1

Infrastructure Management

Setup and scale an optimized Apache Spark environment

2

Collaboration

Workspace for Data Engineers, Data Scientists & Business Analysts

3

Automation

Run data pipelines and workflows on demand or an schedule

4

Security

Built in access control and enterprise grade security

Microsoft Azure Databricks

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

Azure Databricks

Managed 1st Party Azure Service

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

Transparency

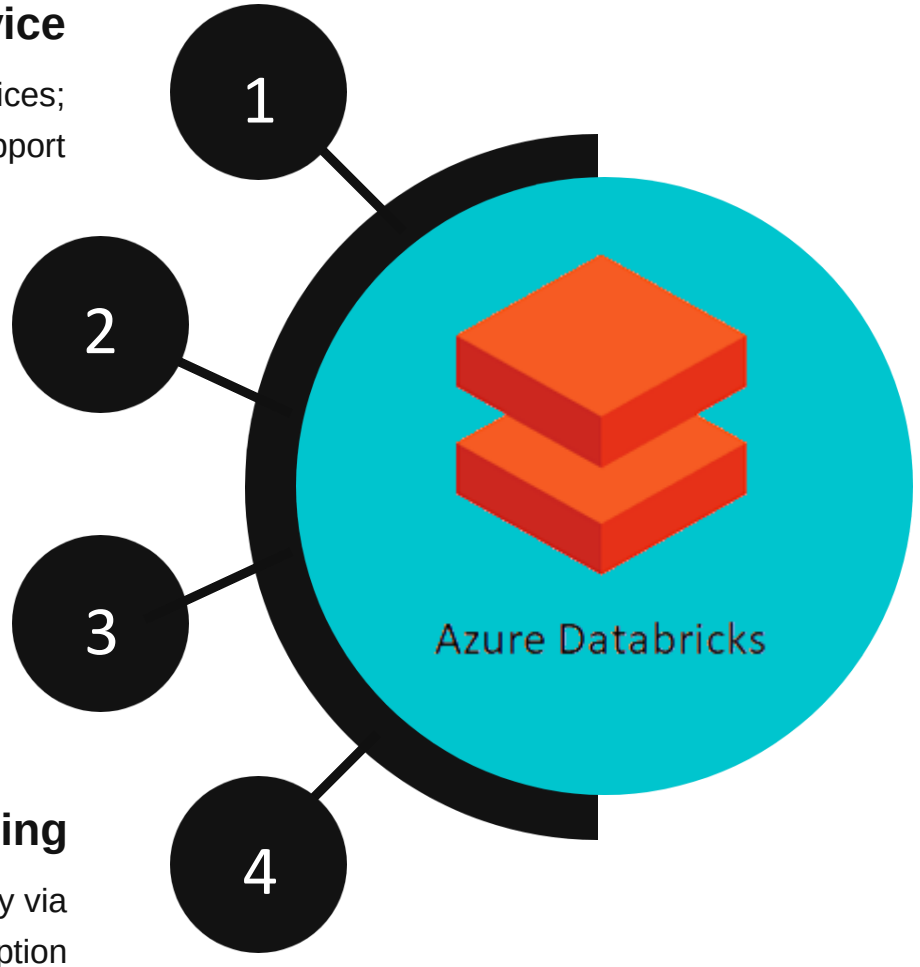
Deploys Databricks workspace and
clusters in customer subscription

Security

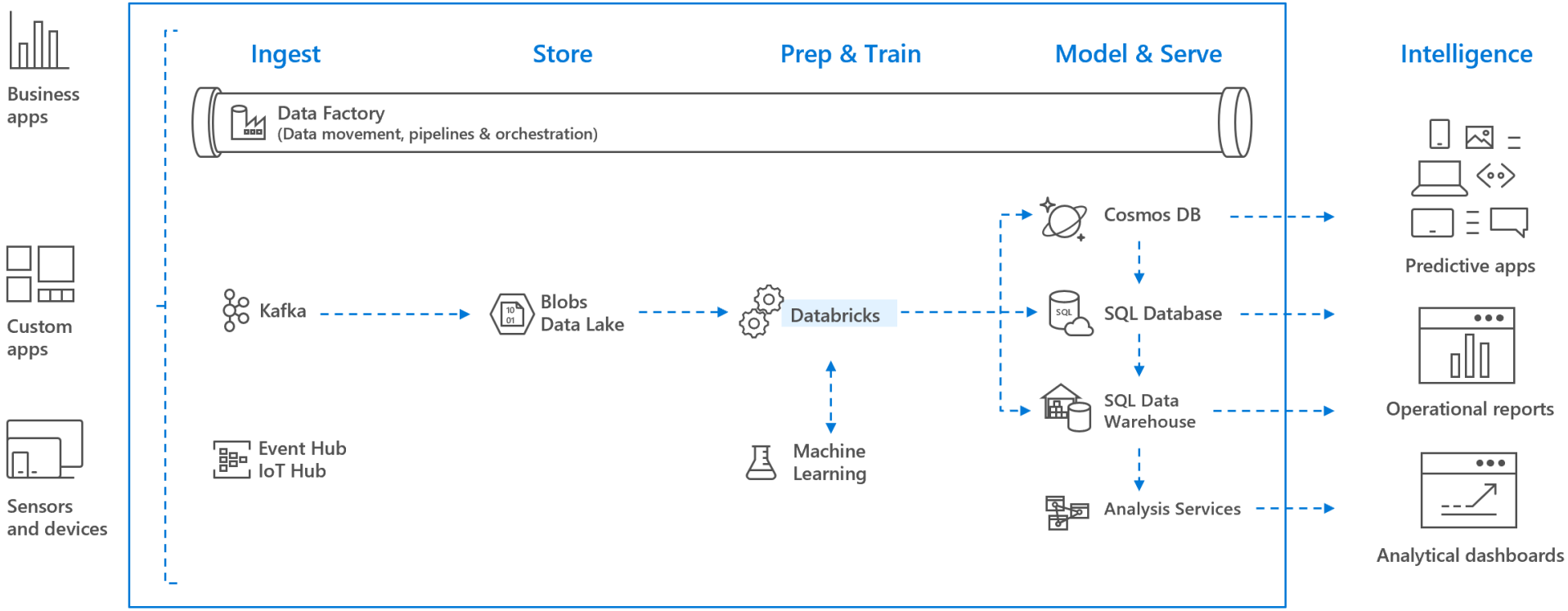
Natively integrates with Azure
Active Directory & Providers RBAC

United Billing

Pay for what you use only via
Azure subscription



Azure Databricks Architecture



Cluster Types

Interactive Cluster

Multiple users interactively analyze the data together

Job Cluster

Created and terminated for running automated jobs

Cluster Types

Interactive Cluster	Job Cluster
Interactively analyze the data	Run automated jobs
Created by users	Auto created when job starts
Manually terminate	Terminates when the job ends
Option to auto terminate, if inactive	Option to auto terminate not applicable
Low execution time	High throughput
Auto scale on demand	Auto scale on demand
Comparatively costly	Comparatively cheaper

Cluster Types

Standard Mode	High Concurrency Mode
Single user	Multiple users
No fault isolation	Fault isolation
No task preemption	Task preemption – fair resource sharing
Each user require separate cluster	Maximum cluster utilization
Supports Scala, Python, SQL, R % Java	Only supports Python, SQL & R

Cluster

There are two types of nodes

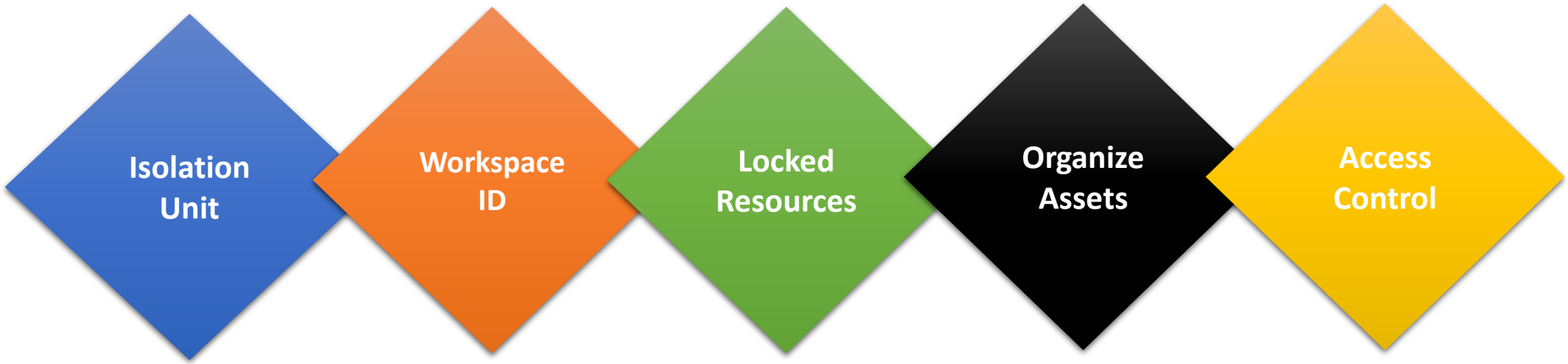
Worker Nodes

Multiple nodes perform data processing task

Driver Node

Distributes task to workers and coordinates execution

Workspace



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

Code in any Spark supported Languages

Invoke notebook from others & pass data

Run directly on clusters or visa jobs

Turn data into graphs or build dashboards

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