Get Started with Data Engineering on Databricks

We will start at 3 minutes past the starting time...

©2023 Databricks Inc. — All rights reserved

Meet your instructor

Venkita Krishnan Mani, Technical Instructor

linkedin.com/in/venkitakrishnan

Now

- Technical Instructor

Then

- Big Data & Spark Consultant
  - – JPMC / Citi Bank / Deutsche Bank / Zarantec
- Hadoop Practice Lead –
  - - Nichetek / Collabera India / CavalierIT

Interests

- Consulting / Teaching & Mentoring

©2023 Databricks Inc. — All rights reserved

databricks

Get Started with Data Engineering on Databricks

©2023 Databricks Inc. — All rights reserved

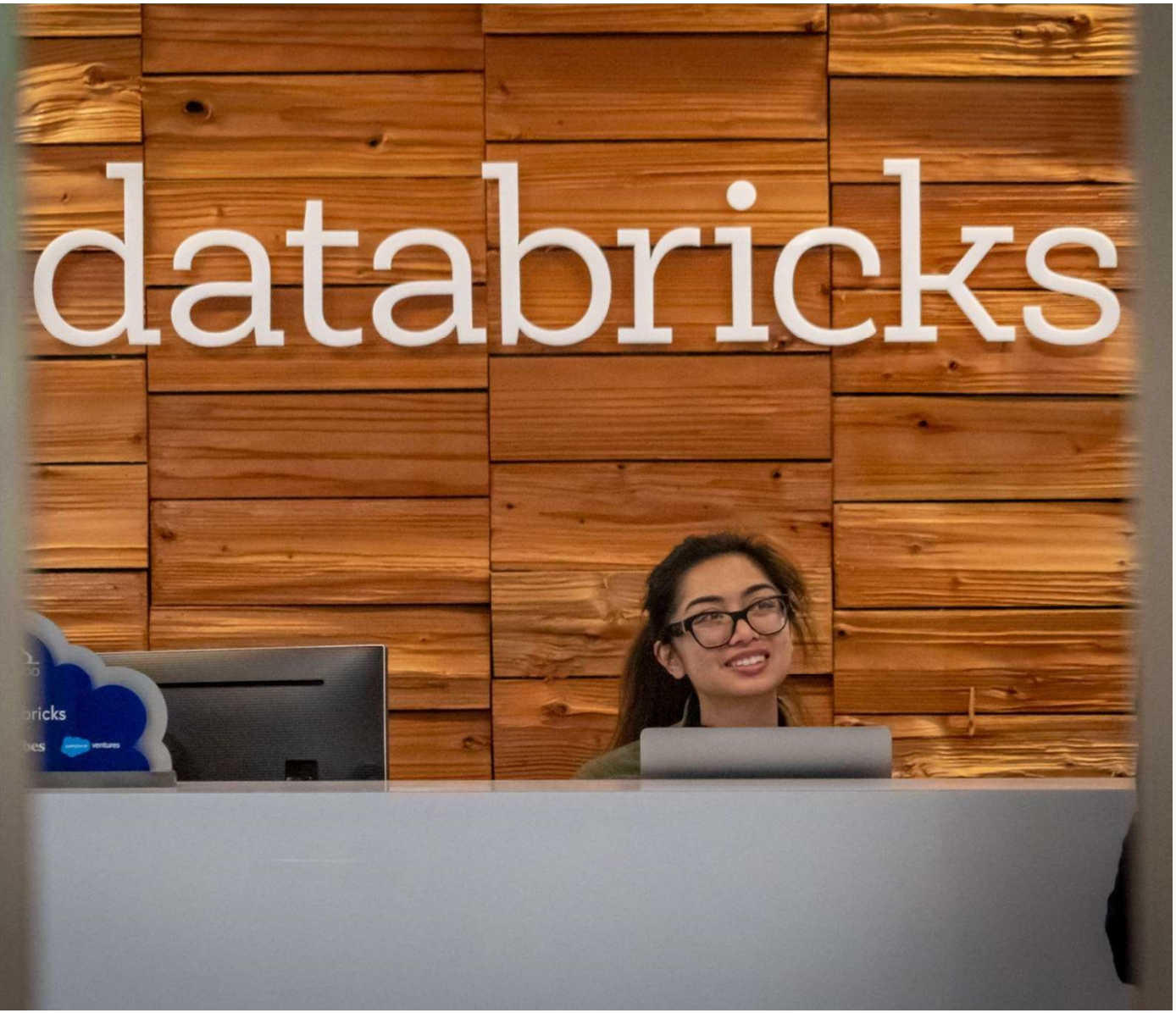
Session topics

- Databricks Lakehouse Platform overview
- Databricks Architecture and Services
- Data Science and Engineering Workspace feature dive/demos:
  - UI tour
  - Creating and configuring clusters
  - Developing code with Databricks Notebooks
  - Git versioning with Databricks Repos
- Managing Data with Delta Lake
  - Delta Lake overview
  - Setting up, versioning, and optimizing Delta tables
  - Loading data into Delta Lake

©2023 Databricks Inc. — All rights reserved

Databricks Lakehouse Platform Overview

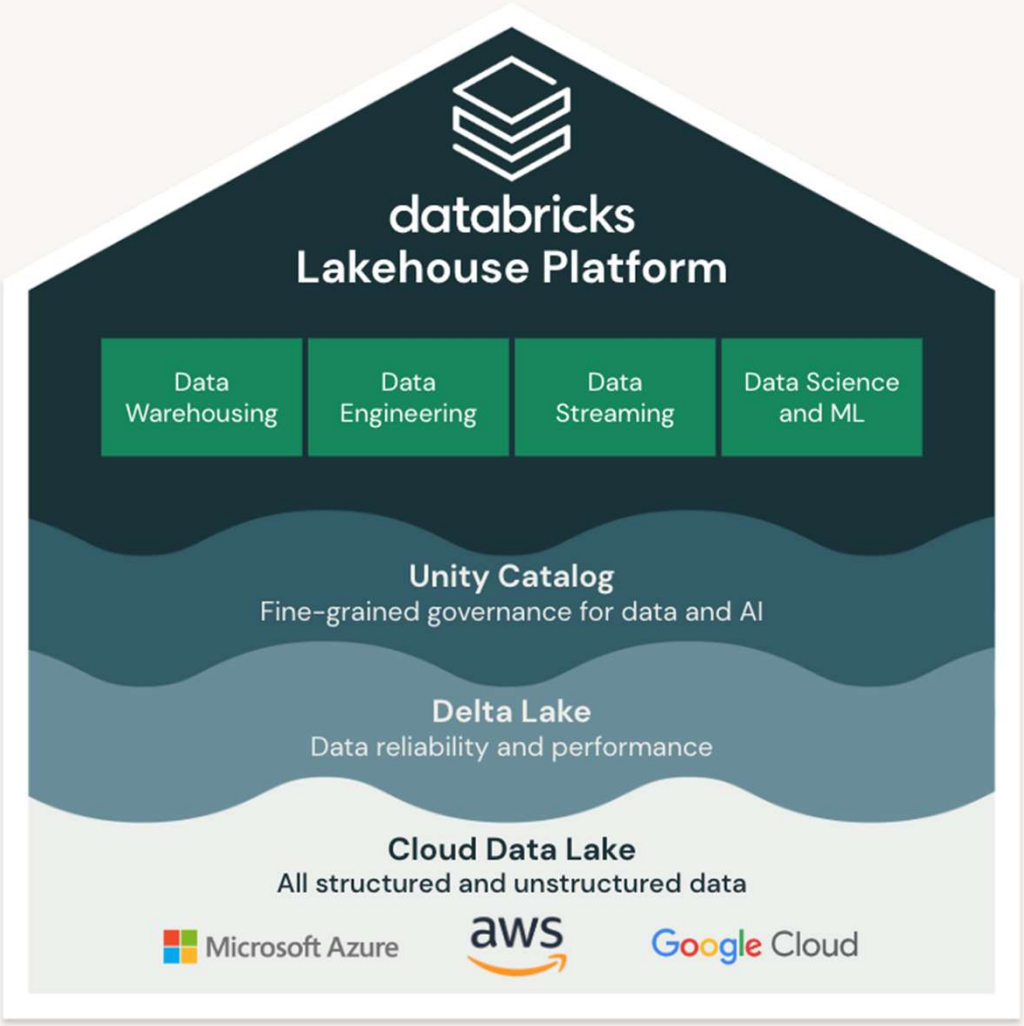
©2023 Databricks Inc. — All rights reserved



APACHE Spark™

DELTA LAKE

mlflow™



Databricks Lakehouse Platform

Simple

Unify your data warehousing and AI use cases on a single platform

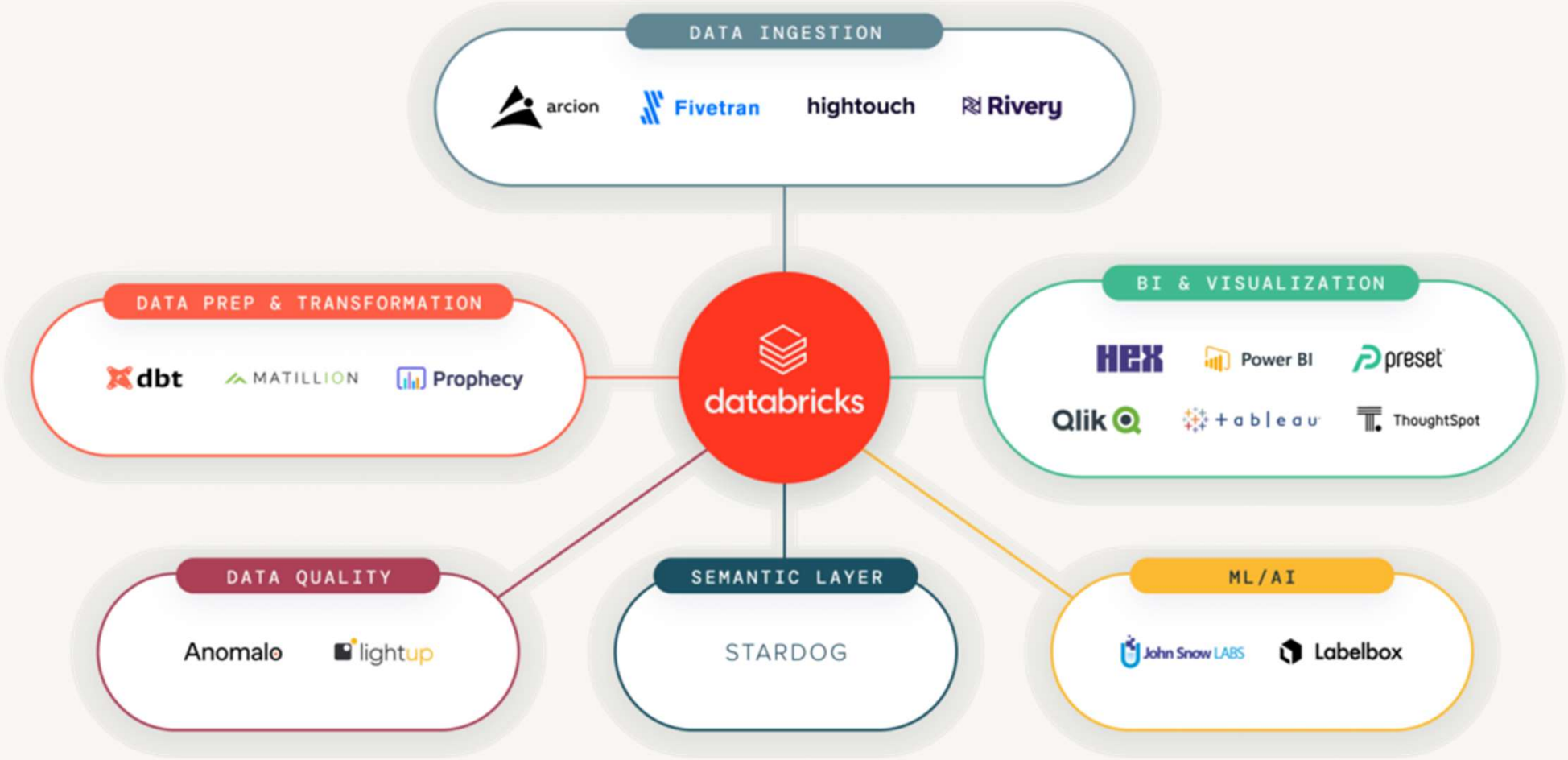
Open

Built on open source and open standards

Multicloud

One consistent data platform across clouds

©2022 Databricks Inc. — All rights reserved

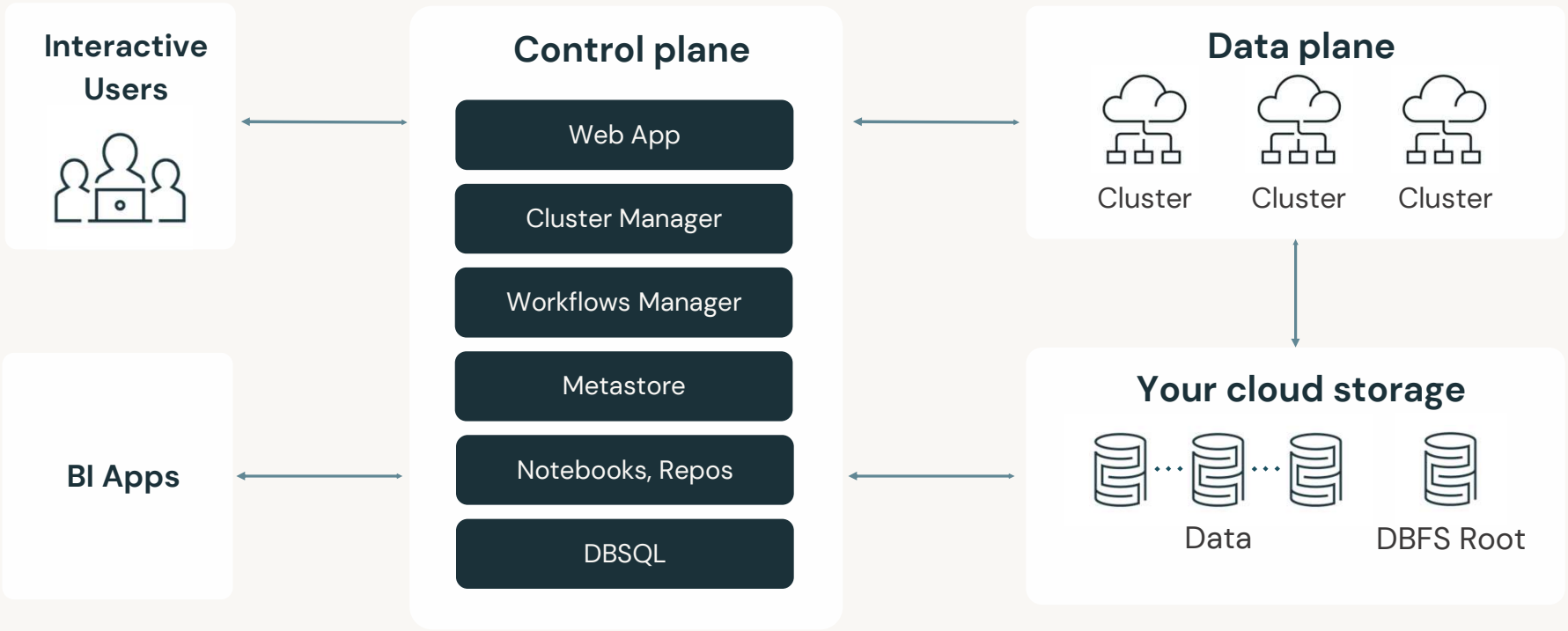


©2022 Databricks Inc. — All rights reserved

Databricks Workspace and Services

©2023 Databricks Inc. — All rights reserved

Databricks architecture



©2023 Databricks Inc. — All rights reserved

Demo: Navigate the Workspace UI

©2023 Databricks Inc. — All rights reserved

Compute Resources

©2023 Databricks Inc. — All rights reserved

Clusters

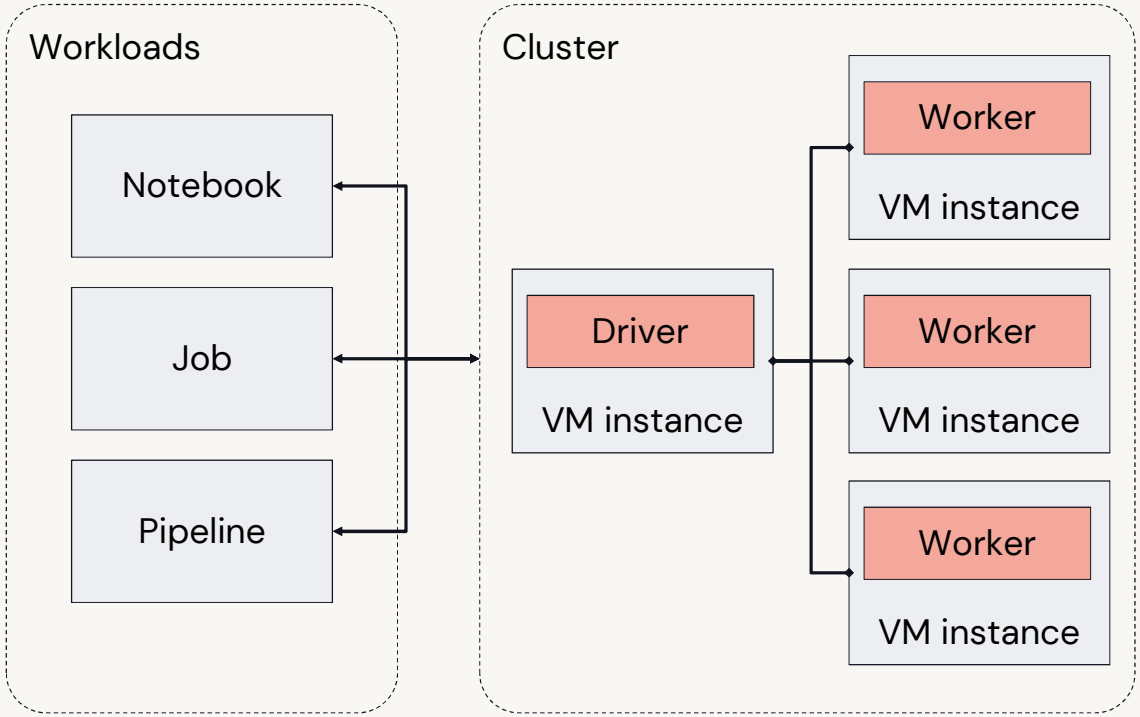
Overview

Collection of VM instances

Distributes workloads across workers

Two main types:

1. **All-purpose** clusters for interactive development
2. **Job** clusters for automating workloads



©2023 Databricks Inc. — All rights reserved

Cluster Types

All-purpose Clusters

Analyze data collaboratively using **interactive** notebooks

Create clusters from the Workspace or API

Configuration information retained for up to 70 clusters for up to 30 days

Job Clusters

Run **automated** jobs

The Databricks job scheduler creates job clusters when running jobs

Configuration information retained for up to 30 most recently terminated clusters

©2023 Databricks Inc. — All rights reserved

Cluster Configuration

©2023 Databricks Inc. — All rights reserved

Cluster Mode

Standard

Default mode for workloads developed in any supported language (requires at least two VM instances)

Single node

Low-cost single-instance cluster catering to single-node machine learning workloads and lightweight exploratory analysis

©2023 Databricks Inc. — All rights reserved

Databricks Runtime Version

Standard

Apache Spark and many other components and updates to provide an optimized big data analytics experiences

Photon

An optional add-on to optimize SQL workloads

Machine learning

Adds popular machine learning libraries like TensorFlow, Keras, PyTorch, and XGBoost.

©2023 Databricks Inc. — All rights reserved

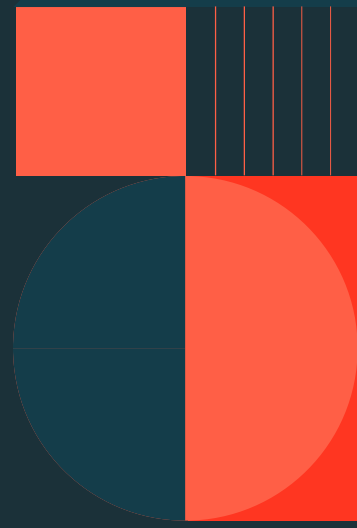
Cluster Policies

Cluster policies can help to achieve the following:

- Standardize cluster configurations
- Provide predefined configurations targeting specific use cases
- Simplify the user experience
- Prevent excessive use and control cost
- Enforce correct tagging

©2023 Databricks Inc. — All rights reserved

Demo: Create and manage interactive clusters



©2023 Databricks Inc. — All rights reserved

Develop Code with Notebooks & Databricks Repos

©2023 Databricks Inc. — All rights reserved

Notebook magic commands

Use to override default languages, run utilities/auxiliary commands, etc.

%python, %r, %scala, %sql	Switch languages in a command cell
%sh	Run shell code (runs only on Spark Driver, and not the Workers)
%fs	Shortcut for dbutils filesystem commands
%md	Markdown for styling the display
%run	Execute a remote Notebook from a Notebook
%pip	Install new Python libraries

©2023 Databricks Inc. — All rights reserved

`dbutils` (Databricks Utilities)

Perform various tasks with Databricks using notebooks

Utility	Description	Example
`fs`	Manipulates the Databricks filesystem (DBFS) from the console	`dbutils.fs.ls()`
`secrets`	Provides utilities for leveraging secrets within notebooks	`dbutils.secrets.get()`
`notebook`	Utilities for the control flow of a notebook	`dbutils.notebook.run()`
`widgets`	Methods to create and get bound value of input widgets inside notebooks	`dbutils.widget.text()`
`jobs`	Utilities for leveraging jobs features	`dbutils.jobs.taskValues.set()`

Available within Python, R, or Scala notebooks

©2023 Databricks Inc. — All rights reserved

Demo: Databricks Notebook Operations

©2023 Databricks Inc. — All rights reserved

databricks

What is Delta Lake?

©2023 Databricks Inc. — All rights reserved

Delta Lake is an open-source project that enables building a data lakehouse on top of existing storage systems

©2023 Databricks Inc. — All rights reserved

Delta Lake Is Not...

- Proprietary technology
- Storage format
- Storage medium
- Database service or data warehouse

©2023 Databricks Inc. — All rights reserved

Delta Lake Is...

- Open source
- Builds upon standard data formats
- Optimized for cloud object storage
- Built for scalable metadata handling

©2023 Databricks Inc. — All rights reserved

Delta Lake brings ACID to object storage

- **A**tomicity
- **C**onsistency
- **I**solation
- **D**urability

©2023 Databricks Inc. — All rights reserved

Problems solved by ACID

1. Hard to append data
2. Modification of existing data difficult
3. Jobs failing mid way
4. Real-time operations hard
5. Costly to keep historical data versions

©2023 Databricks Inc. — All rights reserved

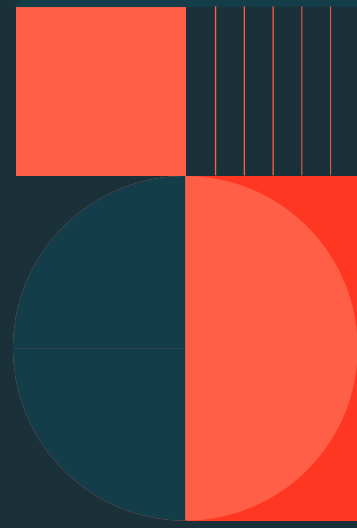
Delta Lake is the default for all tables created in Databricks

©2023 Databricks Inc. — All rights reserved

Demo: Version and Optimize Delta Tables

©2023 Databricks Inc. — All rights reserved

Demo: Load Data into Delta Lake



©2023 Databricks Inc. — All rights reserved

Questions?

©2023 Databricks Inc. — All rights reserved