

Introduction to Databricks Lakehouse

DATABRICKS CONCEPTS



Kevin Barlow
Data Analytics Practitioner



The Data Warehouse

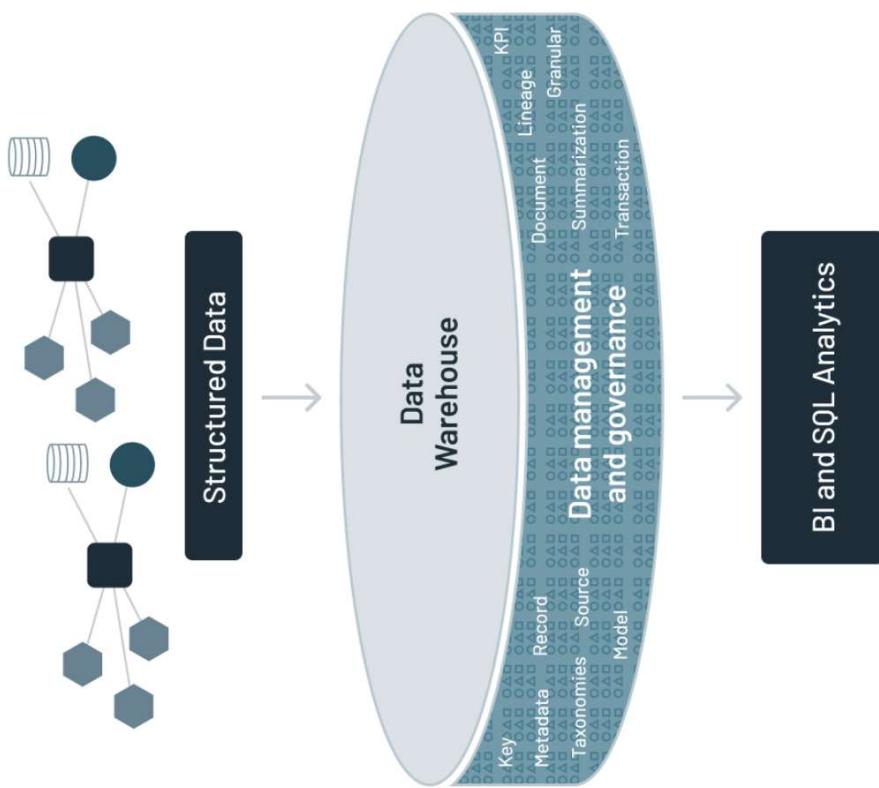
Data Warehouse

Pros

- Great for structured data
- Highly performant
- Easy to keep data clean

Cons

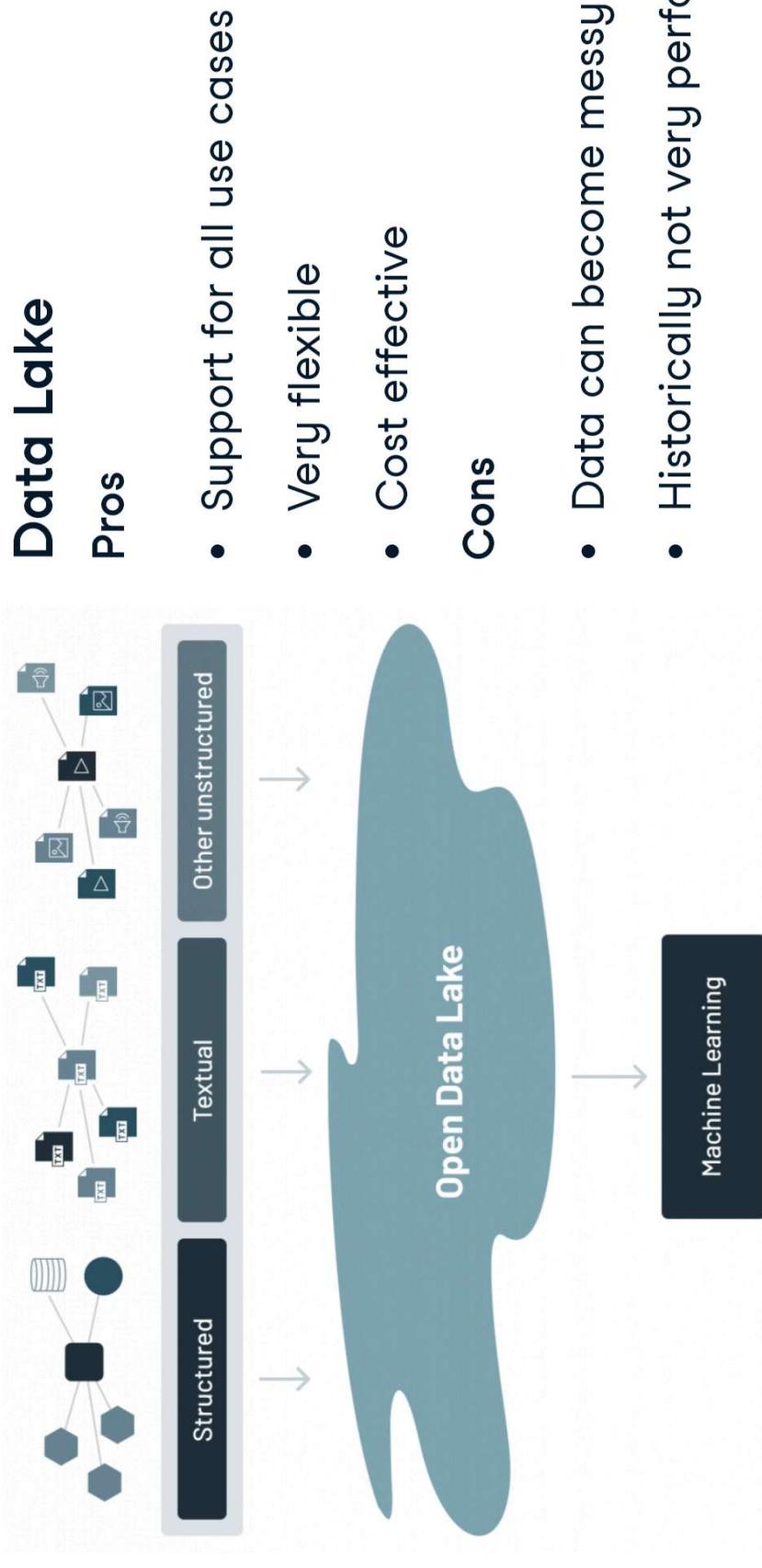
- Very expensive
- Cannot support modern applications
- Not built for Machine Learning



¹ <https://www.databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html>

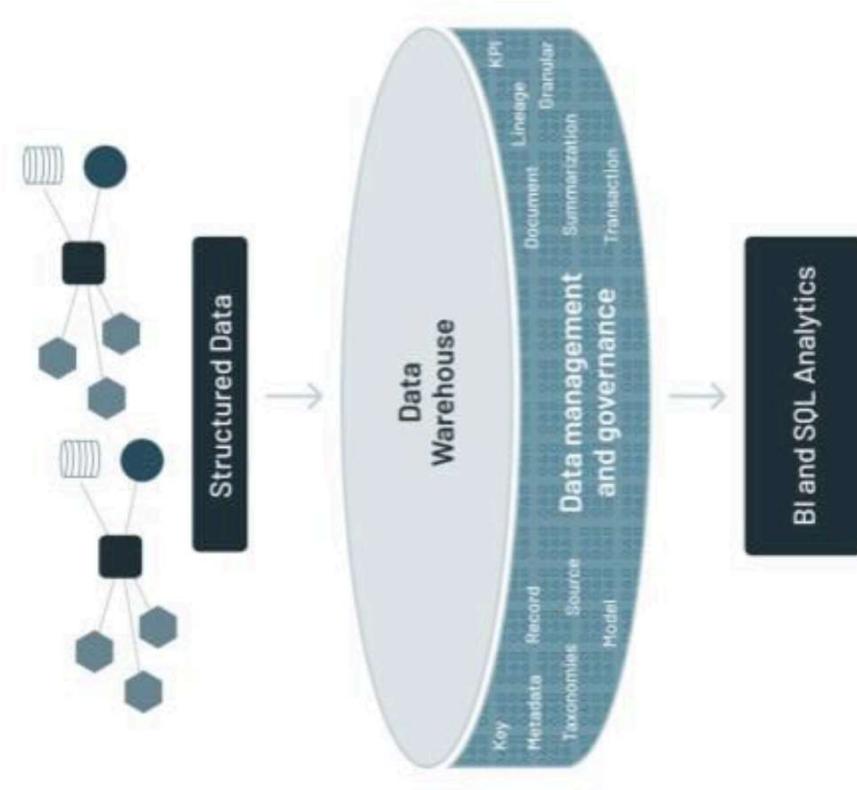


The Data Lake



¹ <https://www.databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html>

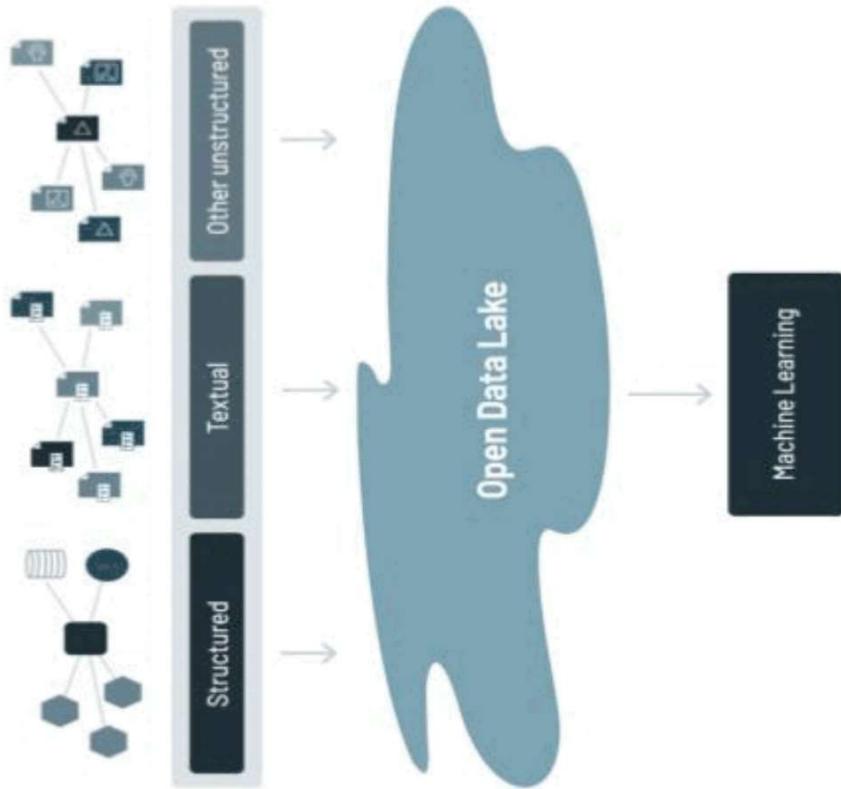
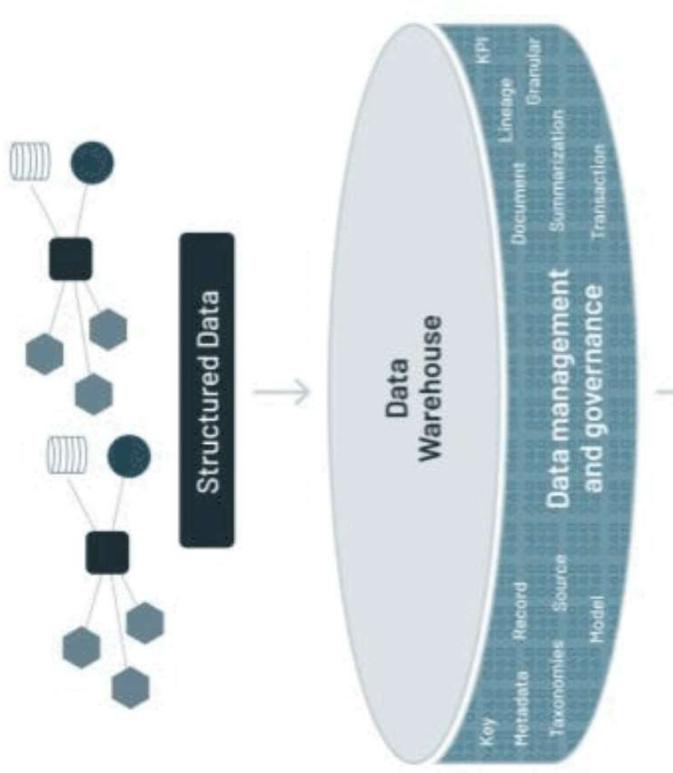
Birth of the Lakehouse



¹ <https://www.databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html>



Birth of the Lakehouse



¹ <https://www.databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html>

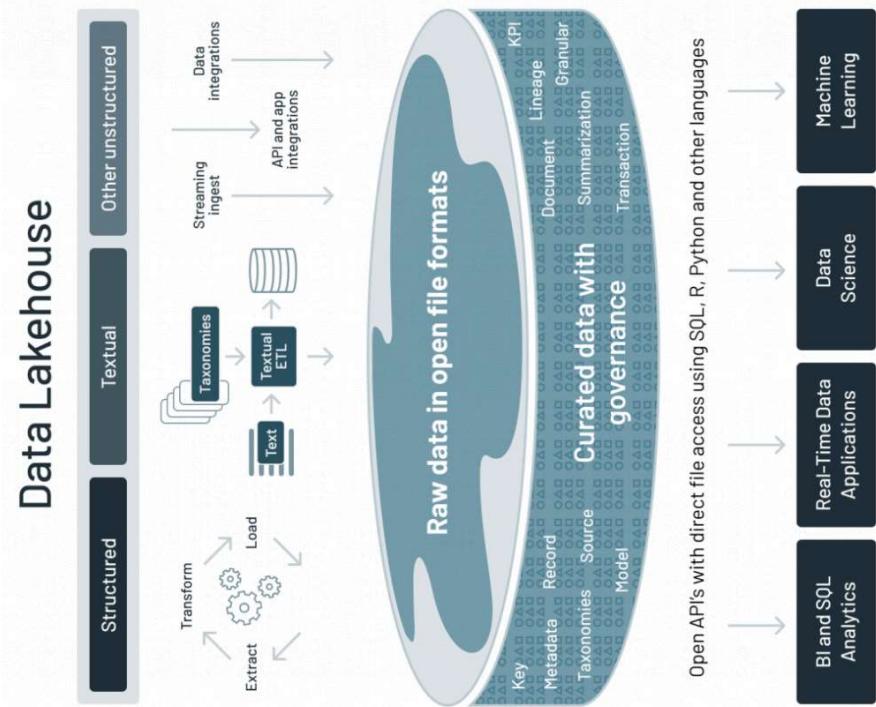


The Databricks Lakehouse

The Databricks Lakehouse Platform

- Single platform for all data workloads
- Built on open source technology
- Collaborative environment
- Simplified architecture

Data Lakehouse

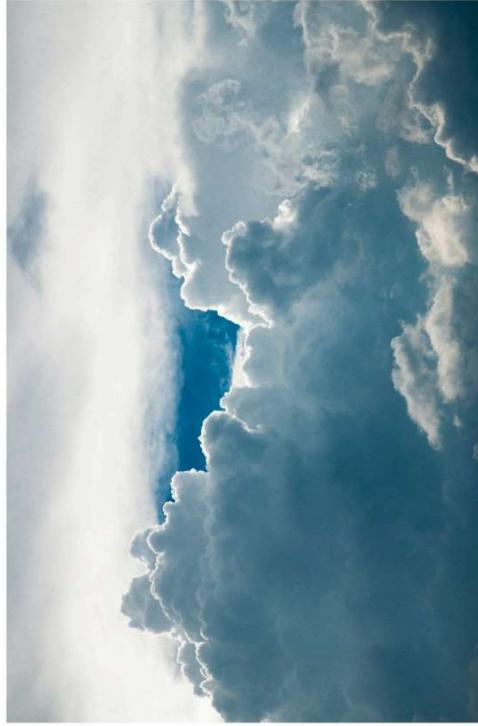


¹ <https://www.databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html>

Databricks Architecture Benefits

Unification

- Every use case from AI to BI
- Benefits of data warehouse and data lake
- Bring powerful platform to your data
- No lock-in to a specific cloud platform



Multi-Cloud

DataBricks Development Benefits

Collaborative

- Every data persona
- Ability to work in same platform in real-time

Open-Source

- Underpinned by Apache Spark
- Support for most popular languages (Python, R, Scala, SQL)

A dark, abstract background featuring several glowing, multi-colored text snippets. These snippets appear to be fragments of code or technical documentation, including words and phrases such as 'digital', 'post', 'global', 'query', 'script', 'while', and 'reviewer', all rendered in various colors like purple, blue, green, and yellow against a dark background.

Let's practice!

DATABRICKS CONCEPTS

Core features of the Databricks Lakehouse Platform

DATABRICKS CONCEPTS



Kevin Barlow
Data Practitioner



Apache Spark

Apache Spark is an open-source data processing framework and is the engine underneath Databricks.

DataCamp Courses

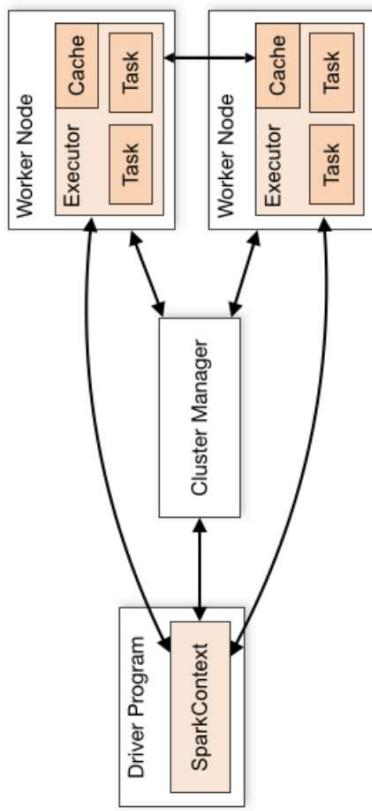
- Introduction to Pyspark
- Big Data Fundamentals with Pyspark
- Cleaning Data with Pyspark
- Machine Learning with Pyspark
- Introduction to Spark SQL in Python



Benefits of Spark

Key Benefits:

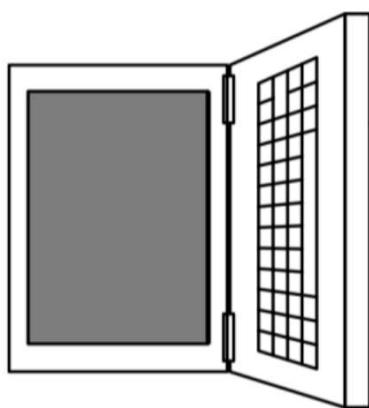
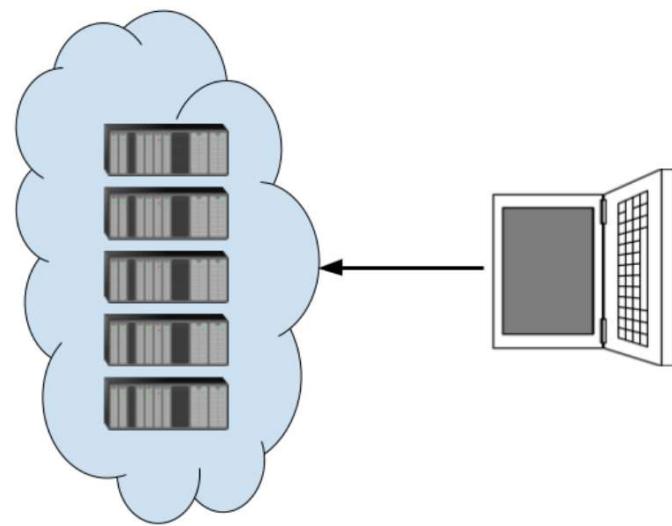
1. Extensible, flexible open-source framework
2. Large developer community
3. High performing
4. Databricks optimizations



¹ <https://spark.apache.org/docs/latest/cluster-overview.html>



Cloud computing basics



Databricks Compute

Clusters

- Collection of computational resources
- All workloads, any use case
- All-purpose vs. Jobs

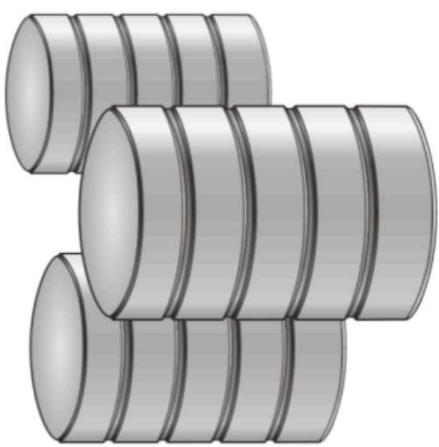


SQL Warehouses

- SQL only
- BI use cases
- Photon



Cloud data storage



Parquet
CSV
JSON

Delta



Delta is an open-source data storage file format, and provides:

- ACID transactions
- Unified batch and streaming
- Schema evolution
- Table history
- Time-travel

¹ delta.io

Unity Catalog

Unity Catalog is an open data governance strategy that controls access to all data assets in the Databricks Lakehouse platform.

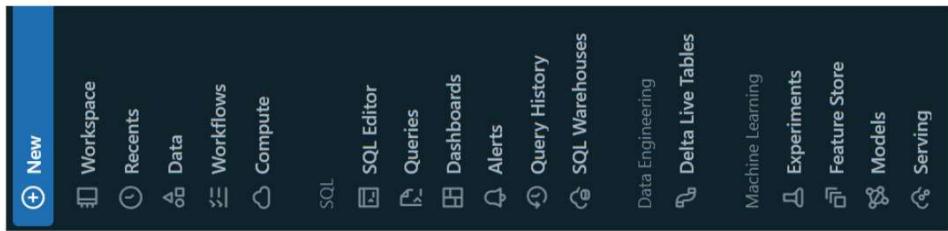
- SQL GRANT , REVOKE statements to control access
- Simple interface for governance



Databricks UI

Designed for easier access to capabilities based on your data workload.

- All users have access to data and compute
- SQL users get a familiar interface for queries and reports
- Data engineers leverage Delta Live Tables
- Machine Learning workloads use models, features, and more



Let's review!

DATABRICKS CONCEPTS

Administering a Databricks workspace

DATABRICKS CONCEPTS



Kevin Barlow
Data Practitioner



Account Admin

Key Responsibilities:

- Creating and managing workspaces
- Enabling Unity Catalog
- Managing identities
- Managing the account subscription



Account Console

Account console

Manage your Databricks account at scale



Workspaces

Configure workspace settings. Workspaces contain notebooks, libraries, queries, and workflows



Data

Manage metastores as your top-level container for data, catalogs, schemas (also called databases), views and tables



Users & groups
Manage identities for use with jobs, automated tools and systems



Settings

Configure your Databricks account user provisioning and other settings



<https://accounts.cloud.databricks.com/>



DATABRICKS CONCEPTS

Account Console - Workspaces

The screenshot shows the Databricks Account Console interface. At the top, there's a navigation bar with icons for Home, Account, Data, Workspaces, and Settings. Below the navigation bar, the title "Account console" is displayed, followed by the subtitle "Manage your Databricks account at scale". The main content area is divided into several sections:

- Workspaces**: This section is highlighted with a red border. It contains the text "Configure workspace settings. Workspaces contain notebooks, libraries, queries, and workflows".
- Data**: This section contains the text "Manage metastores as your top-level container for data, catalogs, schemas (also called databases), views and tables".
- Users & groups**: This section contains the text "Manage identities for use with jobs, automated tools and systems".
- Settings**: This section contains the text "Configure your Databricks account user provisioning and other settings".

<https://accounts.cloud.databricks.com/>



DATABRICKS CONCEPTS

Account Console - Data

The screenshot shows the Databricks Account Console interface. At the top, there's a navigation bar with icons for Home, Account console, Workspaces, Data, and Settings. Below the navigation bar, the main content area has a dark header "Account console" and a sub-header "Manage your Databricks account at scale". The page is divided into several sections:

- Data**: This section is highlighted with a red border. It contains an icon of a database, the word "Data", and a description: "Manage metastores as your top-level container for data, catalogs, schemas (also called databases), views and tables".
- Workspaces**: Contains an icon of a workspace, the word "Workspaces", and a description: "Configure workspace settings. Workspaces contain notebooks, libraries, queries, and workflows".
- Users & groups**: Contains an icon of two people, the word "Users & groups", and a description: "Manage identities for use with jobs, automated tools and systems".
- Settings**: Contains an icon of a gear, the word "Settings", and a description: "Configure your Databricks account user provisioning and other settings".

<https://accounts.cloud.databricks.com/>



DATABRICKS CONCEPTS

Account Console - Users & Groups

The screenshot shows the Databricks Account Console interface. At the top left is the 'Account console' header. Below it are several navigation icons: a gear for account settings, a triangle for workspace management, a square for data management, a double arrow for workspaces, and a gear for settings. The 'Workspaces' icon is currently selected, indicated by a pink background. To the right of these icons are three main sections:

- Users & groups**: A pink box highlights this section. It contains a user icon and a brief description: "Manage identities for use with jobs, automated tools and systems".
- Data**: An orange box contains a data icon and a brief description: "Manage metastores as your top-level container for data, catalogs, schemas (also called databases), views and tables".
- Settings**: A blue box contains a gear icon and a brief description: "Configure your Databricks account user provisioning and other settings".

<https://accounts.cloud.databricks.com/>



Account Console - Settings

The screenshot shows the Databricks Account Console with the 'Settings' page highlighted by a red border. The page has a dark header bar with icons for account, navigation, workspace, and settings. Below the header are four main sections:

- Account console**: Manage your Databricks account at scale.
- Workspaces**: Configure workspace settings. Workspaces contain notebooks, libraries, queries, and workflows.
- Data**: Manage metastores as your top-level container for data, catalogs, schemas (also called databases), views and tables.
- Users & groups**: Manage identities for use with jobs, automated tools and systems.

<https://accounts.cloud.databricks.com/>



Workspace Admin

Key Responsibilities:

- Managing identities in your workspace
- Creating and managing compute resources
- Managing workspace features and settings

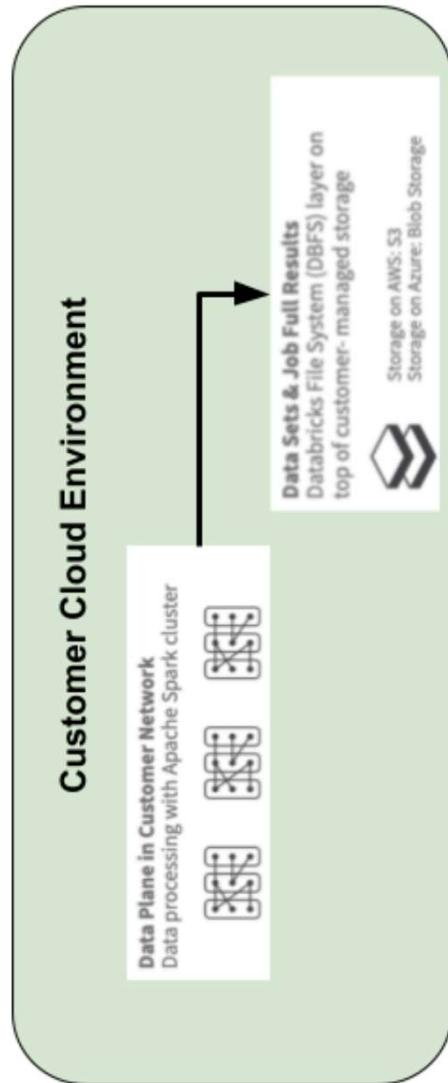
Admin Settings

Users	Service principals	Groups	Global init scripts	Workspace settings	SQL settings	Notification destinations	SQL warehouse settings
-------	--------------------	--------	---------------------	--------------------	--------------	---------------------------	------------------------

Data Plane

Contains all of the customer's assets needed for computation with Databricks.

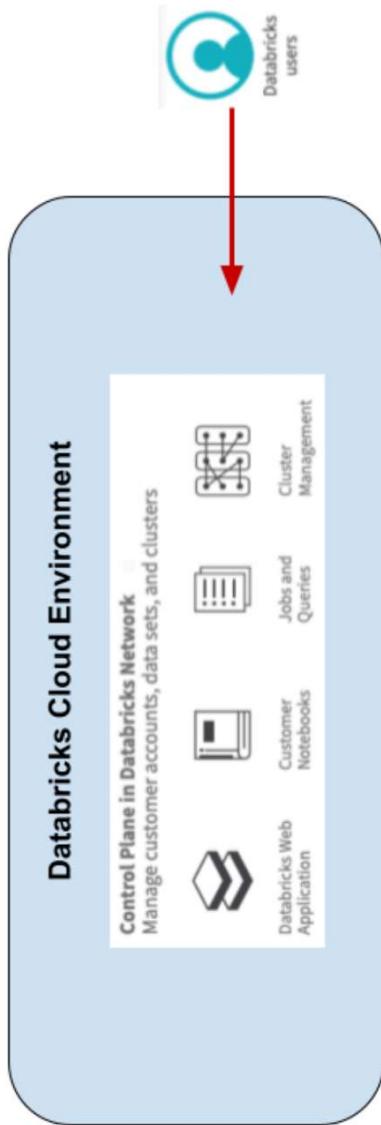
- Data is stored in the customer's cloud environment
- Clusters / SQL Warehouses run in customer's cloud tenant.



Control Plane

The portion of the platform that is managed and hosted by Databricks.

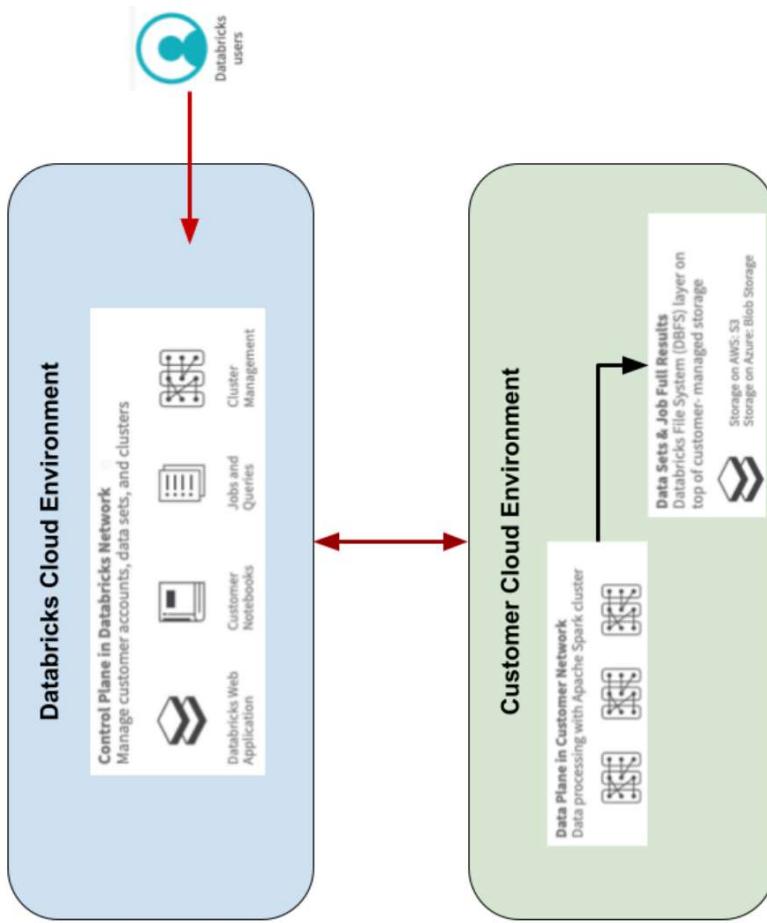
- Orchestrates various background tasks in Databricks
- Sends requests to Data Plane to create clusters, run jobs, etc.



Databricks Platform Architecture

Each cloud will have the same general options to create a workspace:

- Cloud Service Provider marketplace
- Account Console
- Using the Accounts API with Databricks
- Programmatic deployment (e.g., Terraform)



¹ <https://docs.databricks.com/getting-started/overview.html>

Let's review!

DATABRICKS CONCEPTS

Setting up a Databricks workspace example

DATABRICKS CONCEPTS



Kevin Barlow
Data Practitioner



Let's practice!

DATABRICKS CONCEPTS