

Introduction to Databricks Lakehouse

INTRODUCTION TO DATABRICKS



Kevin Barlow

Data Analytics Practitioner

The Data Warehouse

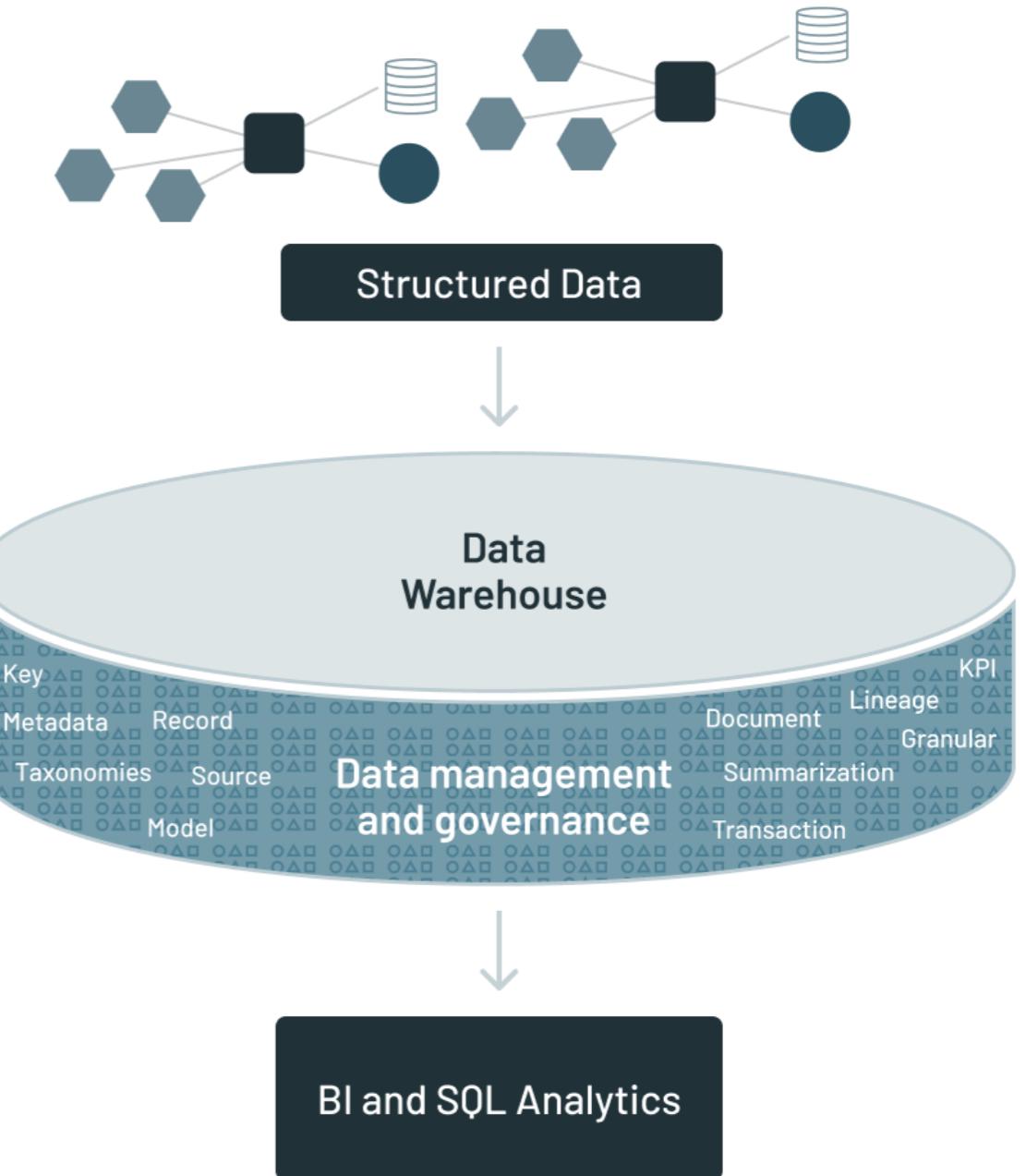
Data Warehouse

Pros

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

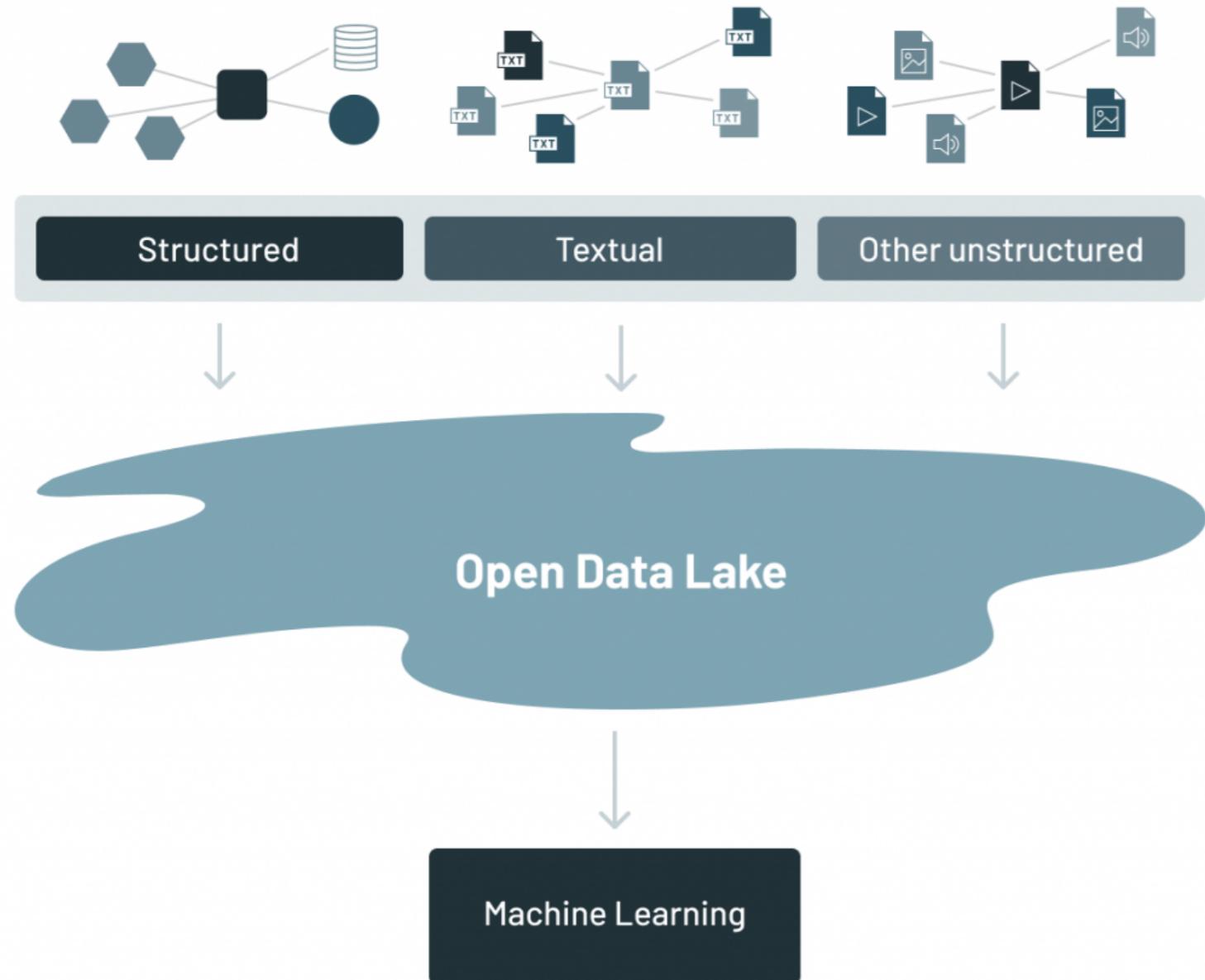
Cons

- 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



Data Lake

Pros

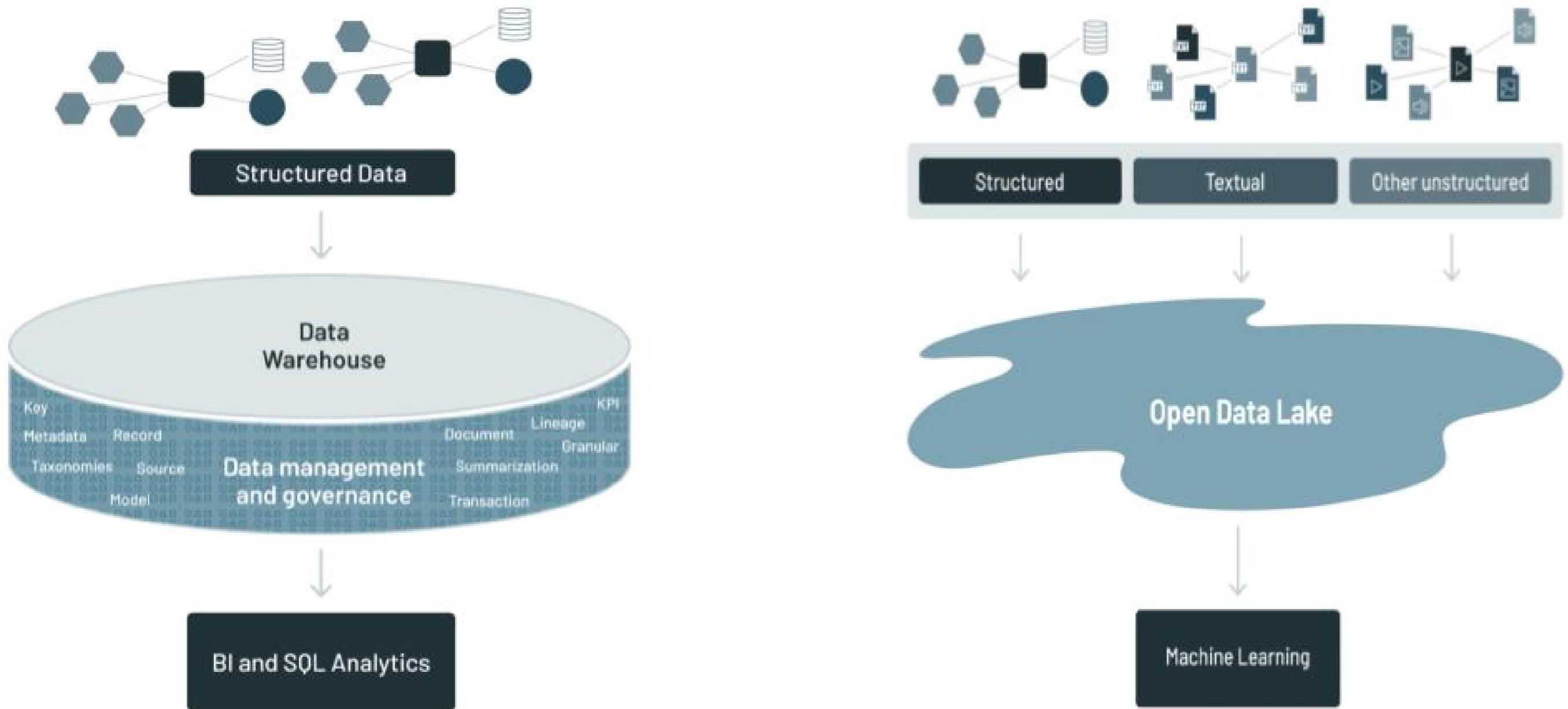
- Support for all use cases
- Very flexible
- Cost effective

Cons

- Data can become messy
- Historically not very performant

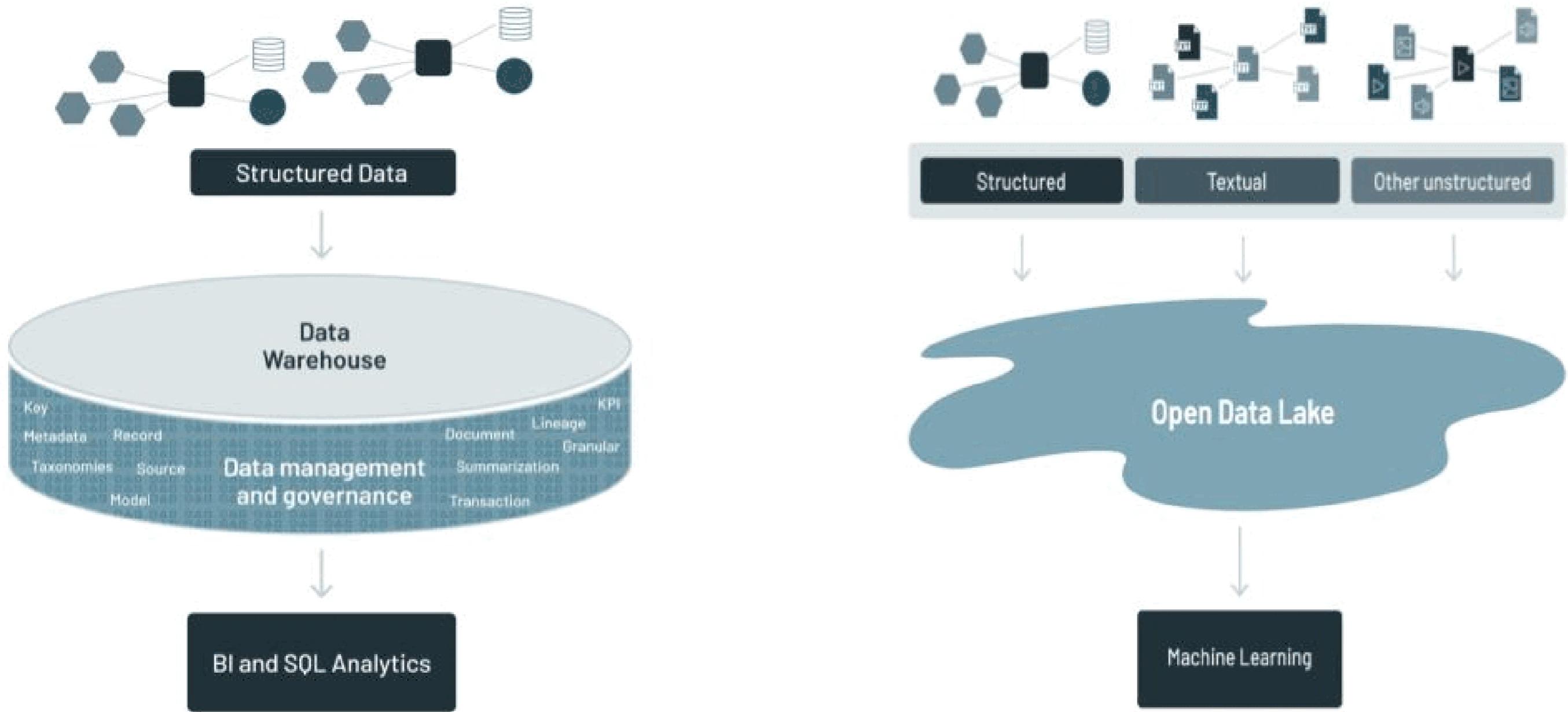
¹ <https://www.databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html>

Birth of the Lakehouse



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

Birth of the Lakehouse



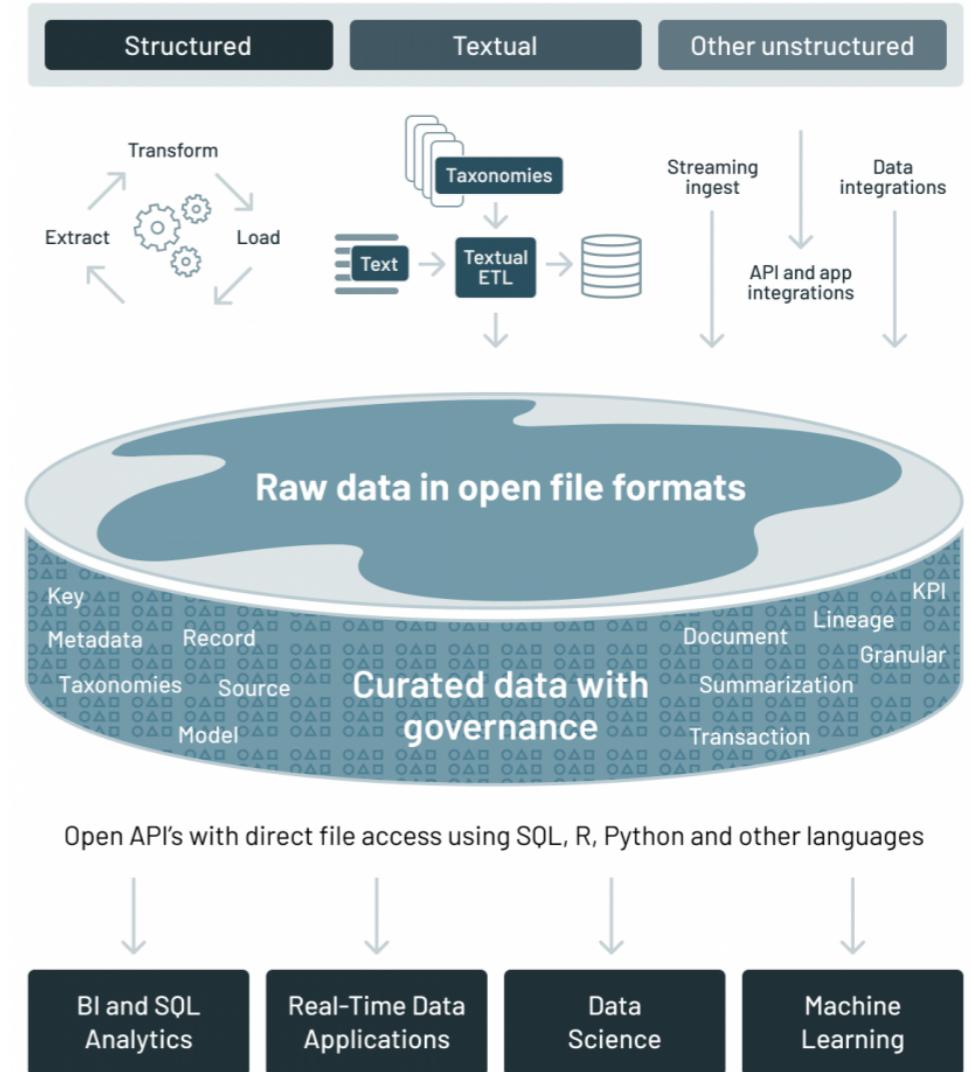
¹ <https://www.databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html>

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



Multi-Cloud

- Bring powerful platform to your data
- No lock-in to a specific cloud platform



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 blurred screenshot of a computer monitor displaying a large amount of code. The code is written in a programming language, likely Python, and includes various functions and variables. The text is in a monospaced font with color coding for different elements like keywords and comments. The background of the code area is dark, and the overall image is slightly out of focus.

Let's practice!

INTRODUCTION TO DATABRICKS

Core features of the Databricks Lakehouse Platform

INTRODUCTION TO DATABRICKS



Kevin Barlow
Data Practitioner

Apache Spark

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

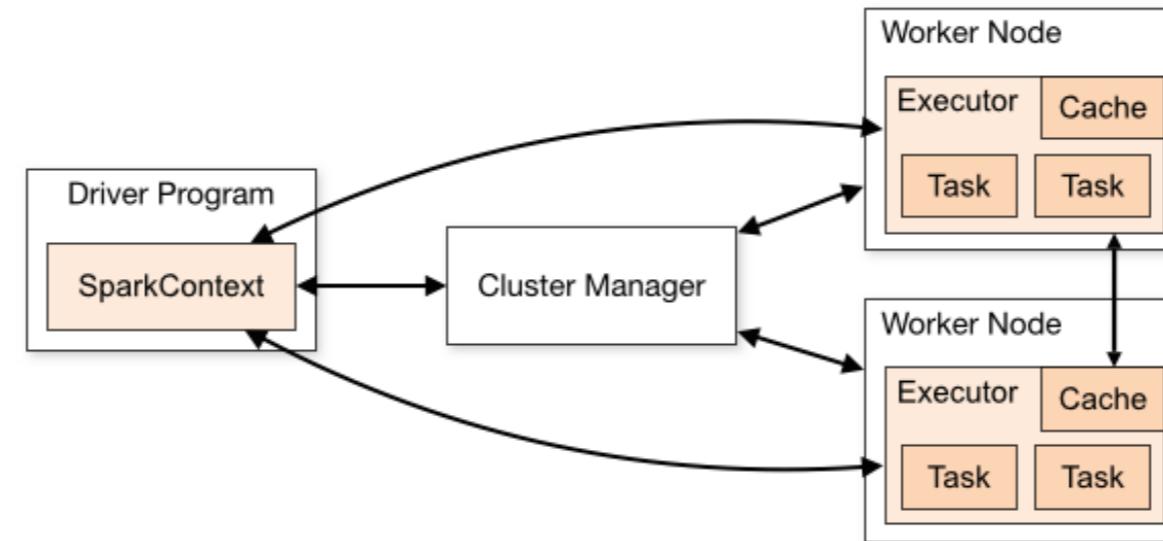
DataCamp Courses

- 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

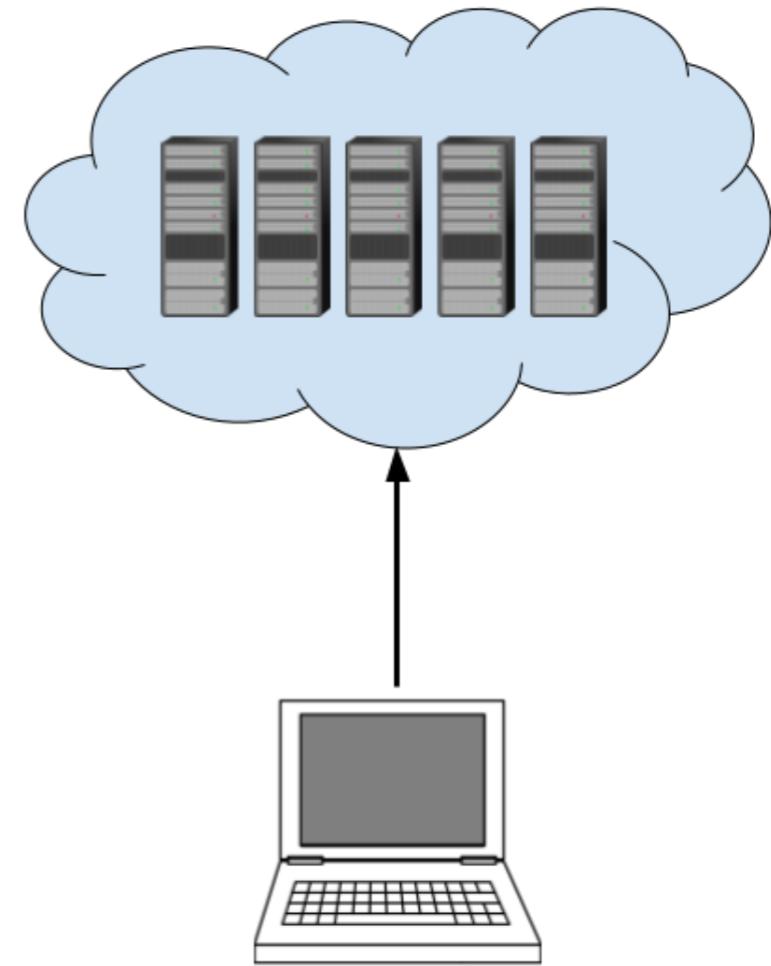
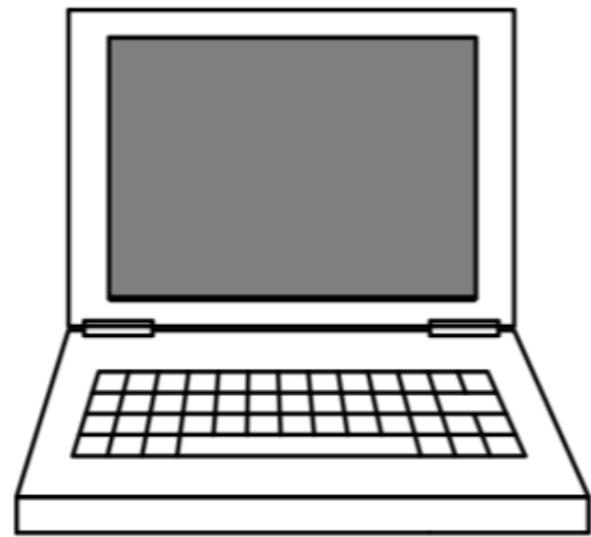
Key Benefits:

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.

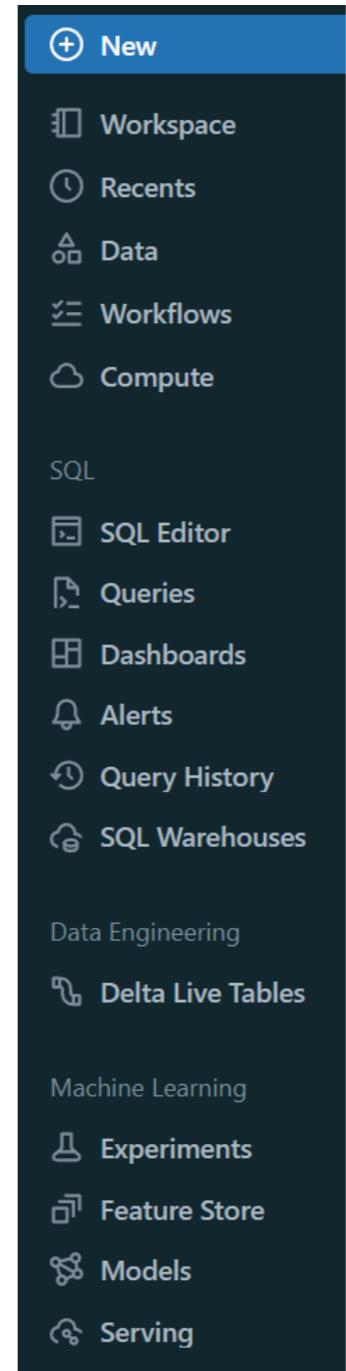
- SQL GRANT , REVOKE statements to control access
- Simple interface for governance



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!

INTRODUCTION TO DATABRICKS

Administering a Databricks workspace

INTRODUCTION TO DATABRICKS



Kevin Barlow
Data Practitioner

Account Admin

Key Responsibilities:

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



Account Console

The screenshot shows the Databricks Account Console. On the left is a vertical sidebar with icons for account management, workspace settings, data, users/groups, and settings. The main area has four cards:

- Workspaces**: Manage 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/>

Account Console - Workspaces

The screenshot shows the Databricks Account Console interface. On the left is a dark sidebar with four icons: a gear (Account), a bar chart (Metrics), a gear and bar chart (Metrics), and a gear (Settings). The main area has a header "Account console" and a subtitle "Manage your Databricks account at scale". Below are four sections: "Workspaces" (highlighted with a red box), "Data", "Users & groups", and "Settings". Each section has an icon and a brief description.

- 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/>

Account Console - Data

The screenshot shows the Databricks Account Console interface. On the left is a vertical sidebar with four icons: a grid (Account console), a bar chart (Metrics), a gear (Settings), and a gear with a checkmark (Integrations). The main area has a header "Account console" and a subtitle "Manage your Databricks account at scale". Below are four cards:

- Workspaces**: Manage 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. This card is highlighted with a red border.
- 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/>

Account Console - Users & Groups

The screenshot shows the Databricks Account Console interface. On the left is a vertical sidebar with icons for Home, Workspaces, Data, Settings, and Help. The main area has a title 'Account console' and a subtitle 'Manage your Databricks account at scale'. It features four main sections: 'Workspaces' (icon of a workspace), 'Data' (icon of a database), 'Users & groups' (icon of two people), and 'Settings' (icon of a gear). The 'Users & groups' section is highlighted with a red border. Each section contains a brief description.

- 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/>

Account Console - Settings

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/>

Workspace Admin

Key Responsibilities:

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

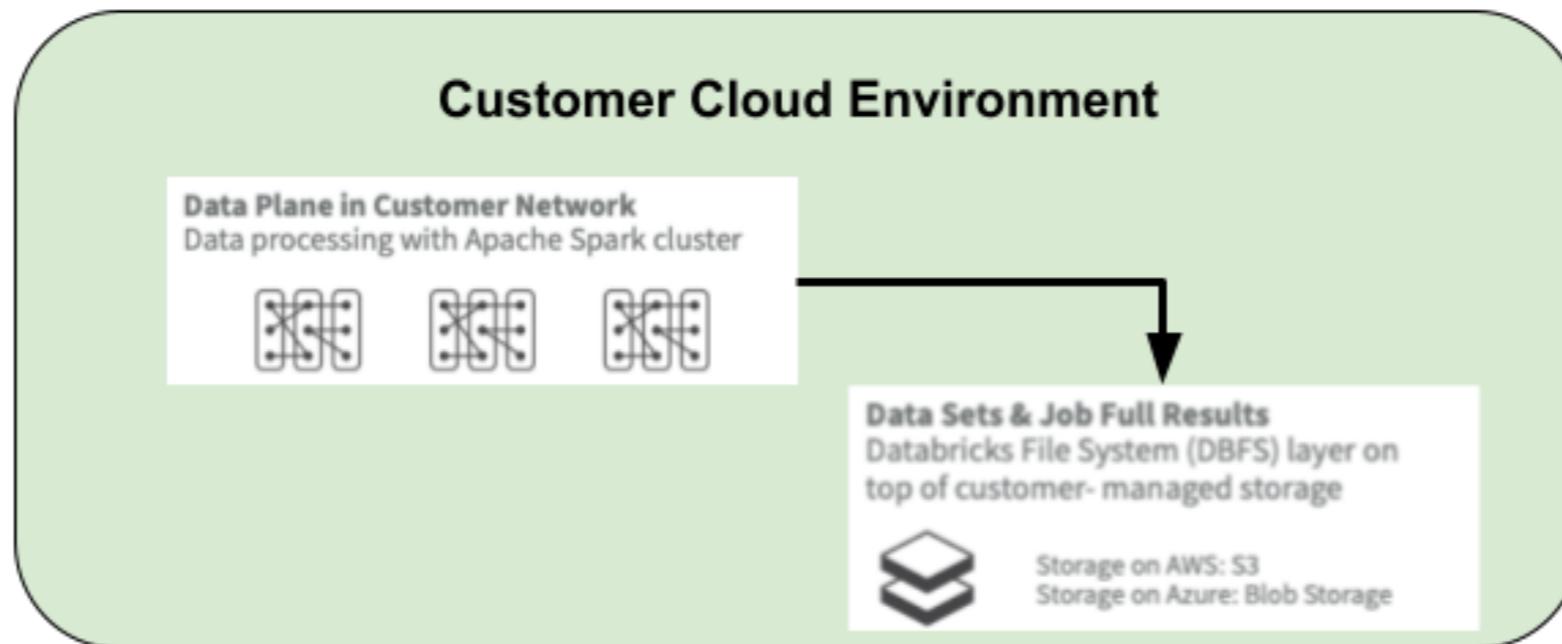
Admin Settings

[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.

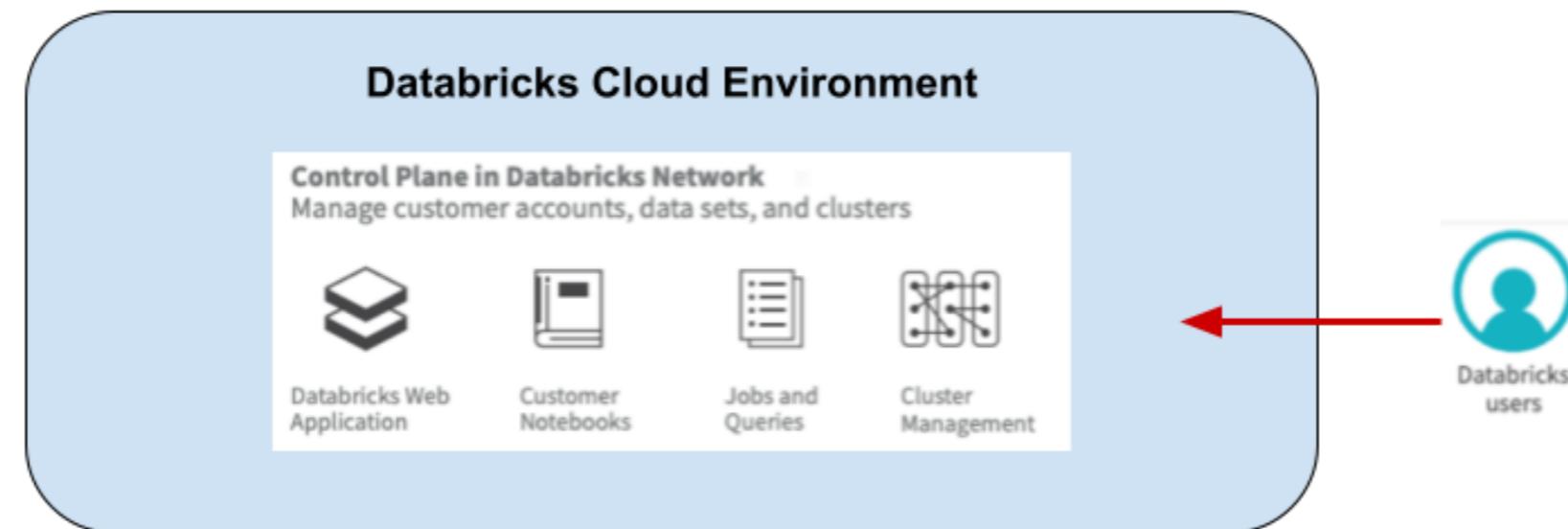
- Data is stored in the customer's cloud environment
- Clusters / SQL Warehouses run in customer's cloud tenant.



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)



Databricks Platform Architecture

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

Let's review!

INTRODUCTION TO DATABRICKS

Setting up a Databricks workspace example

INTRODUCTION TO DATABRICKS



Kevin Barlow
Data Practitioner

Let's practice!

INTRODUCTION TO DATABRICKS