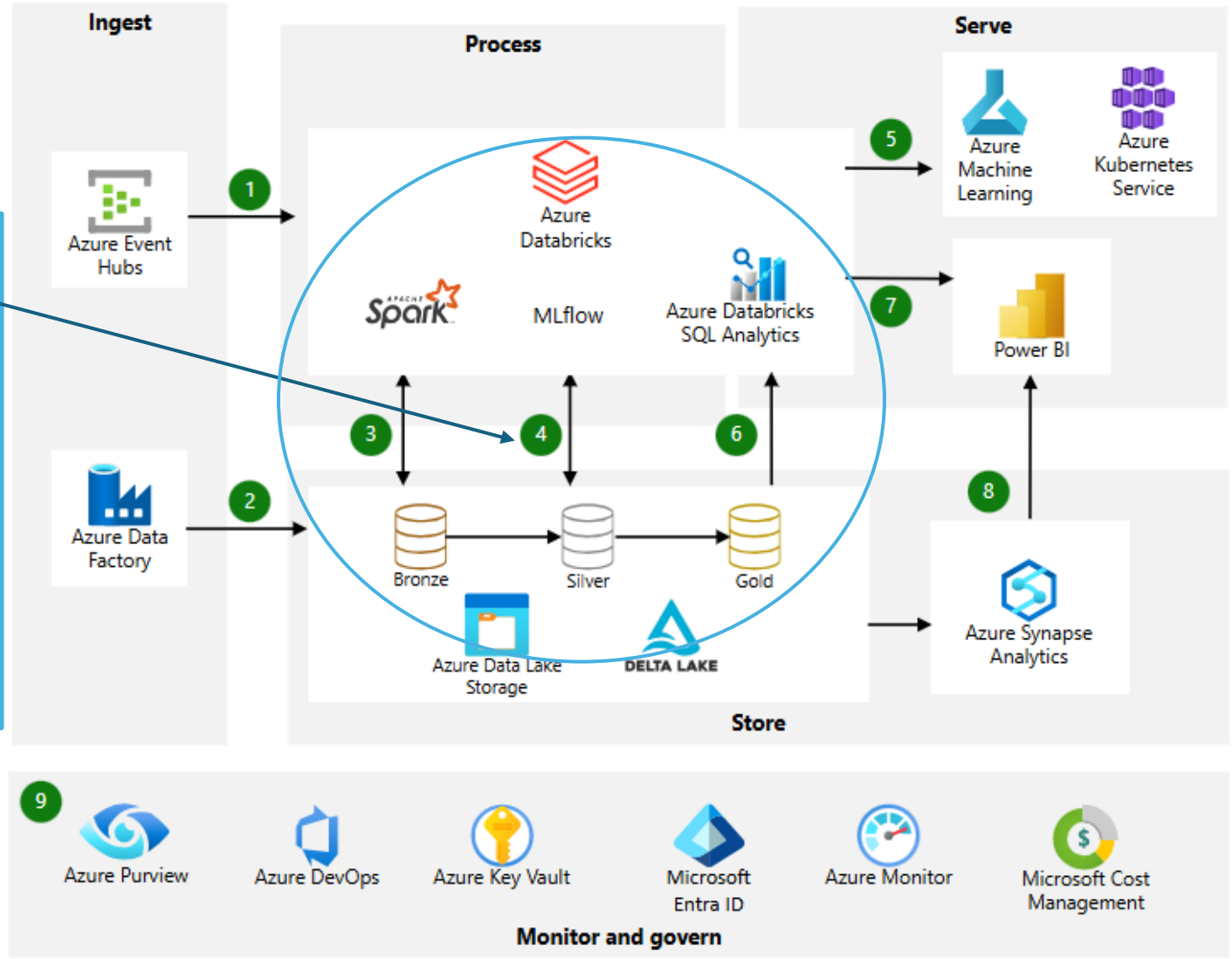
-Azure Databricks works well with a medallion architecture that organizes data into layers: Bronze: Holds raw data. **Silver**: Contains cleaned, filtered data. **Gold**: Stores aggregated data that's useful for business analytics.



Dataflow of → Modern analytics architecture with Azure Databricks

4. The analytical platform ingests data from the disparate batch and streaming sources. Data scientists use this data for these tasks:
- Data Preparation
- Data Exploration
- Model Preparation
- Model training

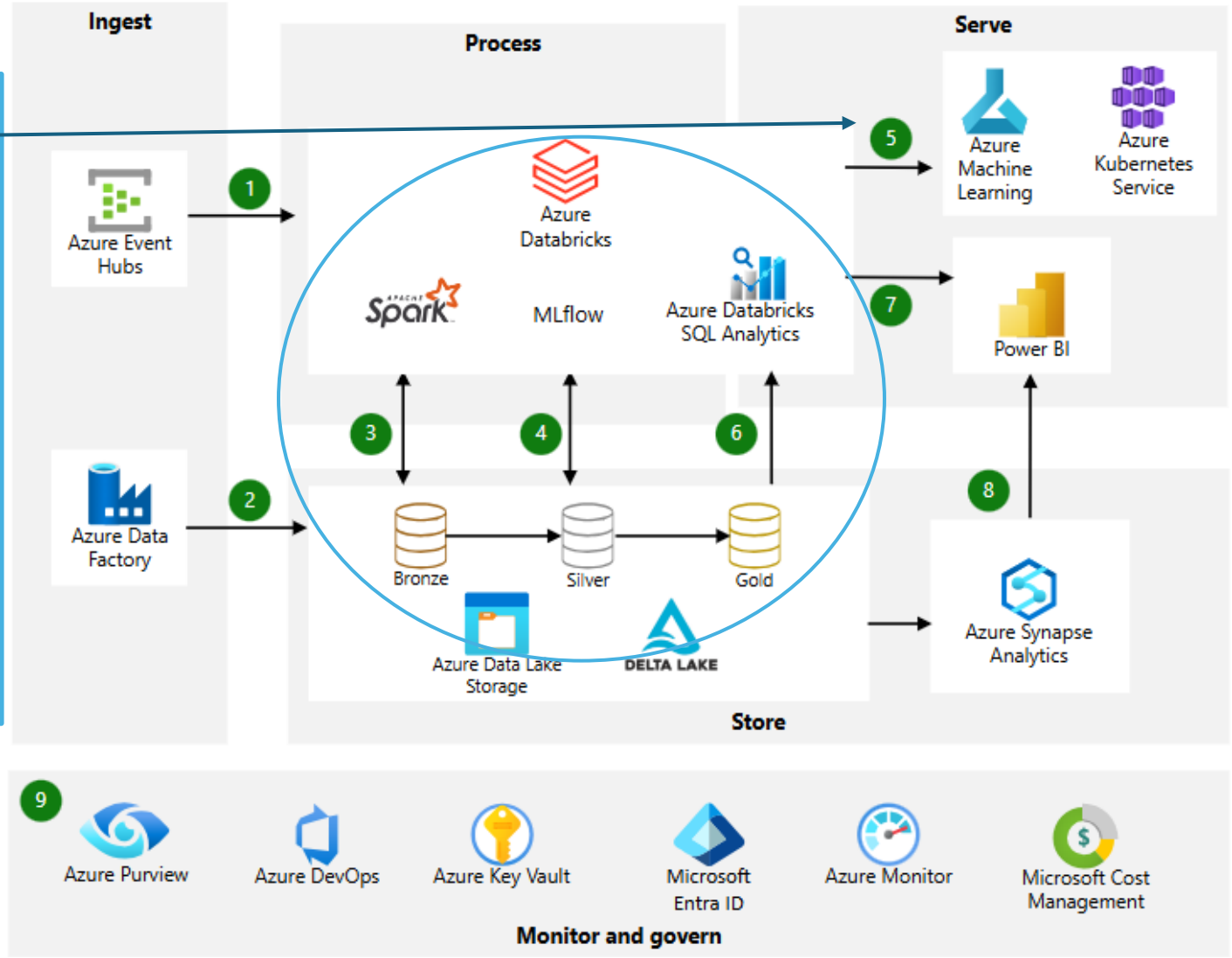
MLflow manages parameter, metric, and model tracking in data science code runs. The coding possibilities are flexible:
- Code can be in SQL, Python, R, and Scala.
- Code can use popular open-source libraries and frameworks such as Koalas, Pandas, and scikit-learn, which are pre-installed and optimize



Dataflow of → Modern analytics architecture with Azure Databricks

5. Machine learning models are available in several formats:

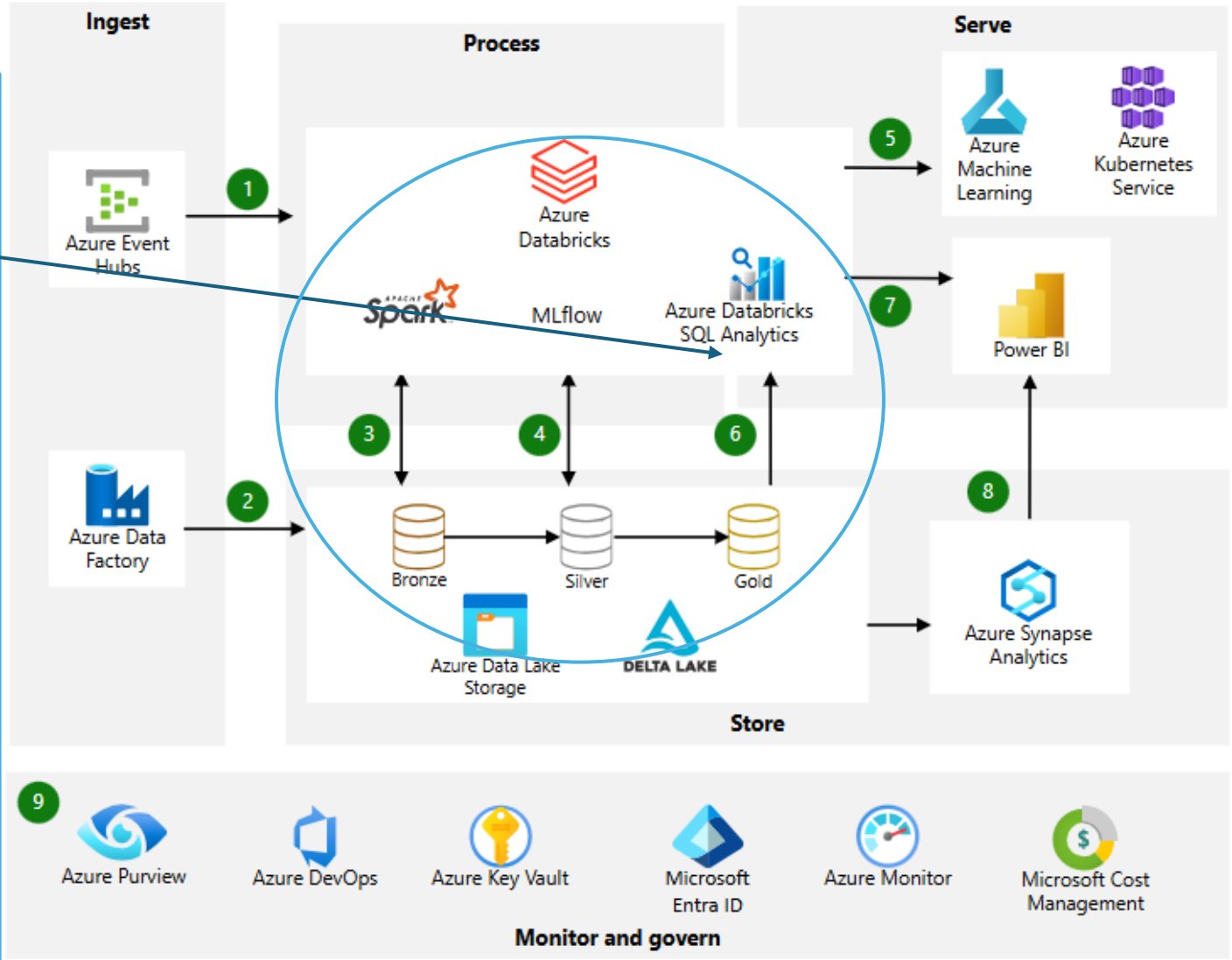
- Azure Databricks stores information about models in the MLflow Model Registry. The registry makes models available through batch, streaming, and REST APIs.
- The solution can also deploy models to Azure Machine Learning web services or Azure Kubernetes Service (AKS).



Dataflow of → Modern analytics architecture with Azure Databricks

6. Services that work with the data connect to a single underlying data source to ensure consistency. For instance, users can run SQL queries on the data lake with **Azure Databricks SQL Analytics**. This service:

- ✓ Provides a query editor and catalog, the query history, basic dashboarding, and alerting.
- ✓ Uses integrated security that includes row-level and column-level permissions.
- ✓ Uses a [Photon-powered Delta Engine to accelerate performance](#).



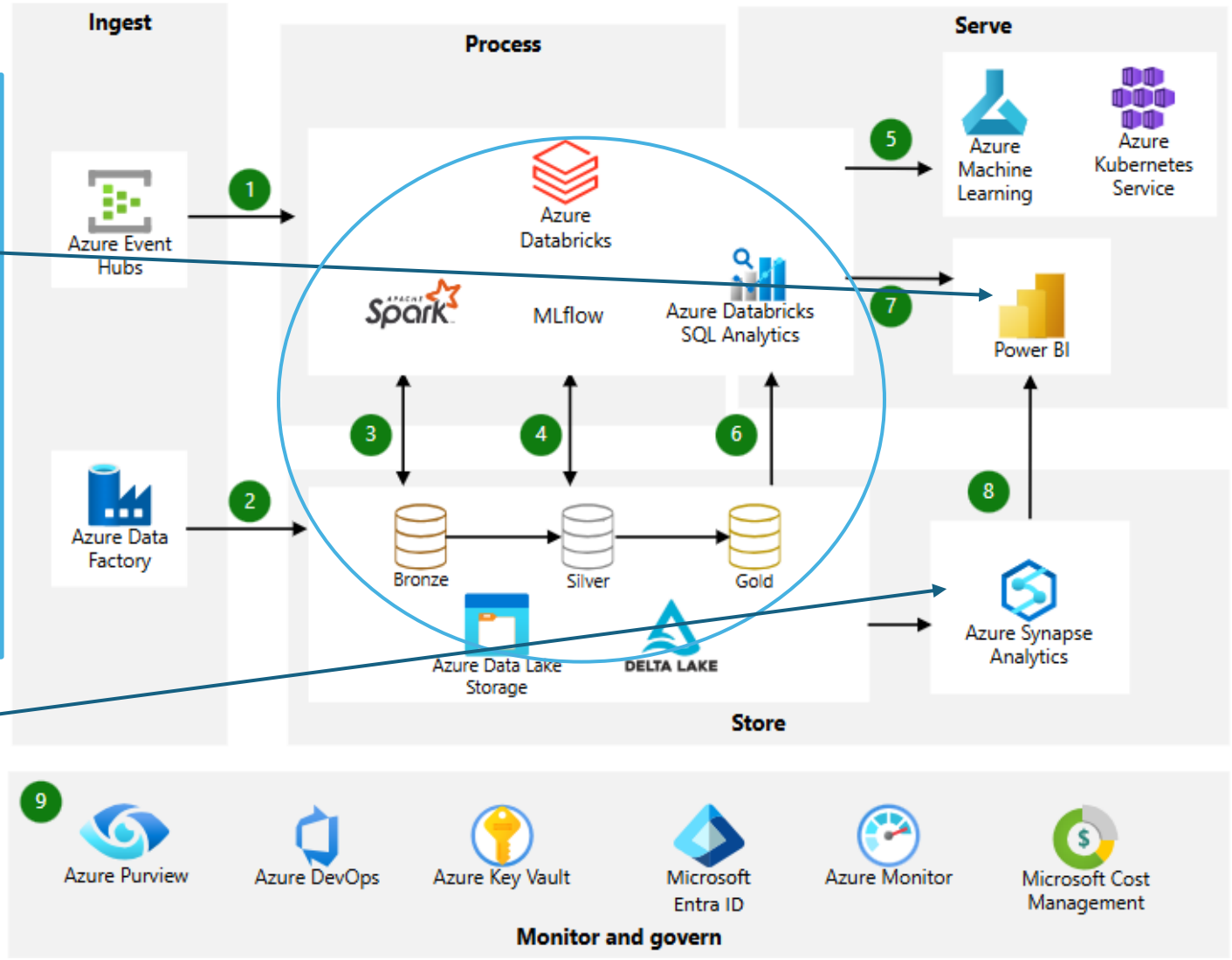
Dataflow of → Modern analytics architecture with Azure Databricks

7. **Power BI** generates analytical and historical reports and dashboards from the unified data platform. This service uses these features when working with Azure Databricks:

   1. A built-in Azure Databricks connector for visualizing the underlying data.

   2. Optimized Java Database Connectivity (JDBC) and Open Database Connectivity (ODBC) drivers

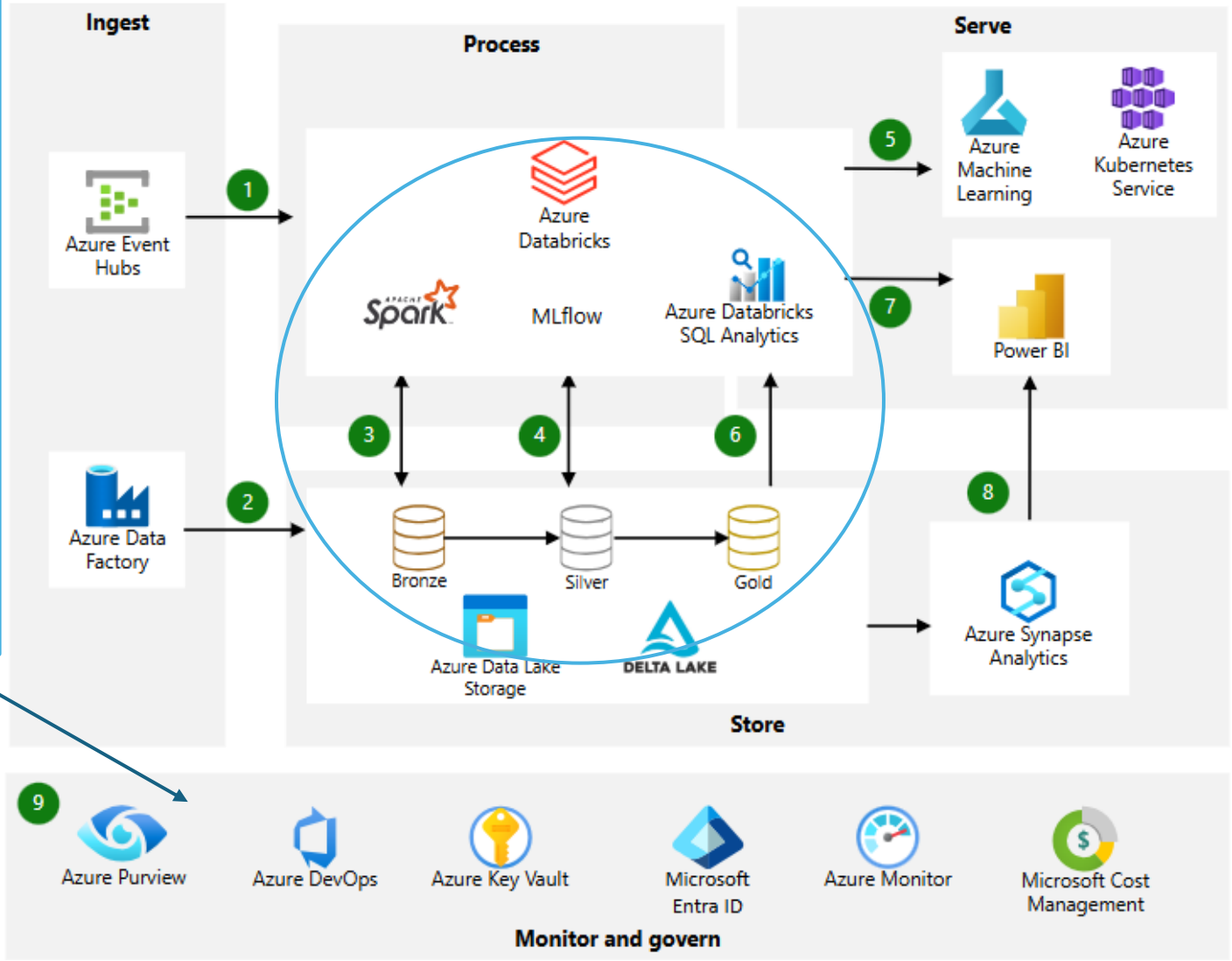
8. Users can export gold data sets out of the data lake into **Azure Synapse** via the optimized Synapse connector. SQL pools in Azure Synapse provide a data warehousing and compute environment.



Dataflow of → Modern analytics architecture with Azure Databricks

- 9.The solution uses Azure services for collaboration, performance, reliability, governance, and security:
- **Microsoft Purview** provides data discovery services, sensitive data classification, and governance insights across the data estate.
- **Azure DevOps** offers continuous integration and continuous deployment (CI/CD) and other integrated version control features.
- **Azure Key Vault** securely manages secrets, keys, and certificates.
- **Microsoft Entra ID** provides single sign-on (SSO) for Azure Databricks users. Azure Databricks supports automated user provisioning with Microsoft Entra ID for these tasks:
- Creating new users.
- Assigning each user an access level.
- Removing users and denying them access.

- **Azure Monitor** collects and analyzes Azure resource telemetry. By proactively identifying problems, this service maximizes performance and reliability.
- **Microsoft Cost Management** provides financial governance services for Azure workloads.



By combining ADLS for storage and Delta Lake for data management, you get a robust and scalable data architecture capable of handling complex data workloads efficiently.

ADLS: For scalable, cost-effective storage of raw data.

Delta Lake: For adding reliability, consistency, and performance optimizations

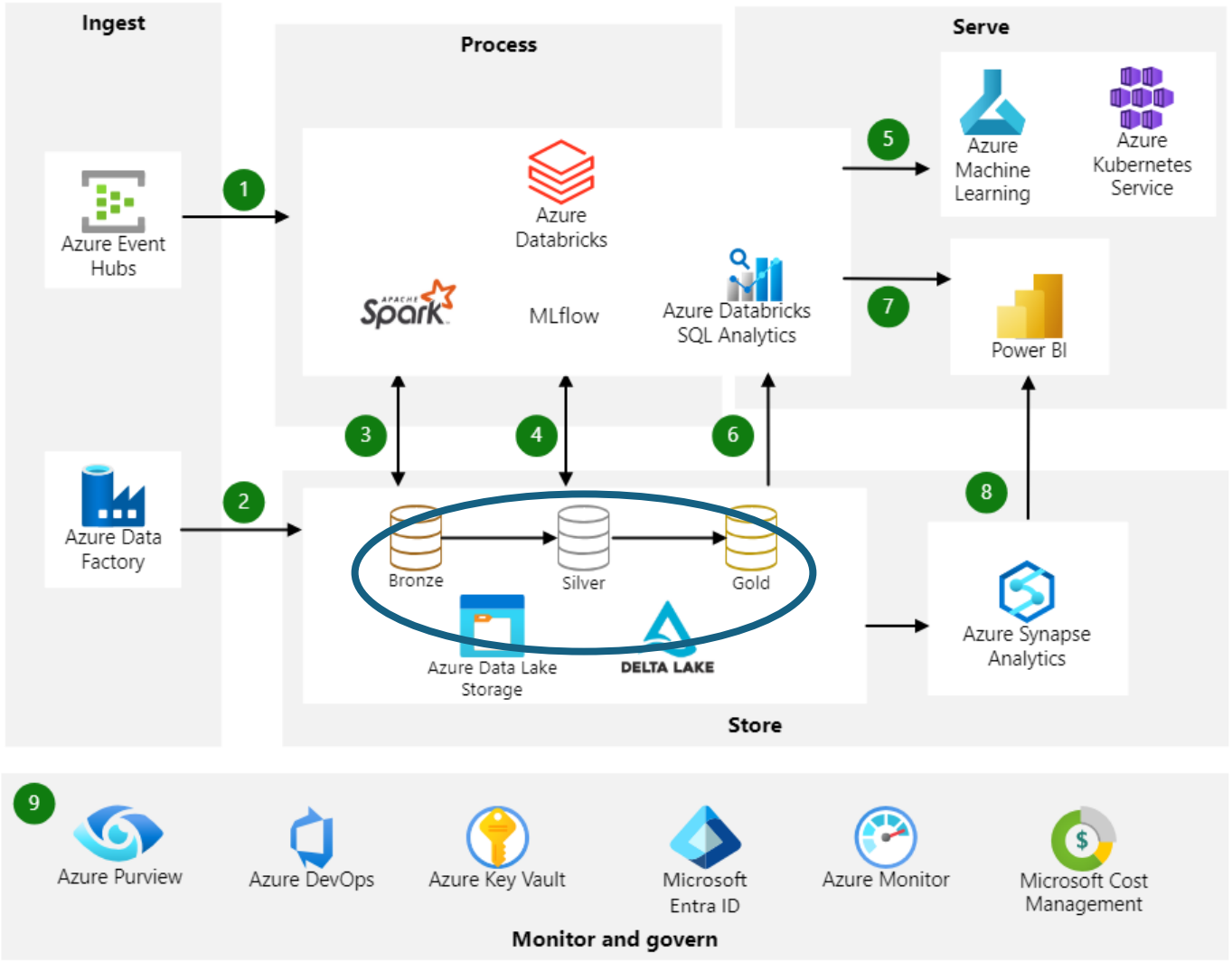
How They Work Together

Data Ingestion: Raw data (Structure and Unstructured) is ingested into **ADLS**, where it can be stored in its native format.

Data Processing: Databricks and other analytics tools can process this **data directly from ADLS.**

Delta Lake Integration: **Delta Lake** can be layered on top of ADLS to manage this data more effectively, providing reliability, consistency, and performance improvements.

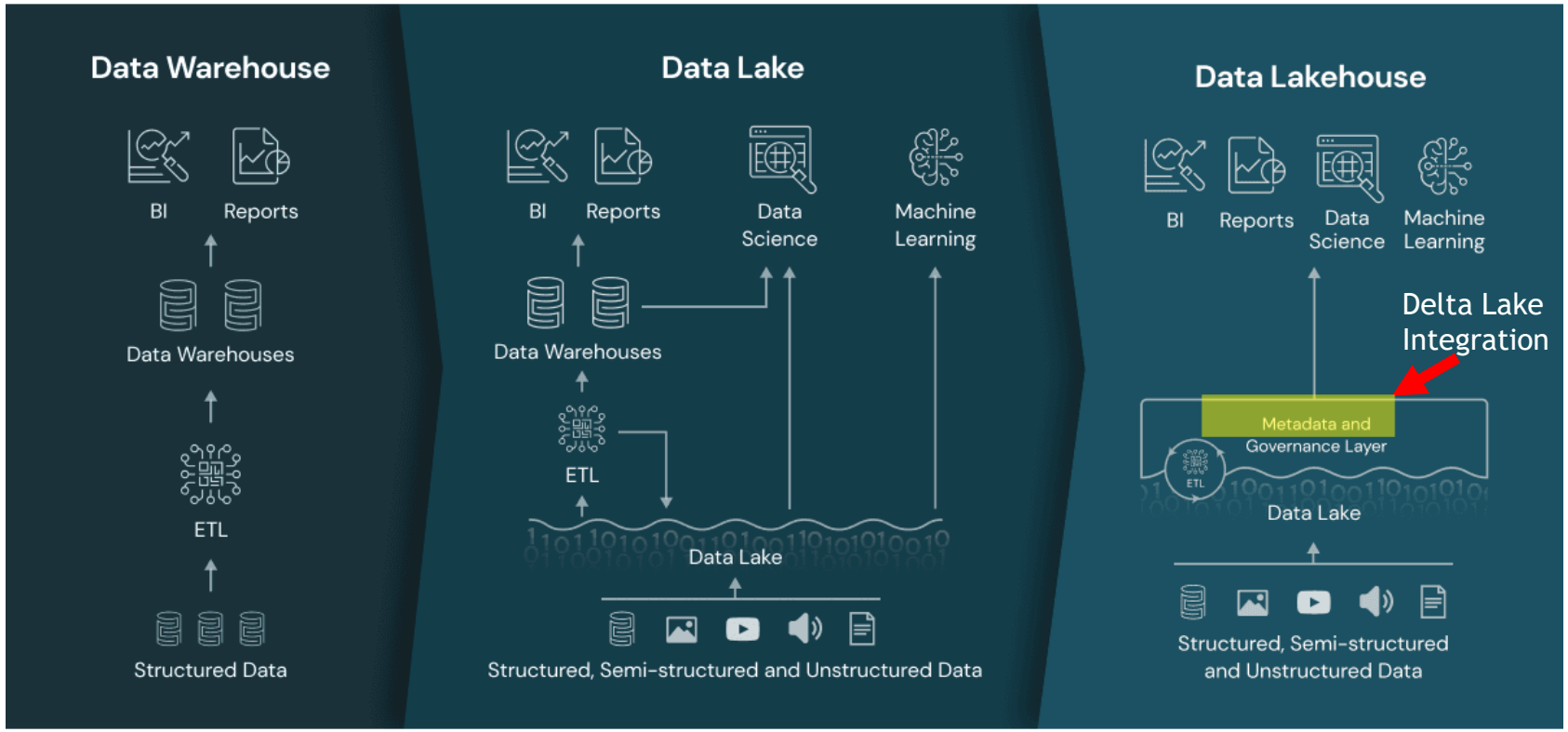
Advanced Analytics: With Delta Lake, you can perform advanced analytics on your data while ensuring data quality and governance.



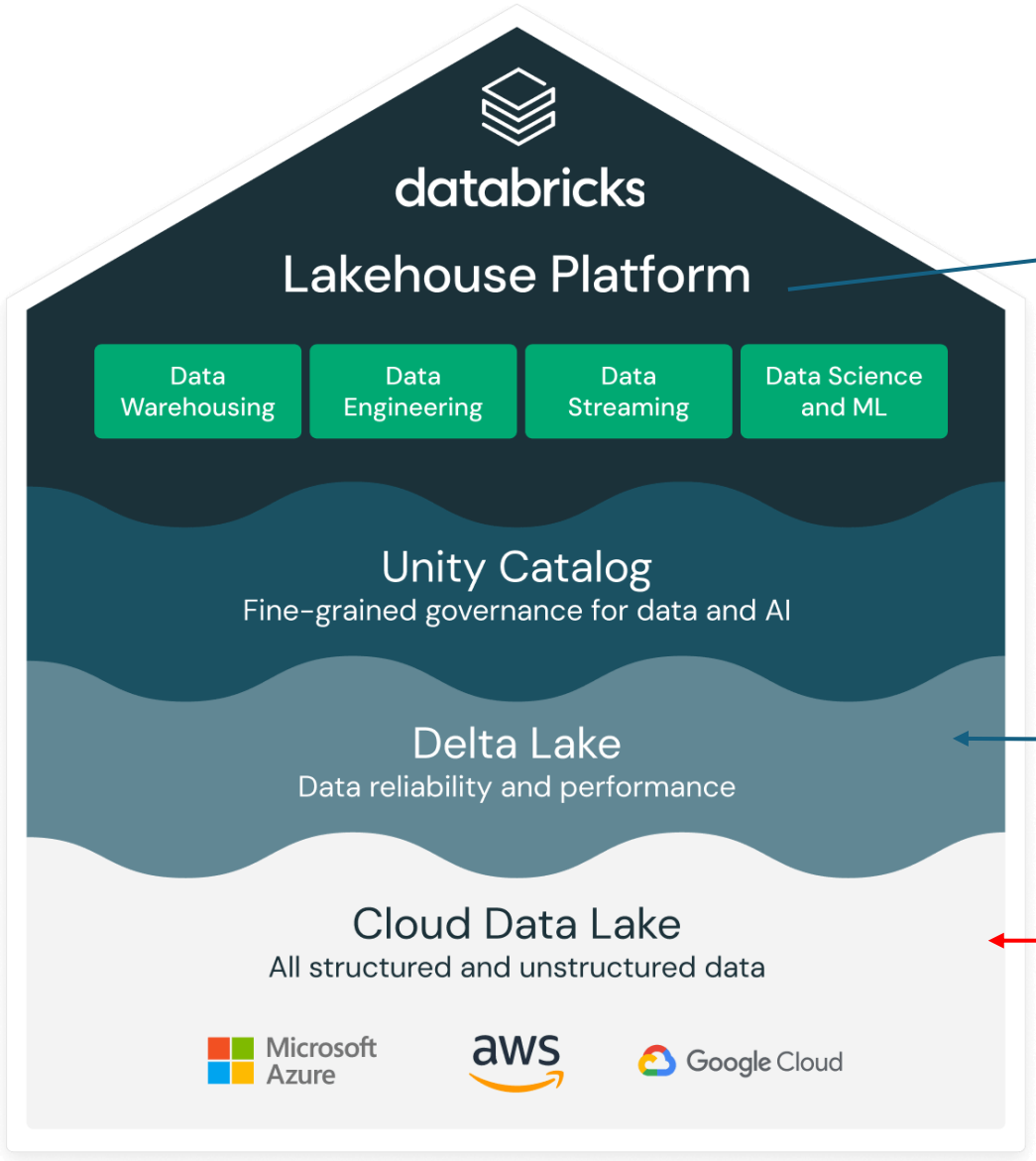
Your data becomes your "single source of truth"

Microsoft Azure

➢ **A data lakehouse can help establish a single source of truth, eliminate redundant costs, and ensure data freshness.**



Data lakehouse leverages **ADLS for storage** while adding robust **data management features through Delta Lake**, creating a powerful and efficient data platform for various analytics and machine learning workloads.



Merging **data lakes** and **data warehouses** into a single system means that data teams can move faster as they are able use data without needing to access multiple systems.

Microsoft

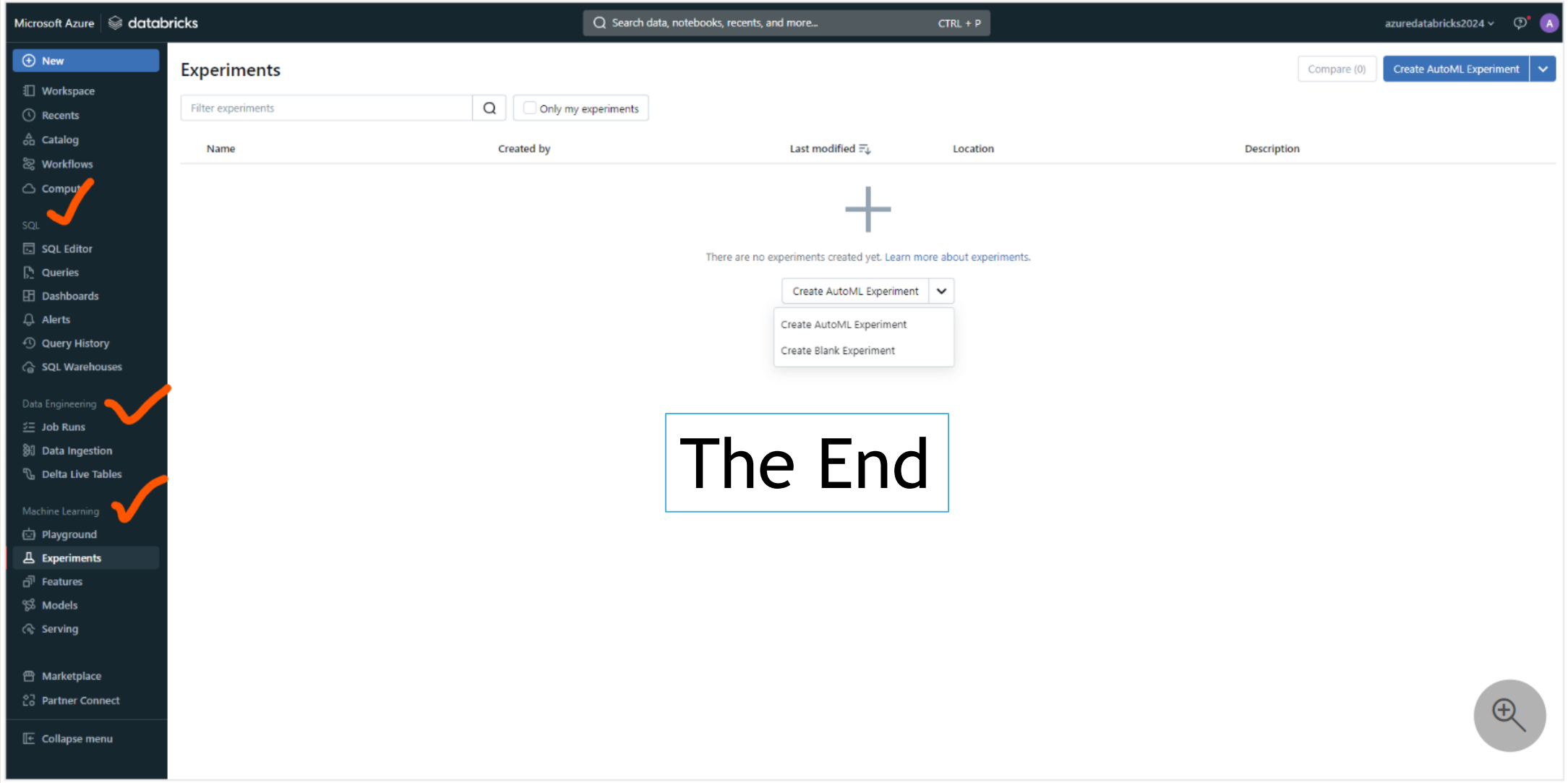
Data governance is important for ensuring that data within an organization is managed securely, efficiently, and in compliance with regulations. Azure Databricks, combined with Unity Catalog and Microsoft Purview, provides a solution for managing and governing data effectively.

Microsoft

Microsoft Purview	Databricks Unity Catalog
•**Ecosystem**: Primarily designed for the Azure ecosystem, including Azure Synapse Analytics, SQL Server, Power BI, and more. •**Features**: Offers data discovery, lineage tracking, data classification, unified data maps, and access controls. •**Use Cases**: Ideal for organizations looking to manage and govern data across various environments, including on-premises, multi-cloud, and SaaS applications	•**Ecosystem**: Tailored for the Databricks ecosystem, providing a cloud-agnostic approach to data lake governance. •**Features**: Includes data search and discovery, automated lineage, granular access controls, AI-powered monitoring, and open data sharing. •**Use Cases**: Best suited for managing data and AI workloads within the Databricks Lakehouse platform

Purview is a comprehensive solution for broader data governance across multiple platforms (saas / paas /Multi cloud), while Unity Catalog focuses on governance within the Databricks environment

Microsoft


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
azuredatabricks2024
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving
Marketplace
Partner Connect
Collapse menu
Experiments
Compare (0)
Create AutoML Experiment
Filter experiments
Only my experiments
Name
Created by
Last modified
Location
Description
There are no experiments created yet. Learn more about experiments.
Create AutoML Experiment
Create AutoML Experiment
Create Blank Experiment
The End