

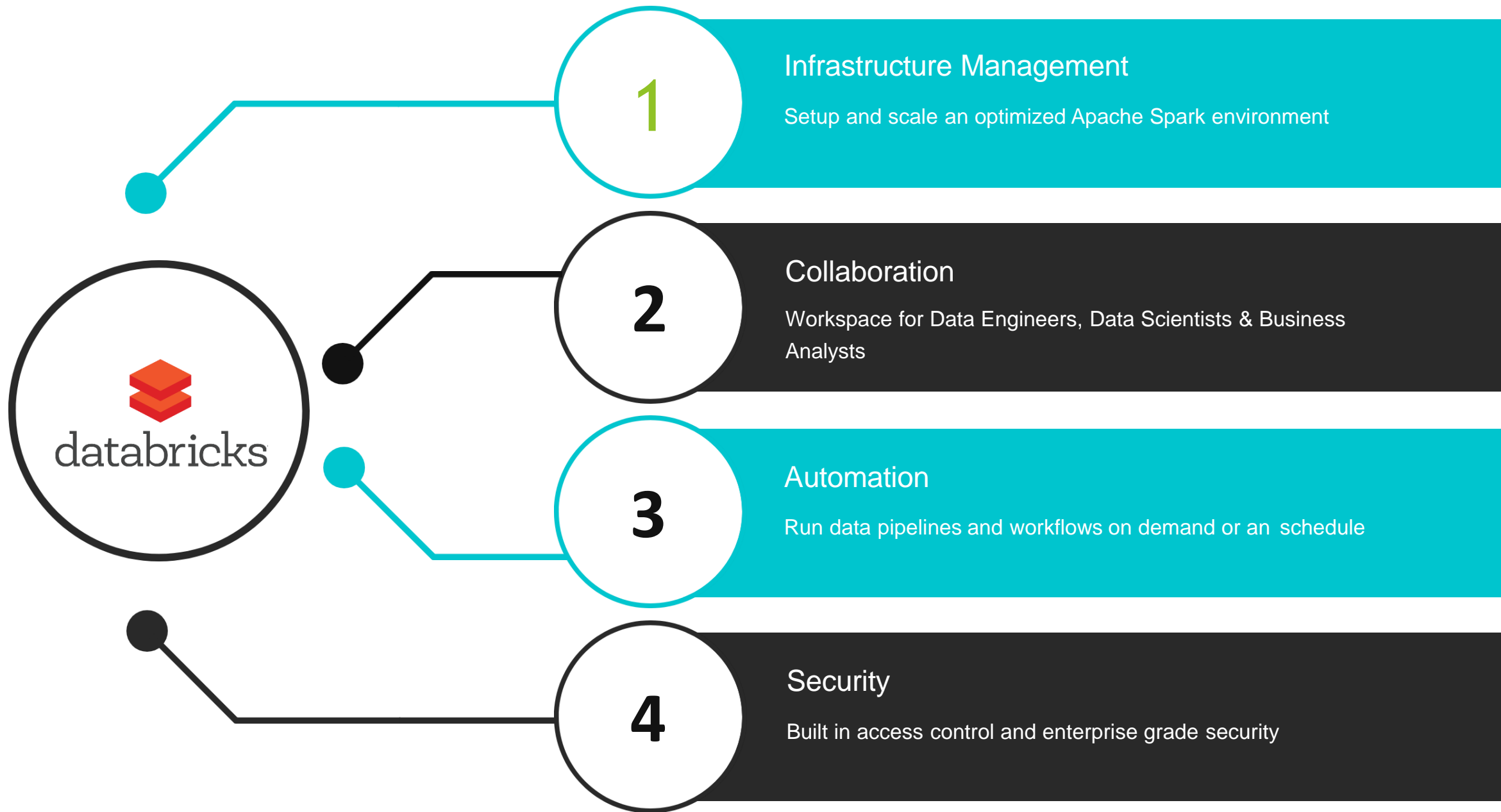
Microsoft Azure Databricks

A fast, easy, and collaborative Apache Spark™ based analytics platform optimized for Azure



Azure Databricks







Azure Databricks

Managed 1st Party Azure Service

Native integration with Azure & Its services;
Azure SLA and support

Transparency

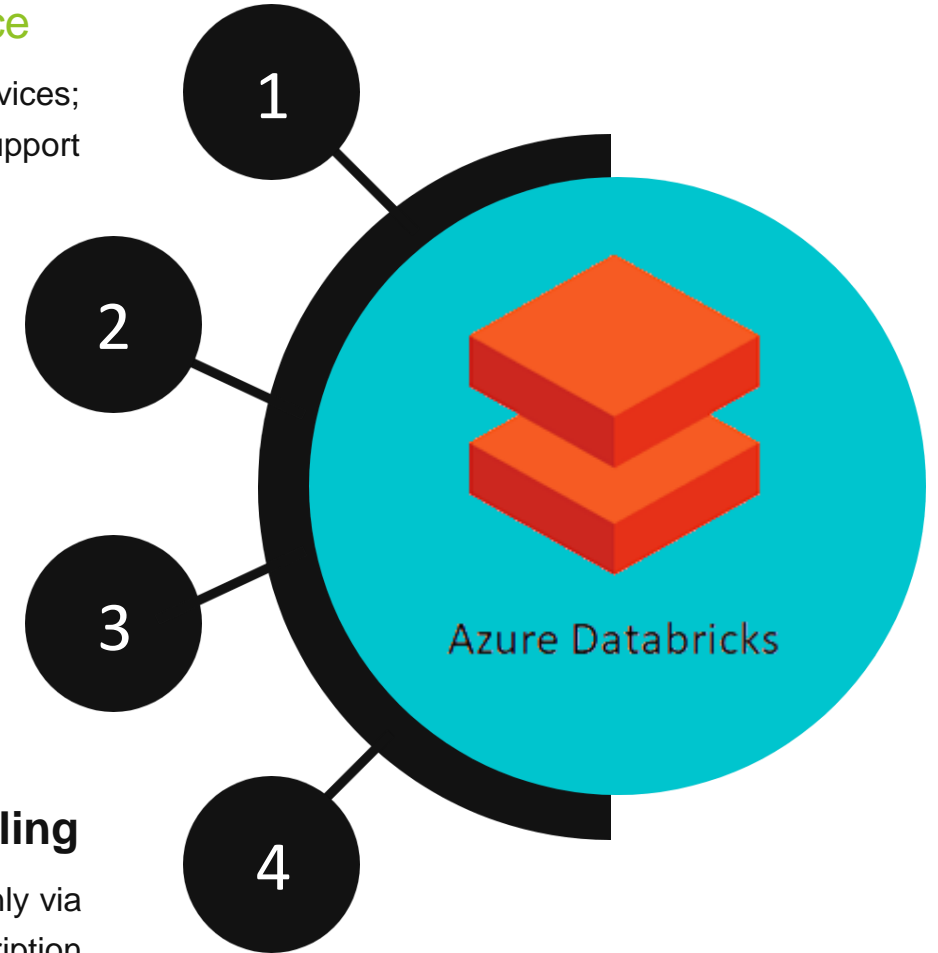
Deploys Databricks workspace and
clusters in customer subscription

Security

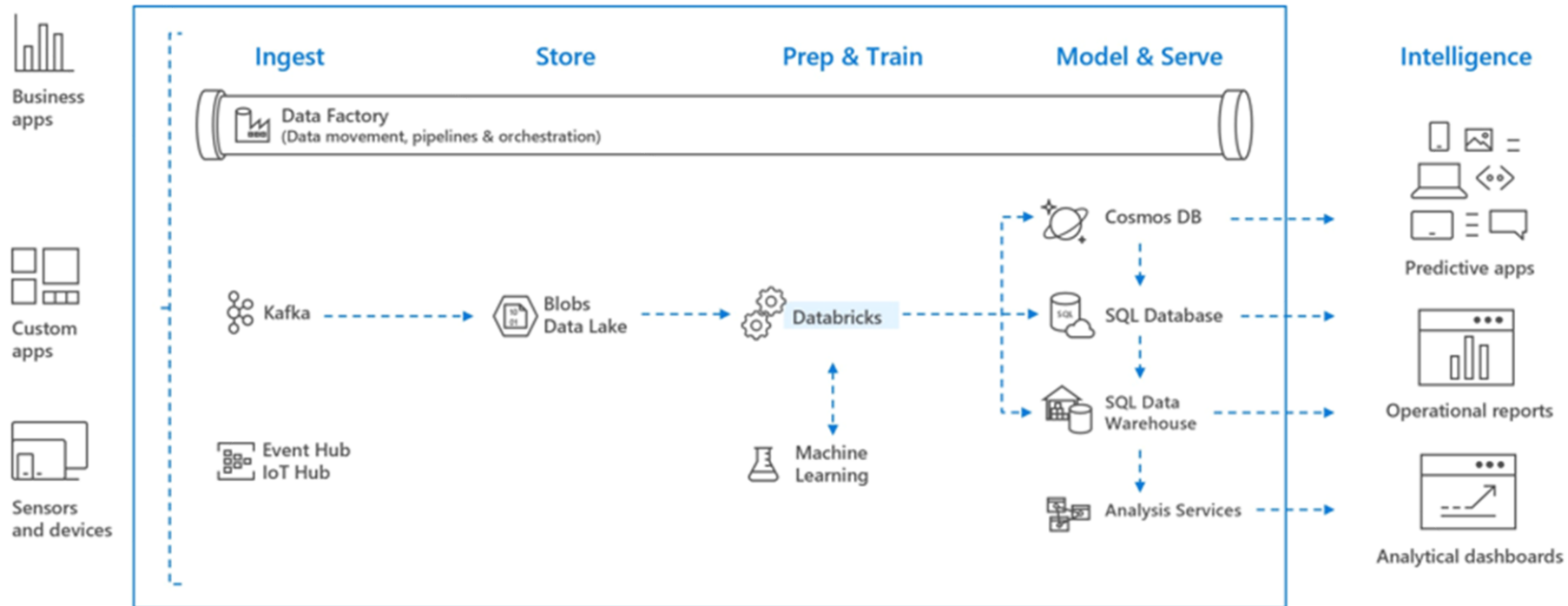
Natively integrates with Azure
Active Directory & Providers RBAC

United Billing

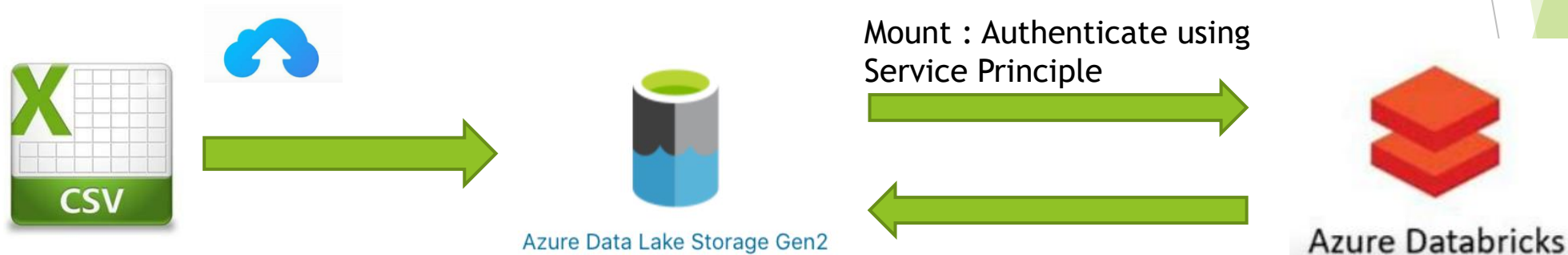
Pay for what you use only via
Azure subscription



Azure Databricks Architecture



Demo Overview



Cluster Types



Interactive Cluster

Multiple users interactively analyze the data together



Job Cluster

Created and terminated for running automated jobs

Cluster Types

Interactive Cluster

- Interactively analyze the data
- Created by users
- Manually terminate
- Option to auto terminate, if inactive
- Low execution time
- Auto scale on demand
- Comparatively costly

Job Cluster

- Run automated jobs
- Auto created when job starts
- Terminates when the job ends
- Option to auto terminate not applicable
- High throughput
- Auto scale on demand
- Comparatively cheaper

Cluster Types

Standard Mode

Single user

No fault isolation

No task preemption

Each user require separate cluster

Supports Scala, Python, SQL, R % Java

High Concurrency Mode

Multiple users

Fault isolation

Task preemption – fair resource sharing

Maximum cluster utilization

Only supports Python, SQL & R

Cluster

There are two types of nodes



Worker Nodes

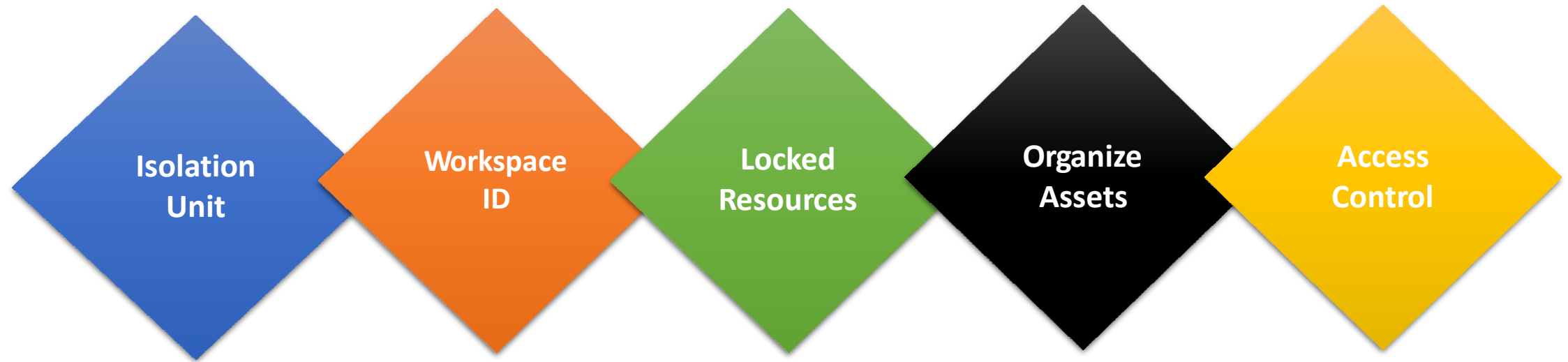
Multiple nodes perform data processing task



Driver Node

Distributes task to workers and coordinates execution

Worksp ace



Each workspace
is isolated from
others

Each workspace
has an identifier

Deployed in
control plane
and data plane

Notebooks,
Libraries,
Dashboard etc.

Define access
control on all
assets

Notebooks

Languages

Code in any
Spark supported
Languages

Workflows

Invoke notebook
from others &
pass data

Execution

Run directly on
clusters or via
jobs

Visualization

Turn data into
graphs or build
dashboards

Collaboration

Multiple users
can edit and
share comments



Jobs

- Execution of a notebook or JAR
- It can run immediately or on schedule
- Create job clusters to run jobs
- Each job can have different cluster configuration
- Monitor job runs and setup alerts

- Install 3rd party libraries
- Can be in any supported language
- Import the library into notebook to work
- Scoped at:
 - Cluster
 - Notebook



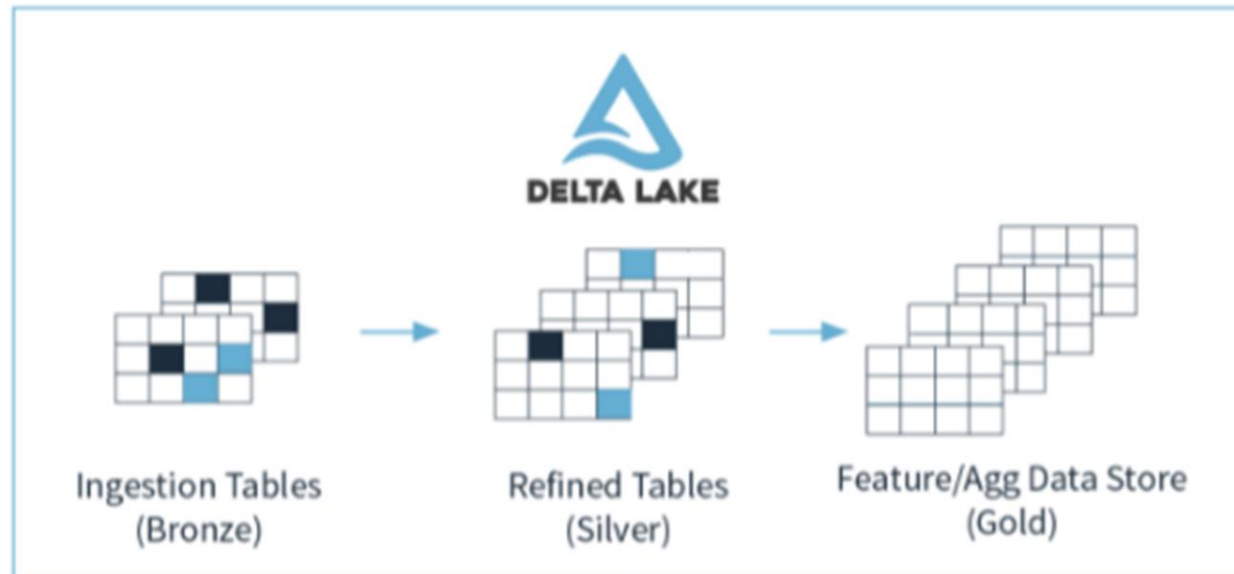
Libraries

- Create databases and tables inside them
- Table:
 - Collection of structured data
 - Equivalent to DataFrame – perform same operations on table
 - Created using files lying on storage
 - Directly query or write to tables



Database & Tables

Delta Lake

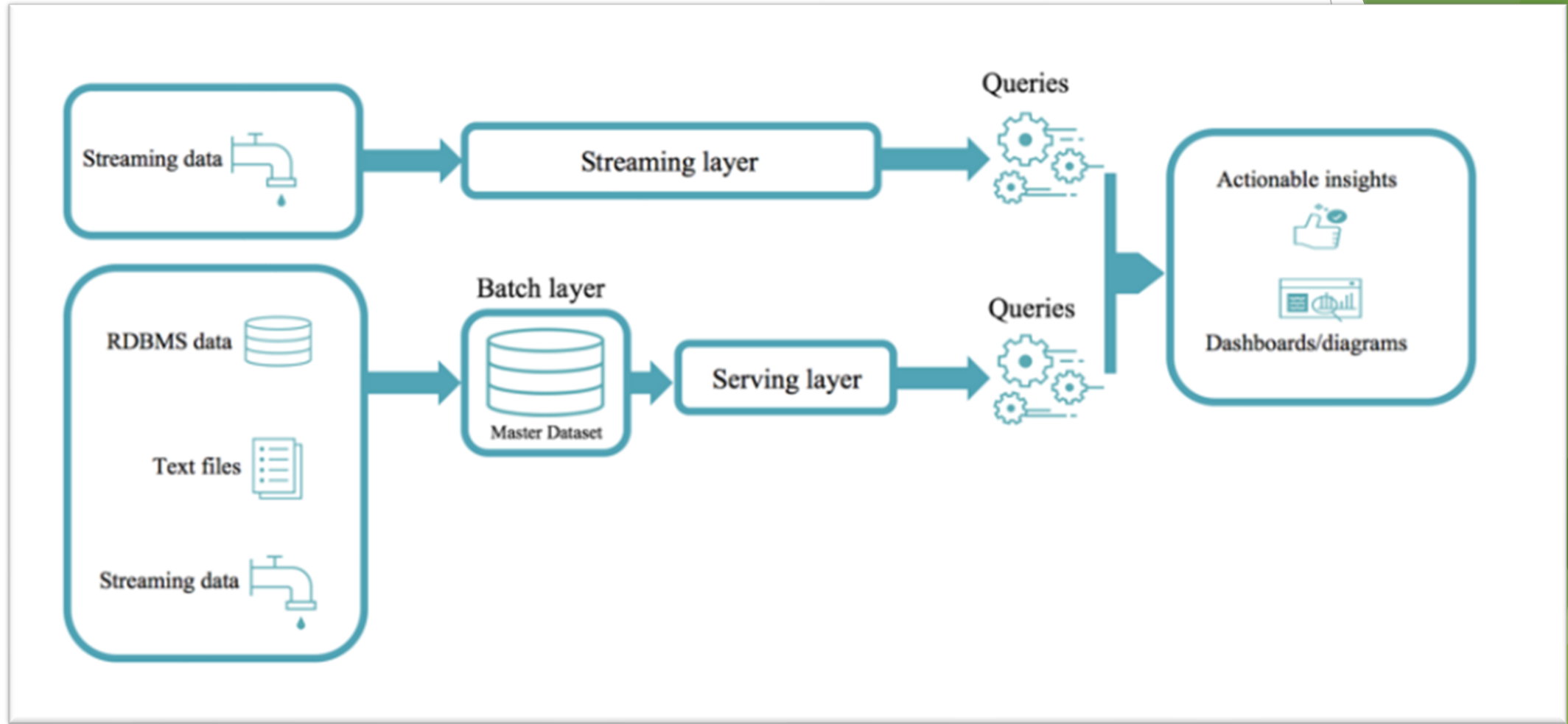


Analytics
and Machine
Learning

Your Existing Data Lake



Lambda Architecture



Azure Delta Lake Architecture

