DP-203 Microsoft Azure Data Engineer

Day7 – Azure Databricks

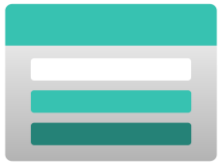
31st July 2021

Vinodkumar Bhovi

databoq®

Data

Data Storage



Azure Storage Accounts

Azure Cosmos DB

Azure Data Lake

Azure SQL

Azure Synapse Analytics

Data Transformation

Azure Data Factory

Azure Stream Analytics

Azure Databricks

Azure HDInsight

databag®

Why Map reduce introduced

Tradition system were failing

Centralized server to store and process data which creates bottleneck

Google introduced Map Reduce

Divide the task into small parts and assign them to many computers

databag®

Challenges with Map Reduce

Data-sharing abstraction

Concurrent data access to memory across the cluster

Inefficient use of resources

Poor memory utilization by spilling to disk after each job

databag®

Apache Spark

Big Data Tool

"Spark is an open source unified analytics engine for large-scale data processing."

databag®

1

Fast and efficient

Supports structured, semi-structured or unstructured.

2

In-memory engine

100 times faster than Hadoop

3

Highly scalable architecture

Process terabytes of data in parallel

4

Unifies variety of use cases

Batch processing, streaming ingestion, machine learning and advanced analytics

databag®

Hadoop

HDFS

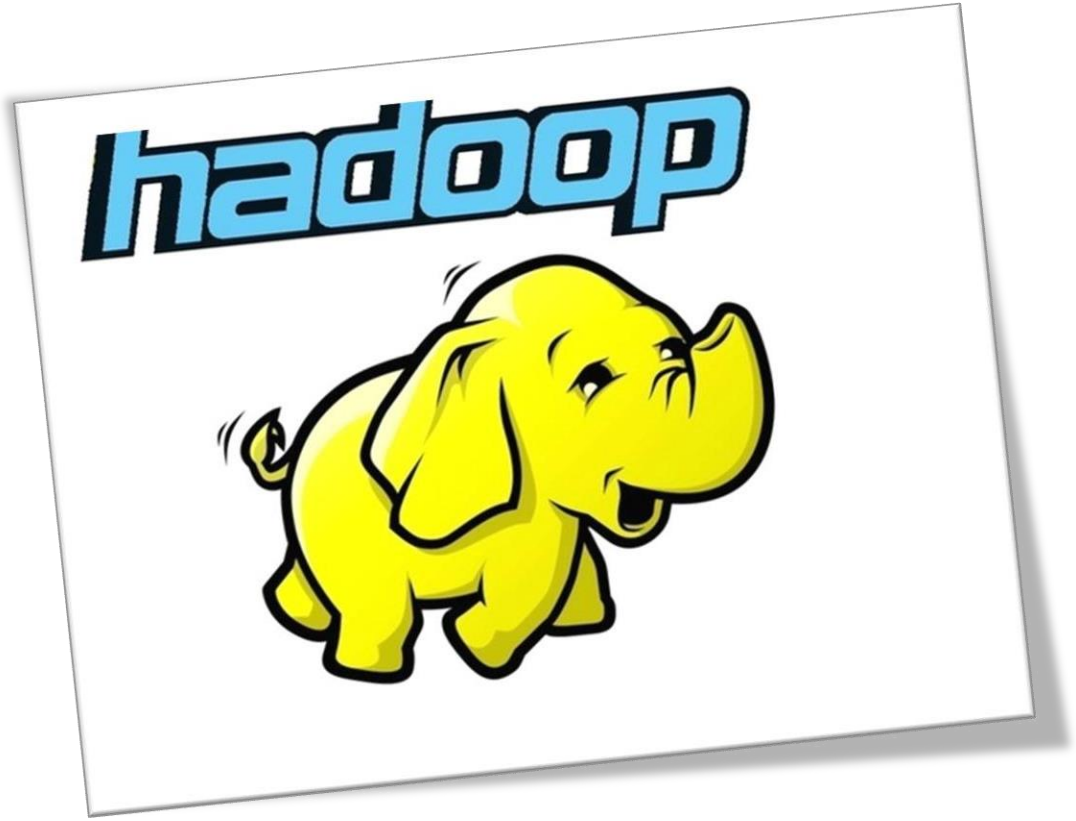
A file system to manage the storage of data

MapReduce

A framework to define a data processing task

YARN

A framework to run the data processing task



databag®

Co-ordination between Hadoop Blocks

1. Step One

User defines map and reduce tasks using the MapReduce API

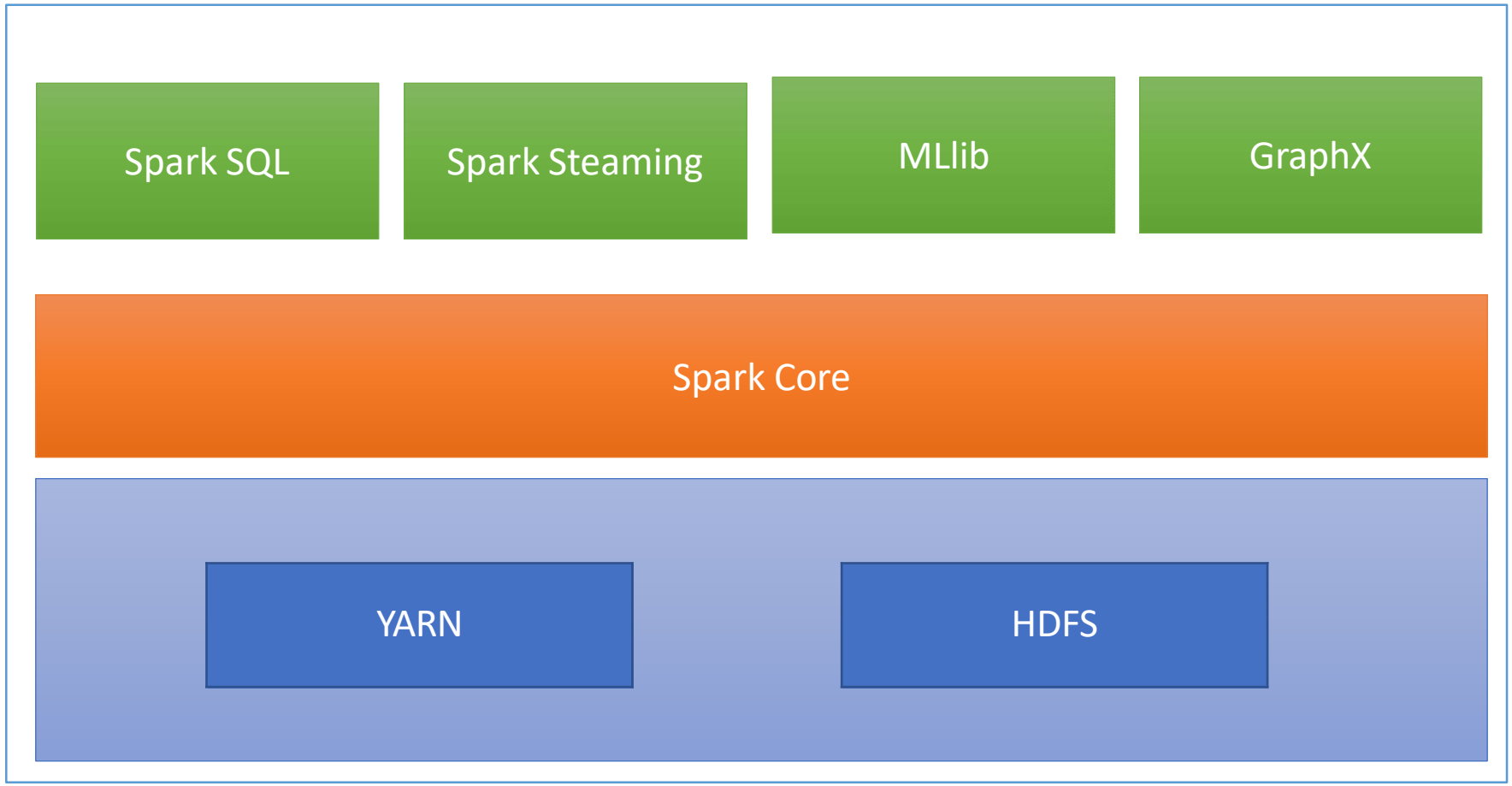
2. Step Two

YARN takes care of resource allocation and figures out where and how to runt he job

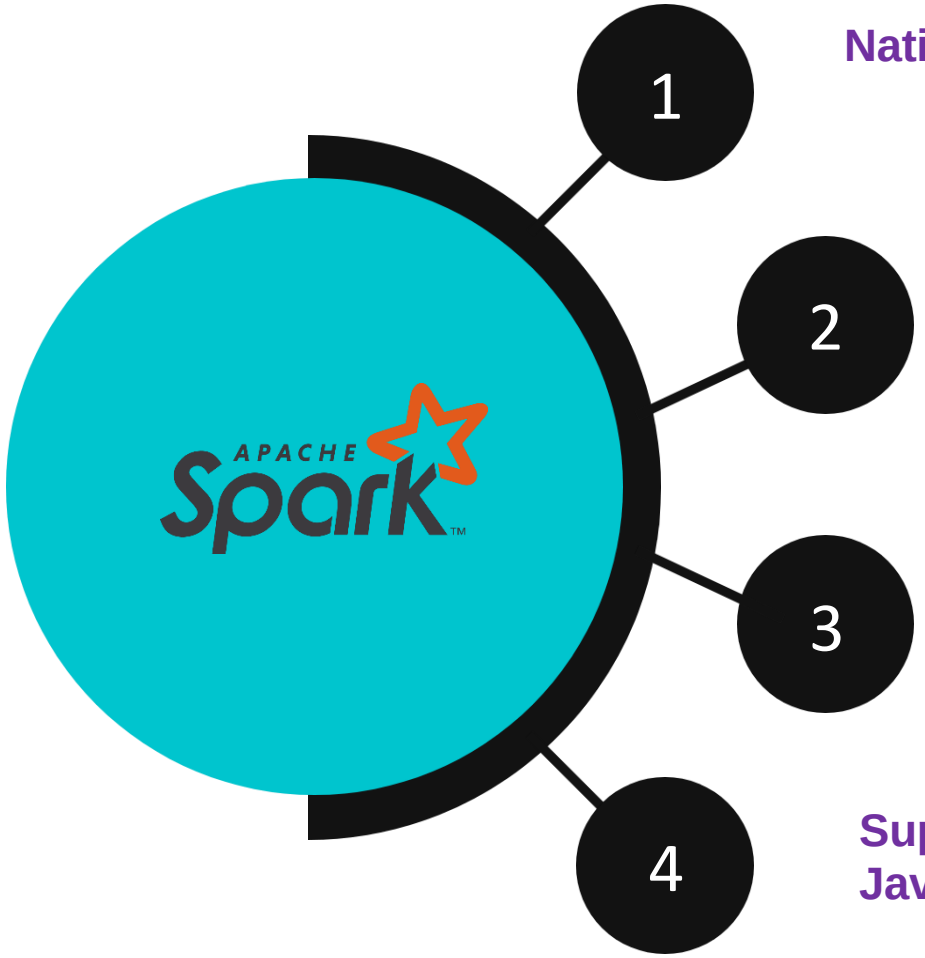
3. Step Three

YARN stores the result in HDFS

Apache Spark Ecosystem



databag®



Native support

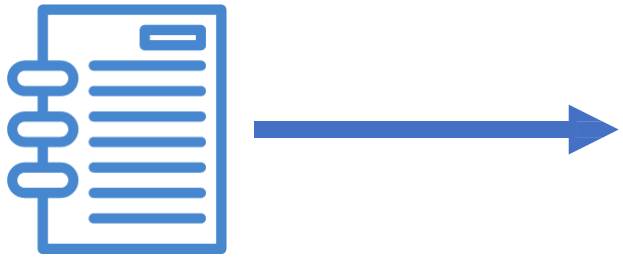
Real-time as well as batch processing

Interactive REPL environment

Support for Phthon, Java, Scala and R

databag®

DataFrame: Data in Rows and Columns



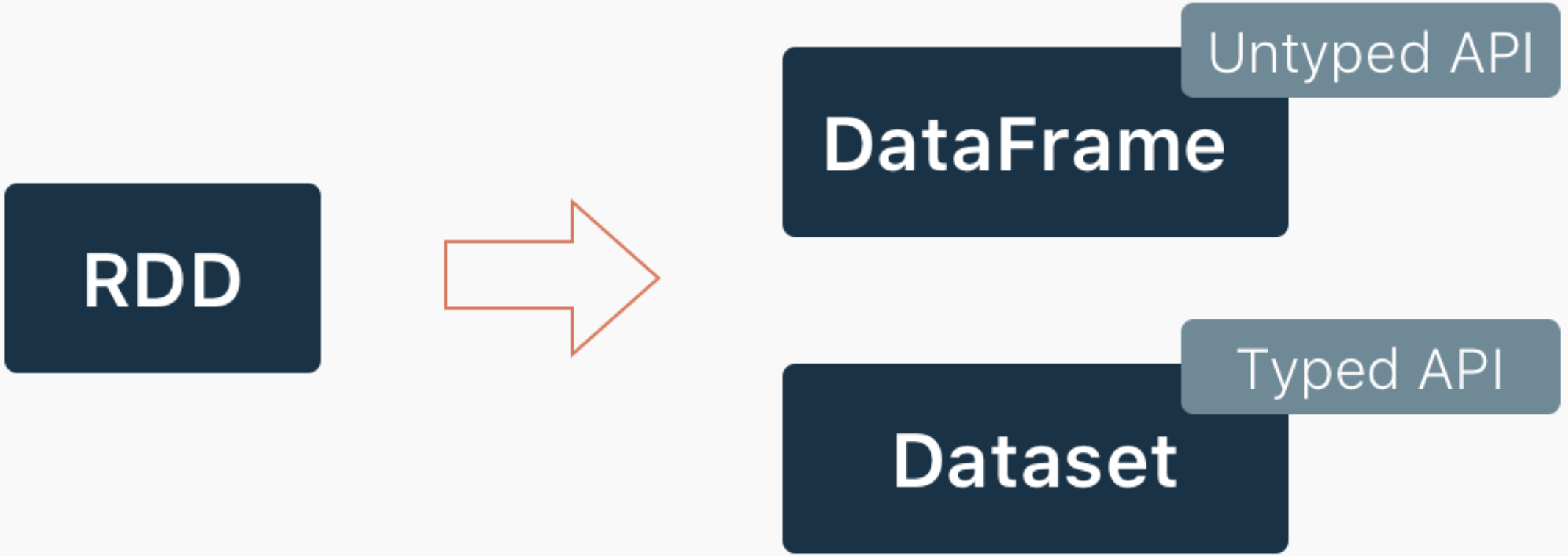
File

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

DataFrame

databag®

Apache Spark API



databag®

RDDs to Dataset

RDDs

Primary abstraction since initial versions

Immutable and distributed

Strong typing, use of Lambda

No optimized execution

Available in all languages

Datasets

Added to Spark in 1.6

Also immutable and distributed

Also support strong typing, lambdas

Leverage optimizers in recent versions

Present in Scala and Java, not python or R

databag®

Datasets to DataFrames

Datasets

Added to Spark in 1.6

Immutable and distributed

No named columns

Extension of DataFrames – OOP interface

Compile time type safety

Present in Scala, Java, not Python, R

DataFrames

Added to Spark in 1.3

Also immutable and distributed

Named columns, like Pandas or R

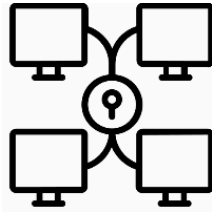
Conceptually equal to a table in an RDBMS

No type safety at compile time

Available in all languages

databag®

What makes Apache Spark difficult to use?

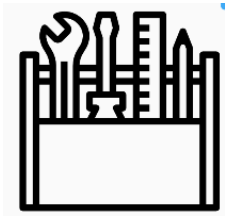


Infrastructure Management

Upgrade Challenge

User Interface

Manual Configuration

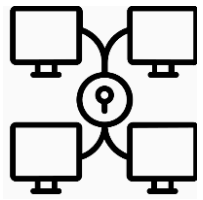


Tooling & Integration Complexity

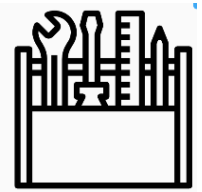
Difficult to Collaborate on Projects

databag®

databricks



Efficient and Interactive Platform



Tools are available

Integrated and Interactive workspace

User Interface to manage Infrastructure (Scalability, failure recovery, upgrades)

databag®

Distributed processing of data

In-memory

Language support

- Scala, Python, SQL, R & Java

Use cases

- Batch & Stream processing
- Machine learning
- Advanced Analytics

An Apache Spark based Unified Analytics Platform, optimized for the cloud

databag®

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

databag®

Microsoft Azure Databricks

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

databag®

Azure Databricks

Managed 1st Party Azure Service

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

Transparency

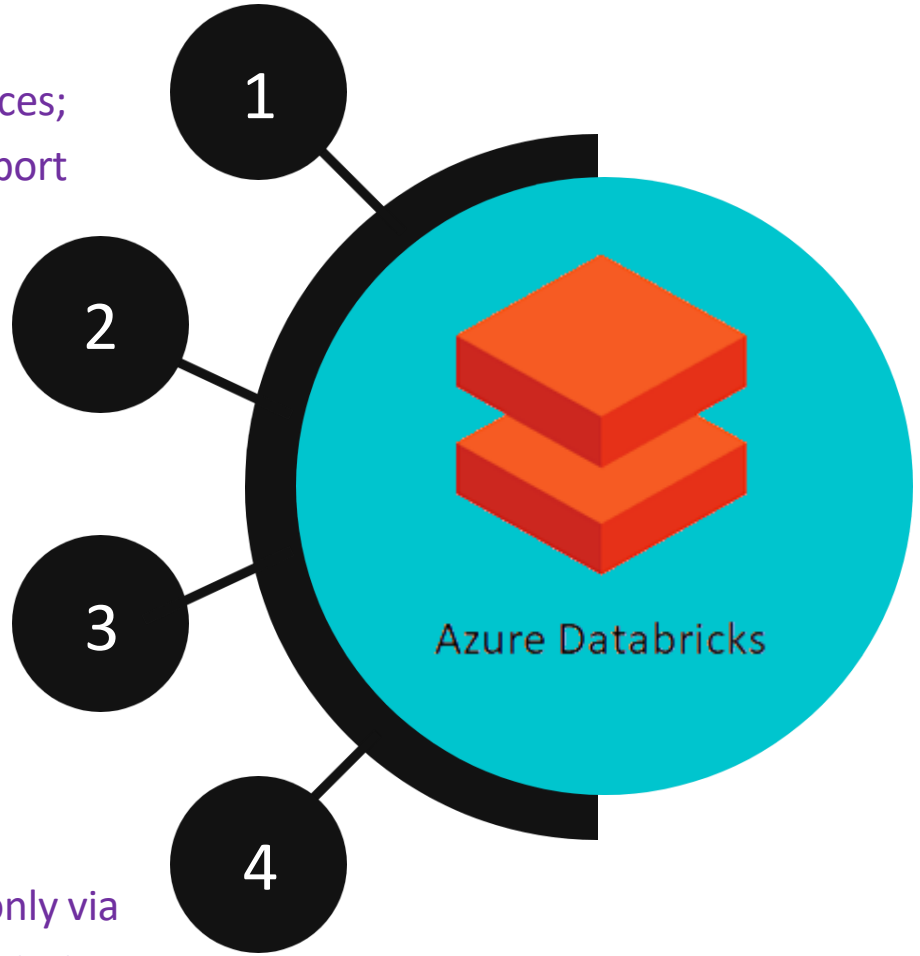
Deploys Databricks workspace and
clusters in customer subscription

Security

Natively integrates with Azure
Active Directory & Providers RBAC

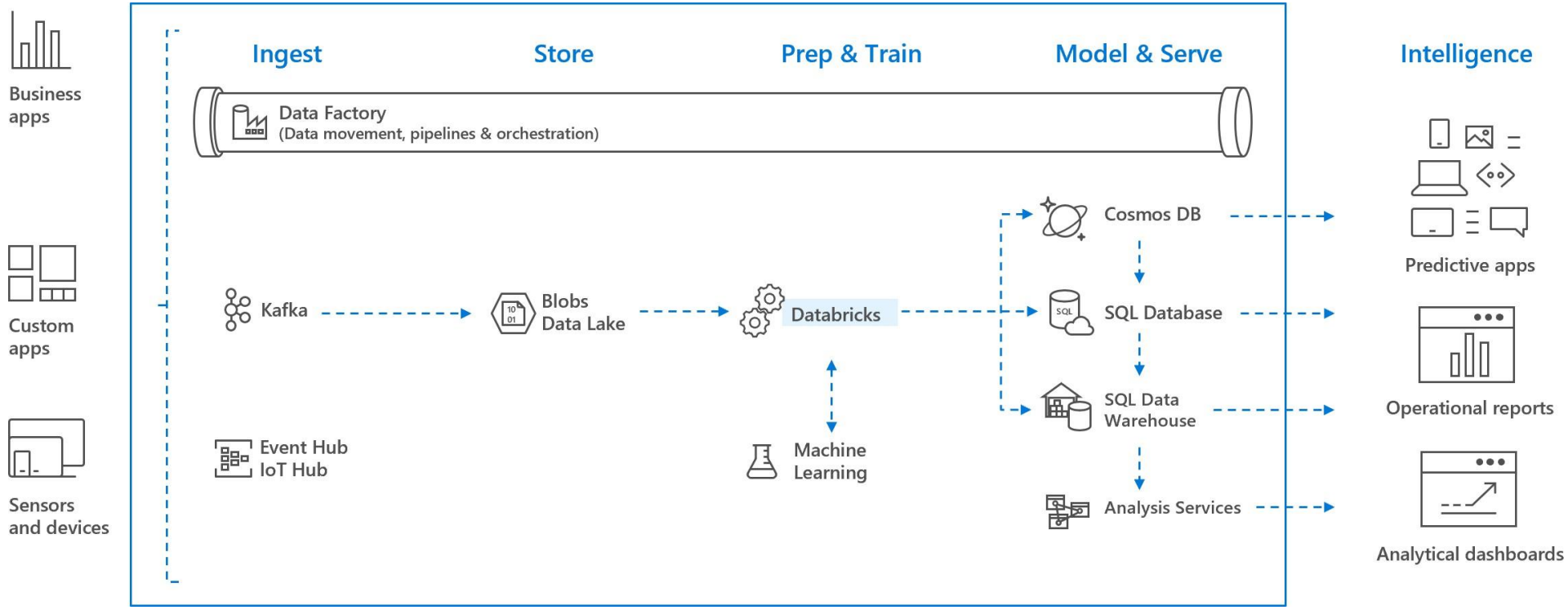
United Billing

Pay for what you use only via
Azure subscription

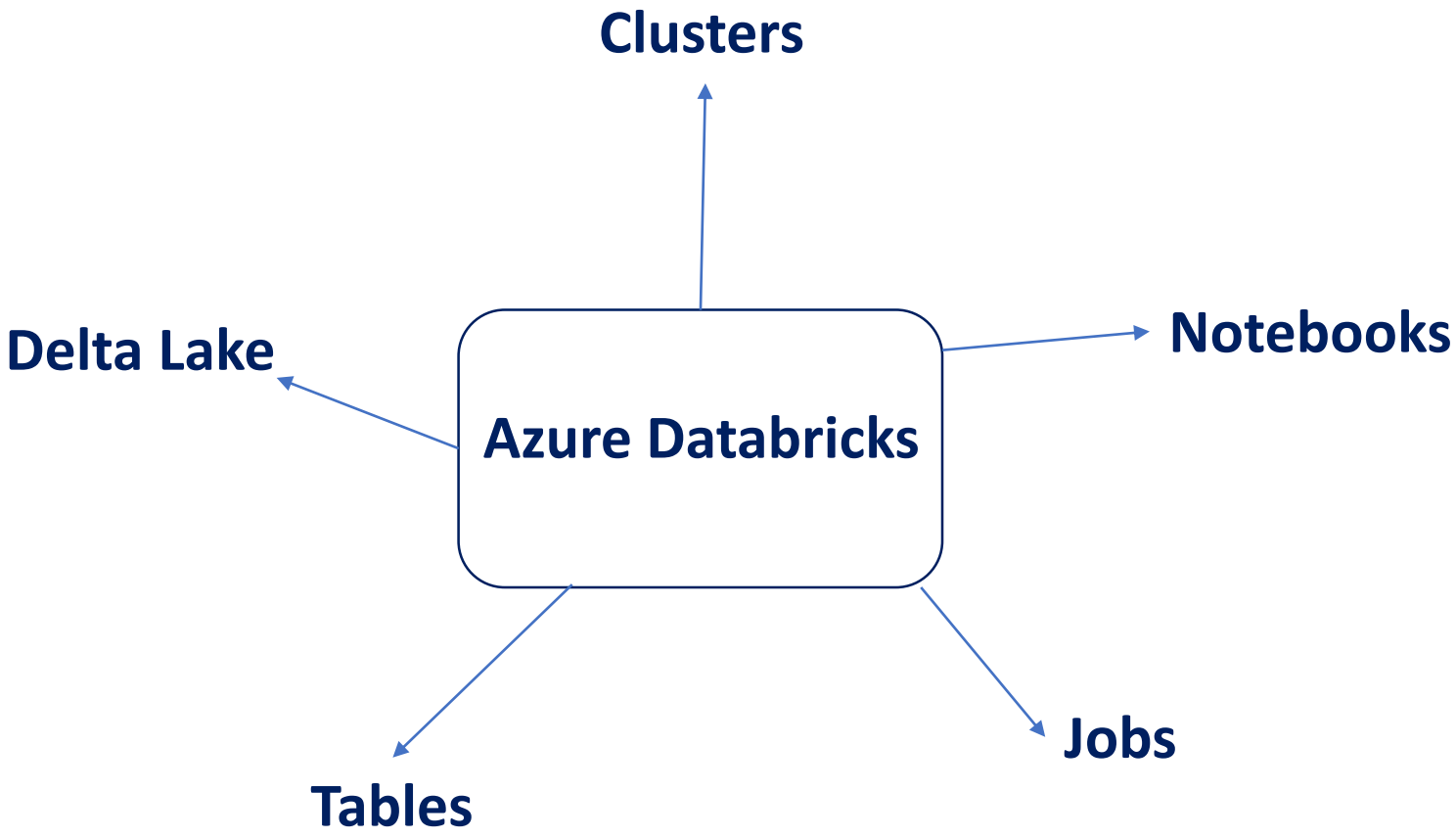


databag®

Azure Databricks Architecture



Azure Databricks



Clusters

The cluster will have the Spark engine and other components installed.

Cluster

There are two types of nodes

Worker Nodes

Multiple nodes perform data processing task

Driver Node

Distributes task to workers and coordinates execution

Cluster Types

Interactive Cluster

Multiple users interactively analyze the data together

Job Cluster

Created and terminated for running automated jobs

databag®

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

databag®

Cluster Interactive Types

Standard Mode

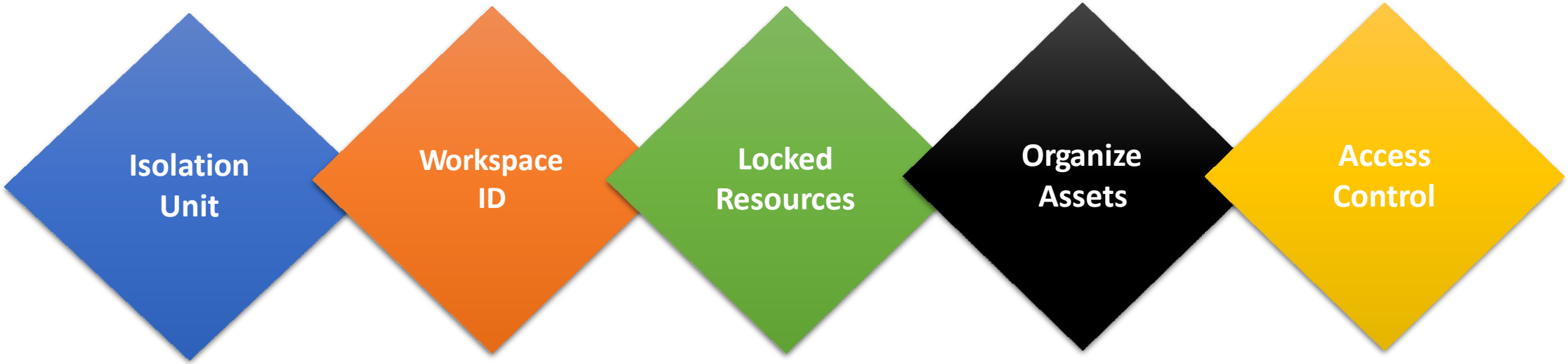
- Single user
- No fault isolation
- No task preemption
- Each user require separate cluster
- Supports Scala, Python, SQL, R % Java

High Concurrency Mode

- Multiple users
- Fault isolation
- Task preemption – fair resource sharing
- Maximum cluster utilization
- Only supports Python, SQL & R

databag®

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

databag®

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

.

databag®

- 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

databag®

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

databag®

- 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

Thank you!

databag®