databricks

Get Started with Databricks Data Science & Engineering Workspace

Module Agenda

Get Started with Databricks Data Science and Engineering Workspace

Introduction to the Databricks Lakehouse Platform

Databricks Architecture and Services

Demo - Navigating the Workspace

DE 1.1 - Create and Manage Clusters Interactively

DE 1.2 - Notebook Basics

Git Versioning with Databricks Repos

Demo - Using Databricks Repos

DE 1.3L - Getting Started with the Databricks Lakehouse Platform Lab

Using the Databricks Lakehouse Platform

Learning Objectives

- Describe the components of the Databricks Lakehouse
- Complete basic code development tasks using services of the Databricks Data Science and Engineering Workspace
- Perform common table operations using Delta Lake in the Lakehouse

Using the Databricks Lakehouse Platform

Agenda

- Introduction to the Databricks Lakehouse Platform
- Introduction to the Databricks Workspace and Services
  - Using clusters, files, notebooks, and repos
- Introduction to Delta Lake
  - Manipulating and optimizing data in Delta tables

databricks

The Databricks Lakehouse Platform

Lakehouse

One simple platform to unify all of your data, analytics, and AI workloads

Customers

7000+

across the globe

Original creators of:

Apache Spark™ · Delta Lake · mlflow™

Supporting enterprises in every industry

Healthcare & Life Sciences

Humana · gsk · AMGEN · OPTUM · NHS · Biogen · REGENERON · AstraZeneca

Manufacturing & Automotive

Schneider Electric · Rolls-Royce · DAIMLER · COLAS · ZEISS · thyssenkrupp · BECHTEL

Media & Entertainment

CONDÉ NAST · RIOT GAMES · COMCAST · SHOWTIME · viacom · PUBLICIS GROUPE · zynga

Financial Services

HSBC · BNP PARIBAS · Nasdaq · Nationwide · S&P Global · CREDIT SUISSE · ABN·AMRO

Public Sector

cfpb · CMS · U.S. Citizenship and Immigration Services · LOS ANGELES FIRE DEPARTMENT · New York Power Authority

Retail & CPG

sam's club · CVSHealth · Foot Locker · 7-ELEVEN · Henkel · MARS · H&M

Energy & Utilities

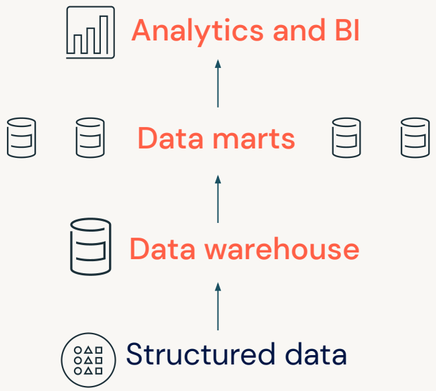
TOTAL · devon · ExxonMobil · aggreko · Quby

Digital Native

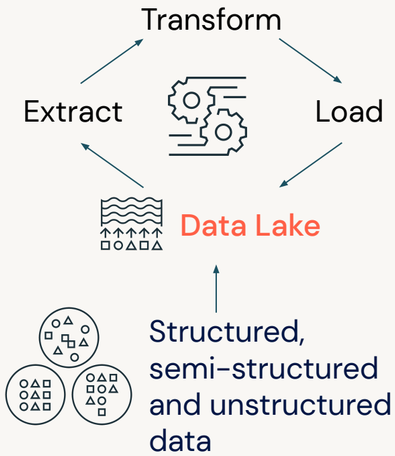
gousto · DOLLAR SHAVE CLUB · Grab · zalando · BUTCHER BOX · wehkamp · SCRIBD

Most enterprises struggle with data

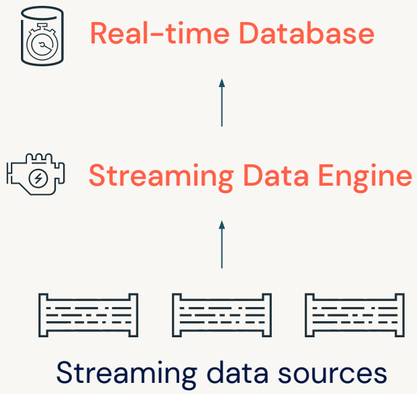
Data Warehousing



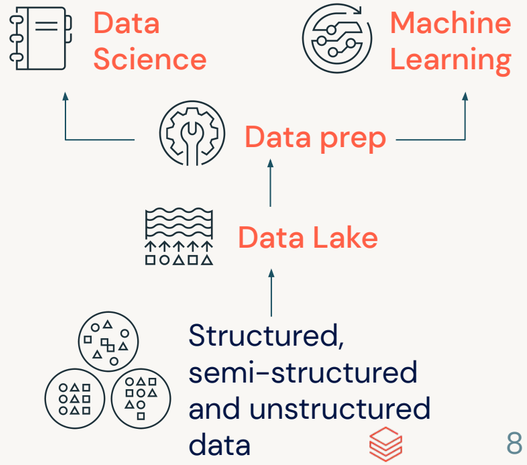
Data Engineering



Streaming



Data Science and ML



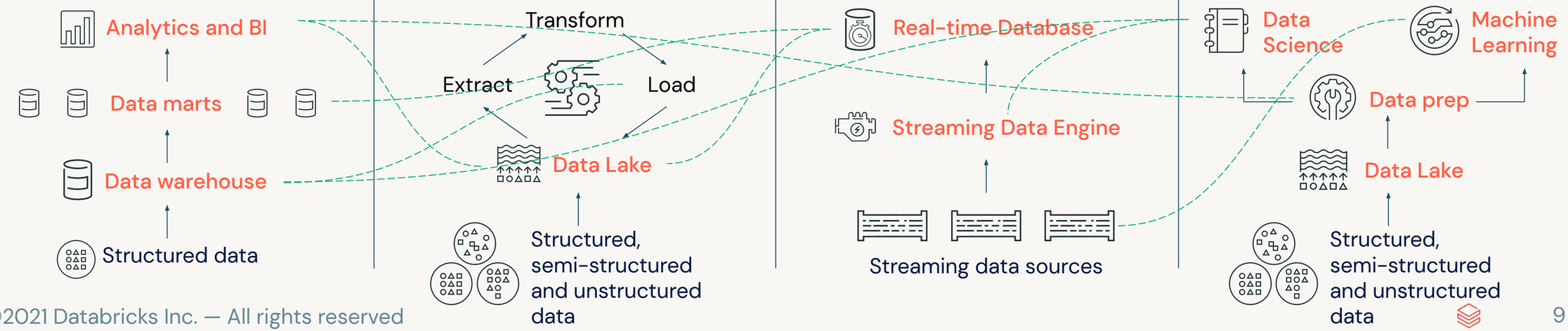
Siloed stacks increase data architecture complexity

Most enterprises struggle with data

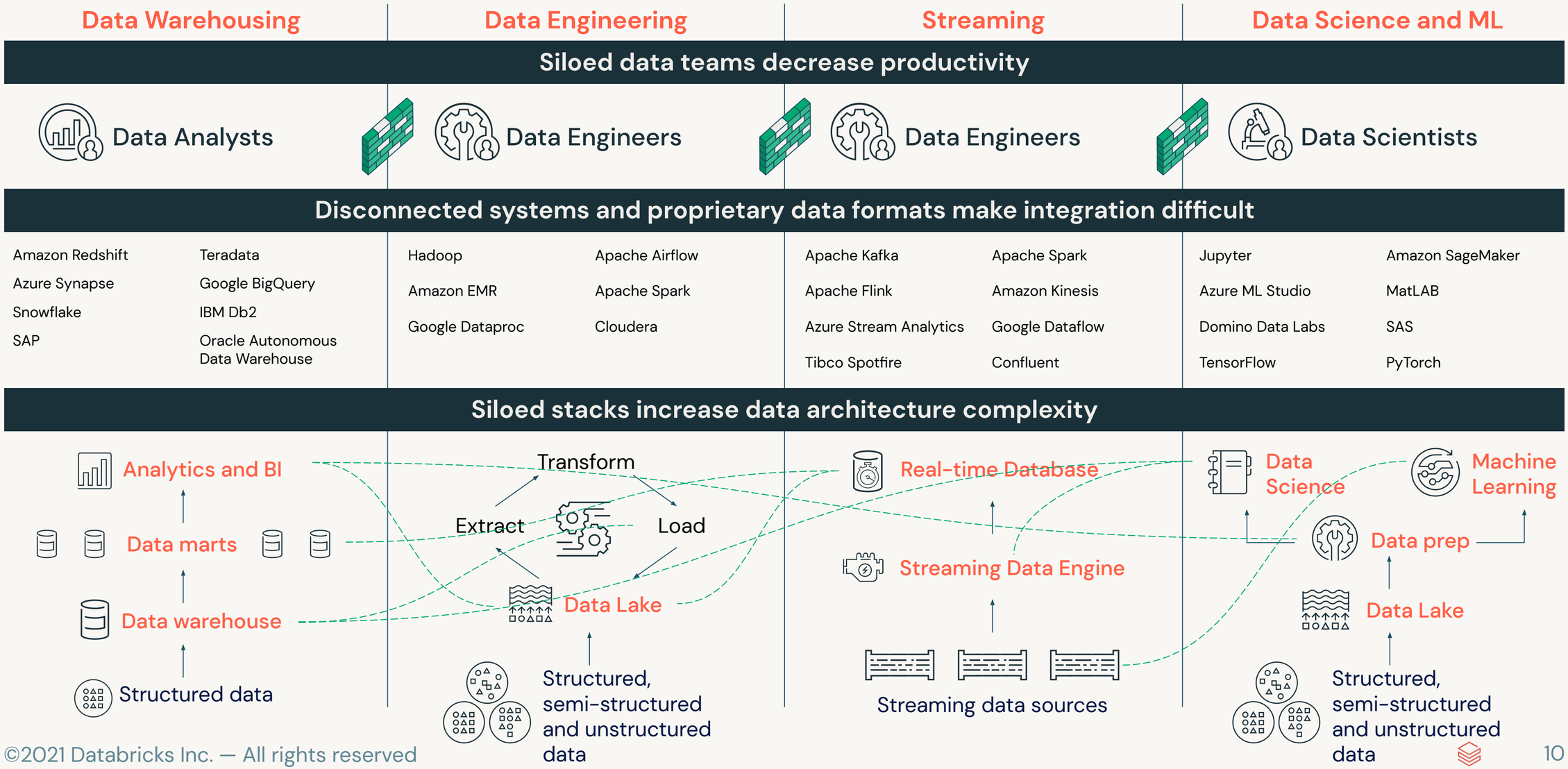
Data Warehousing		Data Engineering		Streaming		Data Science and ML	
Amazon Redshift	Teradata	Hadoop	Apache Airflow	Apache Kafka	Apache Spark	Jupyter	Amazon SageMaker
Azure Synapse	Google BigQuery	Amazon EMR	Apache Spark	Apache Flink	Amazon Kinesis	Azure ML Studio	MatLAB
Snowflake	IBM Db2	Google Dataproc	Cloudera	Azure Stream Analytics	Google Dataflow	Domino Data Labs	SAS
SAP	Oracle Autonomous Data Warehouse			Tibco Spotfire	Confluent	TensorFlow	PyTorch

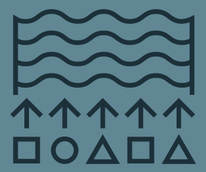
Disconnected systems and proprietary data formats make integration difficult

Siloed stacks increase data architecture complexity



Most enterprises struggle with data



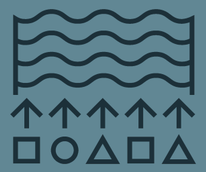


Data Lake

Lakehouse

One platform to unify all of your data, analytics, and AI workloads

Data Warehouse



Data Lake

DELTA LAKE

An open approach to bringing **data management and governance** to data lakes

Better reliability with transactions

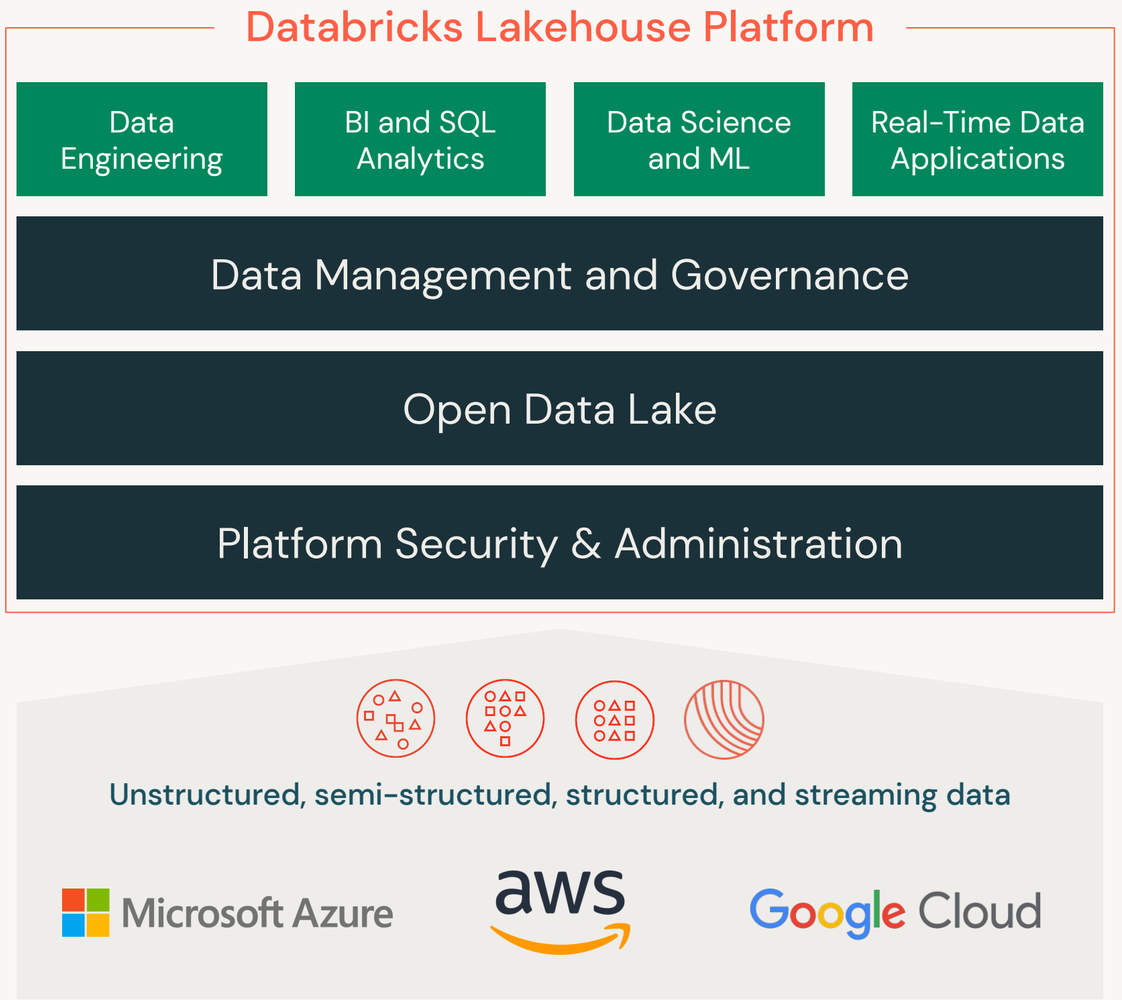
48x faster data processing with indexing

Data governance at scale with fine-grained access control lists

Data Warehouse

The Databricks Lakehouse Platform

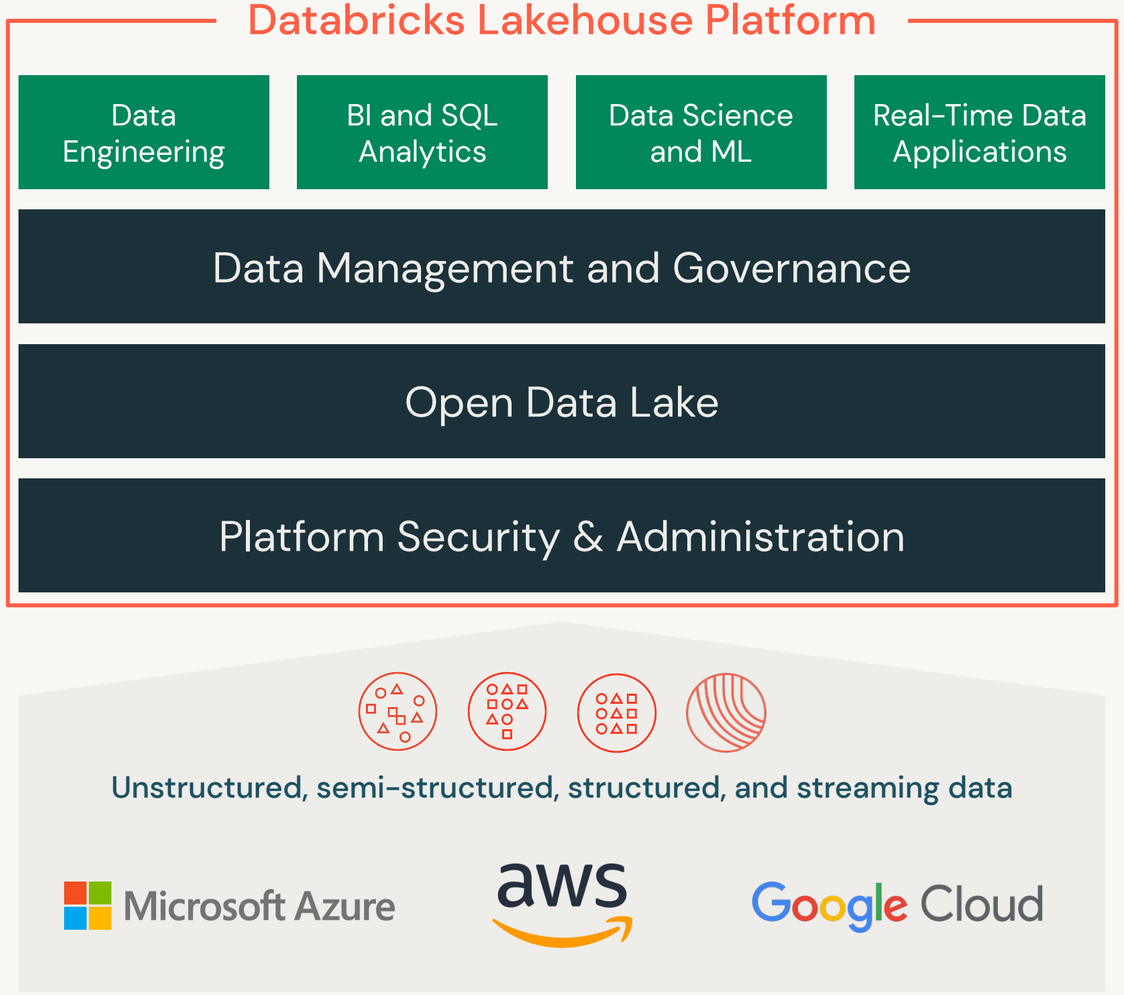
- Simple
- Open
- Collaborative



The Databricks Lakehouse Platform

Simple

Unify your data, analytics, and AI on one common platform for all data use cases



The Databricks Lakehouse Platform

Open

Unify your data ecosystem with open source standards and formats.

Built on the innovation of some of the most successful open source data projects in the world

30 Million+

Monthly downloads

The Databricks Lakehouse Platform

Open

Unify your data ecosystem with open source standards and formats.

450+

Partners across the data landscape

Visual ETL & Data Ingestion

Azure Data Factory, Fivetran, CONFLUENT, Informatica, talend, Qlik, StreamSets, precisely, Infoworks, wanDISCO, TRIFACTA, dbt, mongoDB., MATILLION

Business Intelligence

+ableau, Power BI, Azure Synapse, Qlik, plotly, looker, Google BigQuery, ThoughtSpot, TIBCO Spotfire, Amazon Redshift

Machine Learning

Amazon SageMaker, Azure Machine Learning, MathWorks, Google AI Platform, Studio, H2O.ai

databricks
Lakehouse Platform
Microsoft Azure, aws, Google Cloud

Data Providers

dun & bradstreet, SAFEGRAPH, experian, yipitDATA, arm INSIGHT

Centralized Governance

AWS Glue, Informatica, Alation, PRIVACERA, IMMUTA, collibra

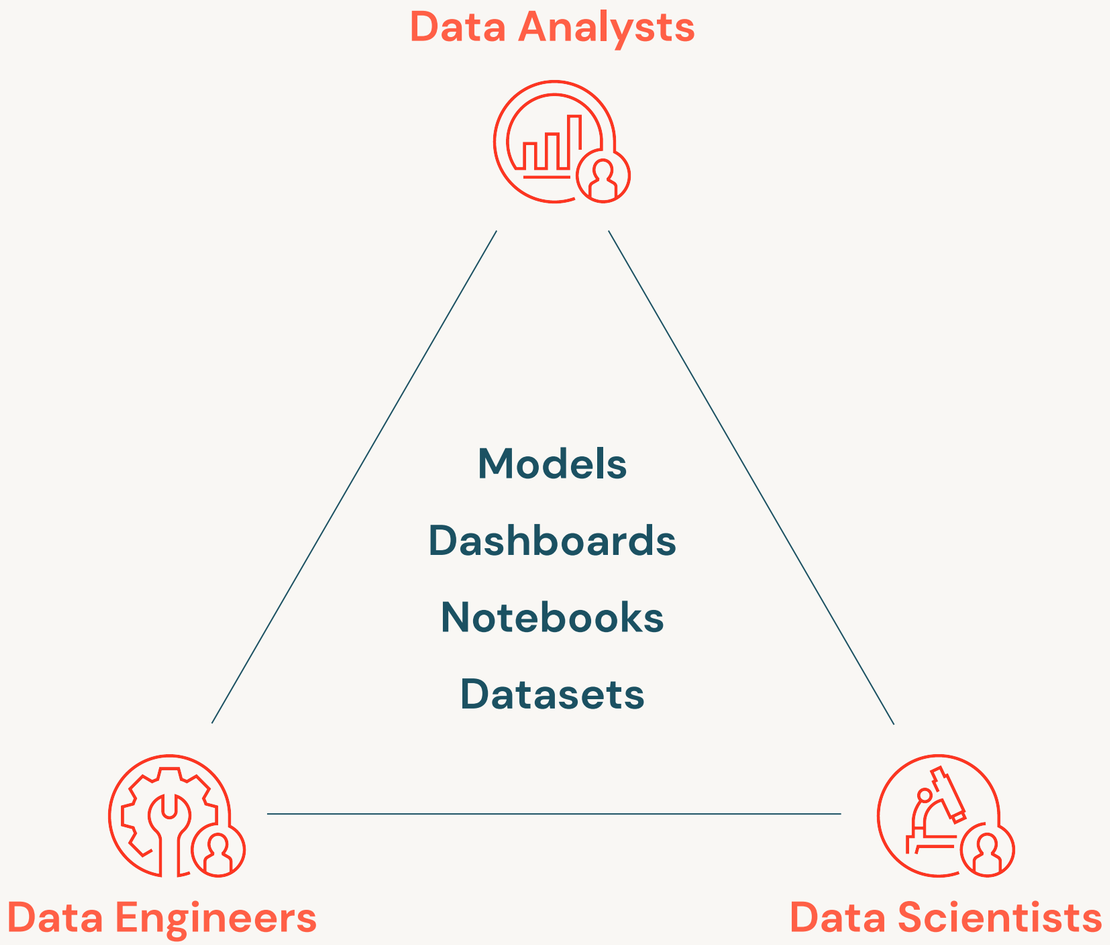
Top Consulting & SI Partners

Insight, NEUDESIC, PARIVEDA, accenture, avanade, wipro, Infosys, Tech Mahindra, GO DATA DRIVEN, ZEAL, AGILISIUM, BLUEGRANITE, Hitachi Solutions, ONICA, Capgemini, Cognizant, Deloitte., ELASTACLOUD, OPENVALUE, b.telligent, kumulus, BlueShift, deep_dive, Mindtree, slalom, TATA CONSULTANCY SERVICES, JUST ANALYTICS, Intellify, Versor

The Databricks Lakehouse Platform

Collaborative

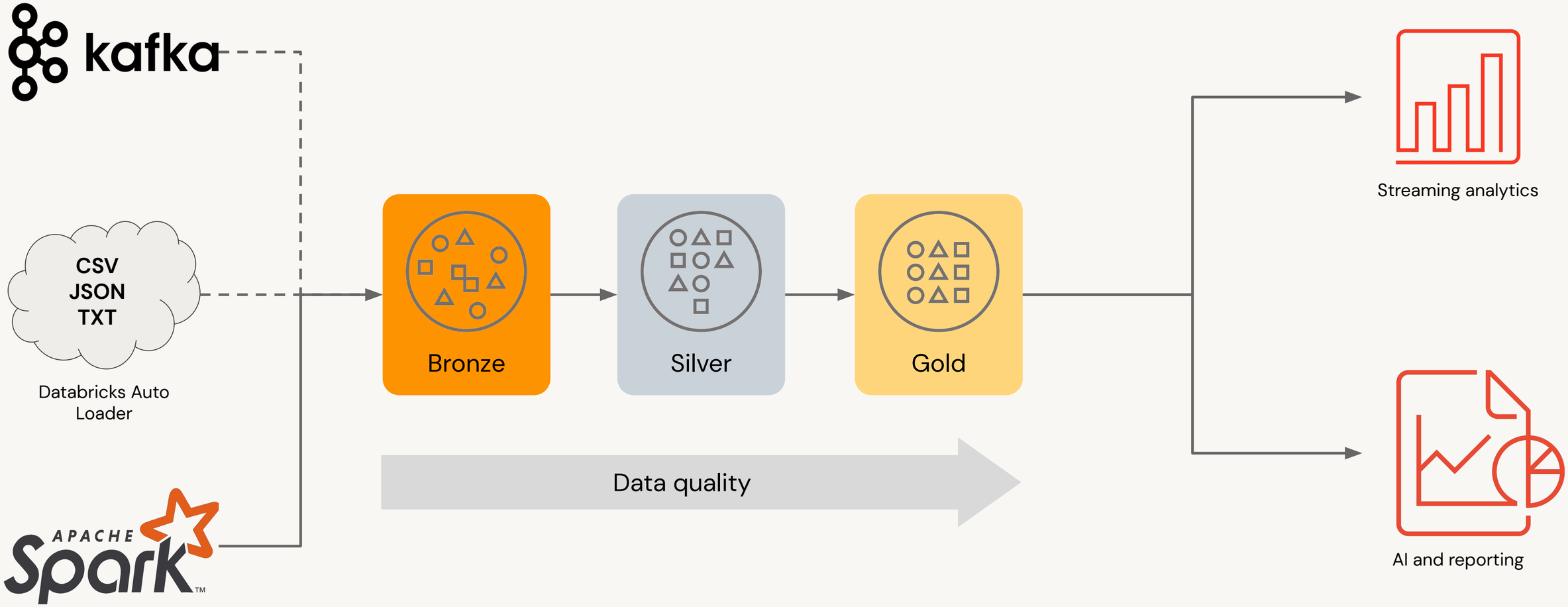
Unify your data teams to collaborate across the entire data and AI workflow



databricks

The Medallion Architecture

Multi-Hop in the Lakehouse



Multi-Hop in the Lakehouse

Bronze Layer

Typically just a raw copy of ingested data

Replaces traditional data lake

Provides efficient storage and querying of full, unprocessed history of data

Multi-Hop in the Lakehouse

Silver Layer

Reduces data storage complexity, latency, and redundancy

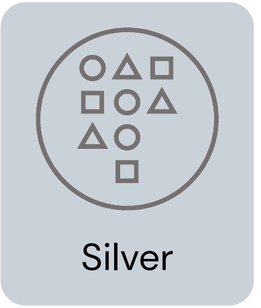
Optimizes ETL throughput and analytic query performance

Preserves grain of original data (without aggregations)

Eliminates duplicate records

Production schema enforced

Data quality checks, corrupt data quarantined



Multi-Hop in the Lakehouse

Gold Layer

Powers ML applications, reporting, dashboards, ad hoc analytics

Refined views of data, typically with aggregations

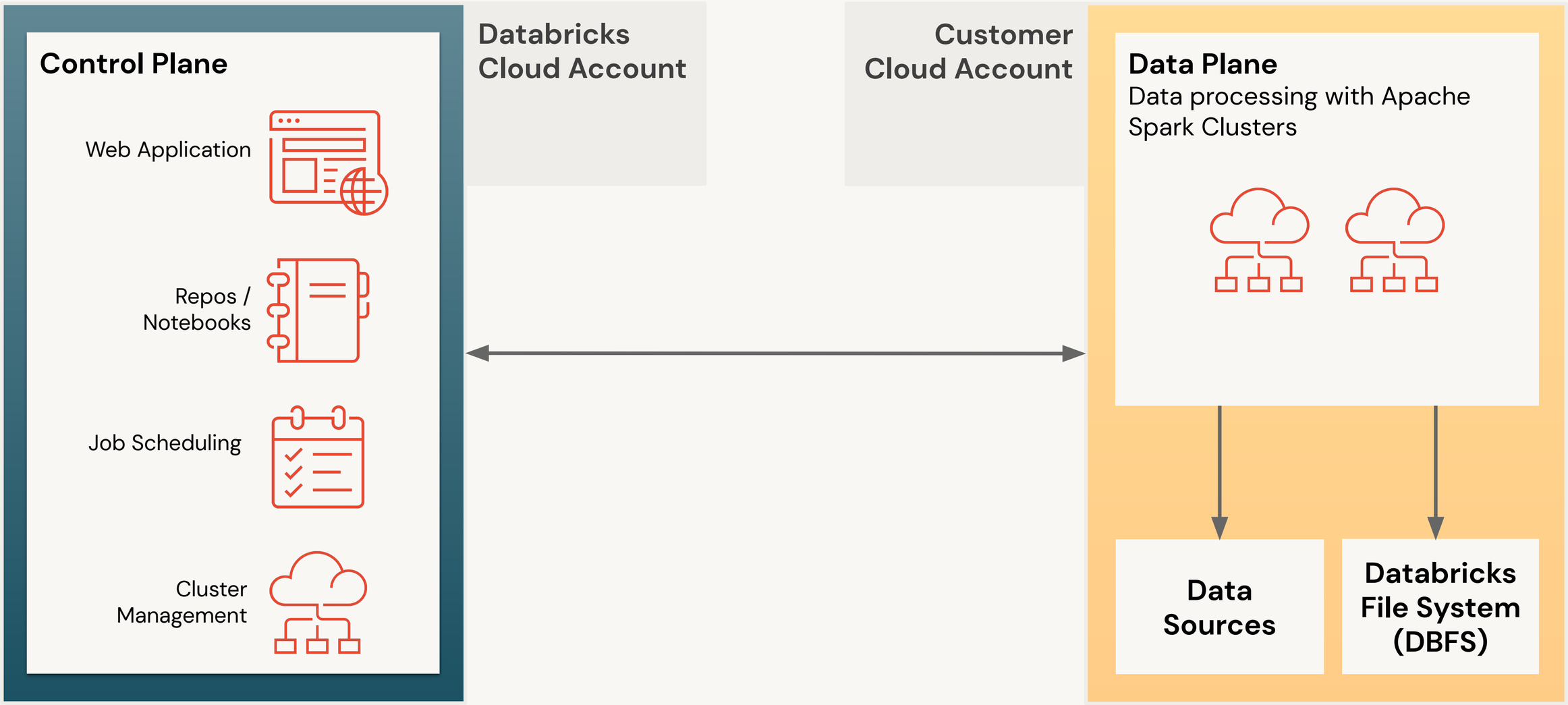
Reduces strain on production systems

Optimizes query performance for business-critical data

databricks

Databricks Architecture and Services

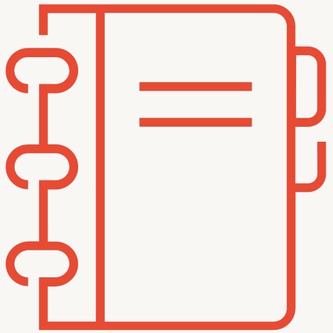
Databricks Architecture



Databricks Services

Control Plane in Databricks
Manage customer accounts, datasets, and clusters

Databricks Web Application

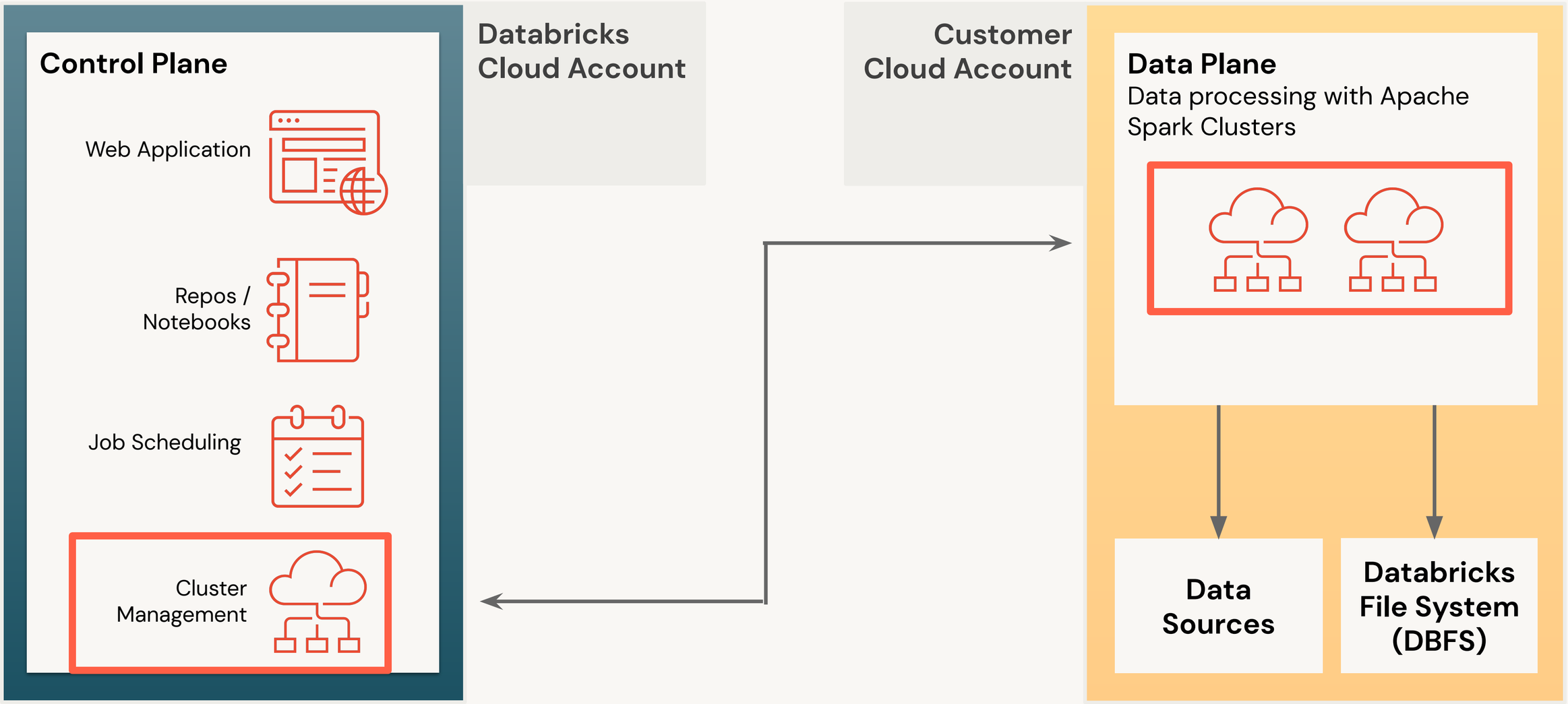


Repos / Notebooks

Jobs

Cluster Management

Clusters



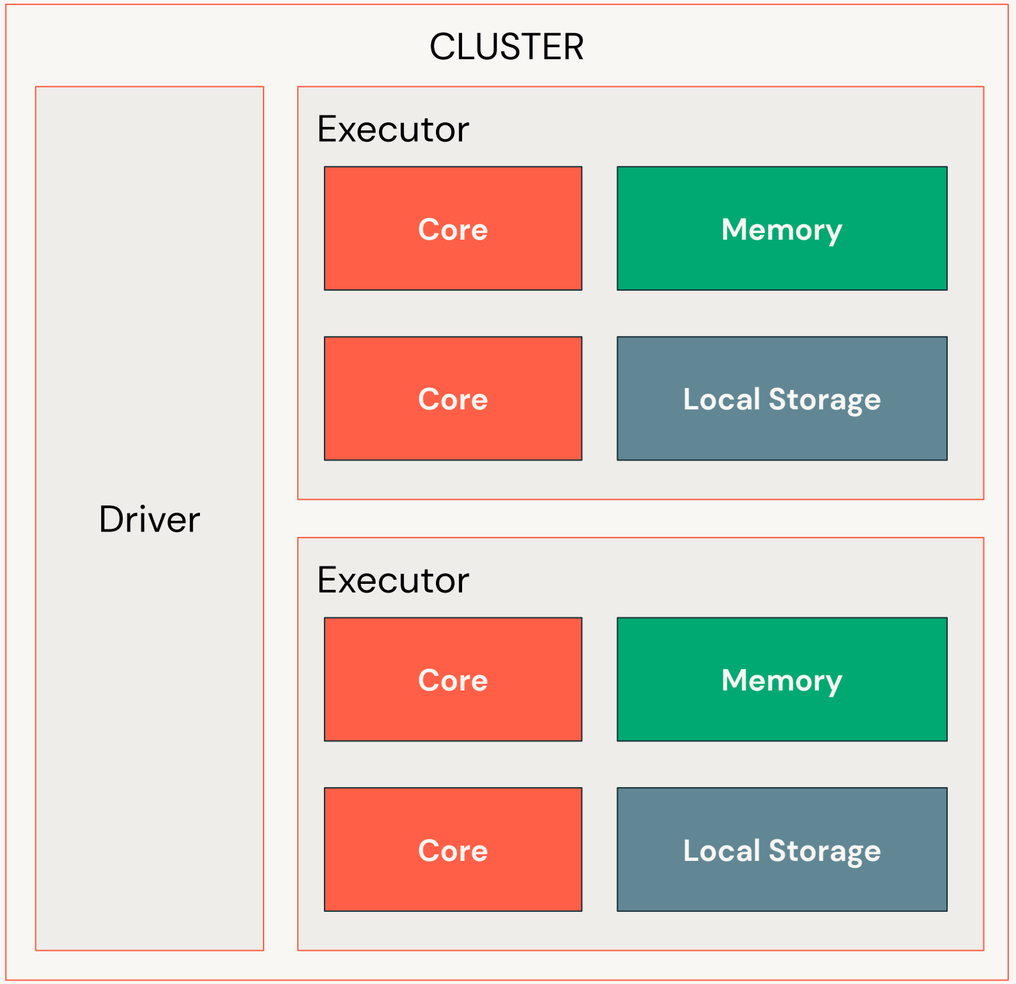
Clusters

Overview

Clusters are made up of one or more virtual machine (VM) instances

Driver coordinates activities of executors

Executors run tasks composing a Spark job



Clusters

Types

All-purpose Clusters

Analyze data collaboratively using interactive notebooks

Create clusters from the Workspace or API

Retains up to 70 clusters for up to 30 days.

Job Clusters

Run automated jobs

The Databricks job scheduler creates job clusters when running jobs.

Retains up to 30 clusters.

databricks

Git Versioning with Databricks Repos

Databricks Repos

Overview

Git Versioning

Native integration with Github, Gitlab, Bitbucket and Azure Devops

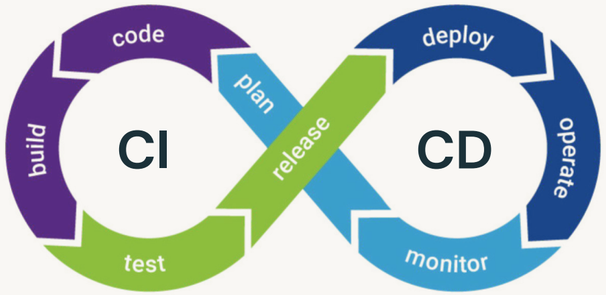
UI-based workflows

Azure DevOps GitHub GitLab Bitbucket

CI/CD Integration

API surface to integrate with automation

Simplifies the dev/staging/prod multi-workspace story



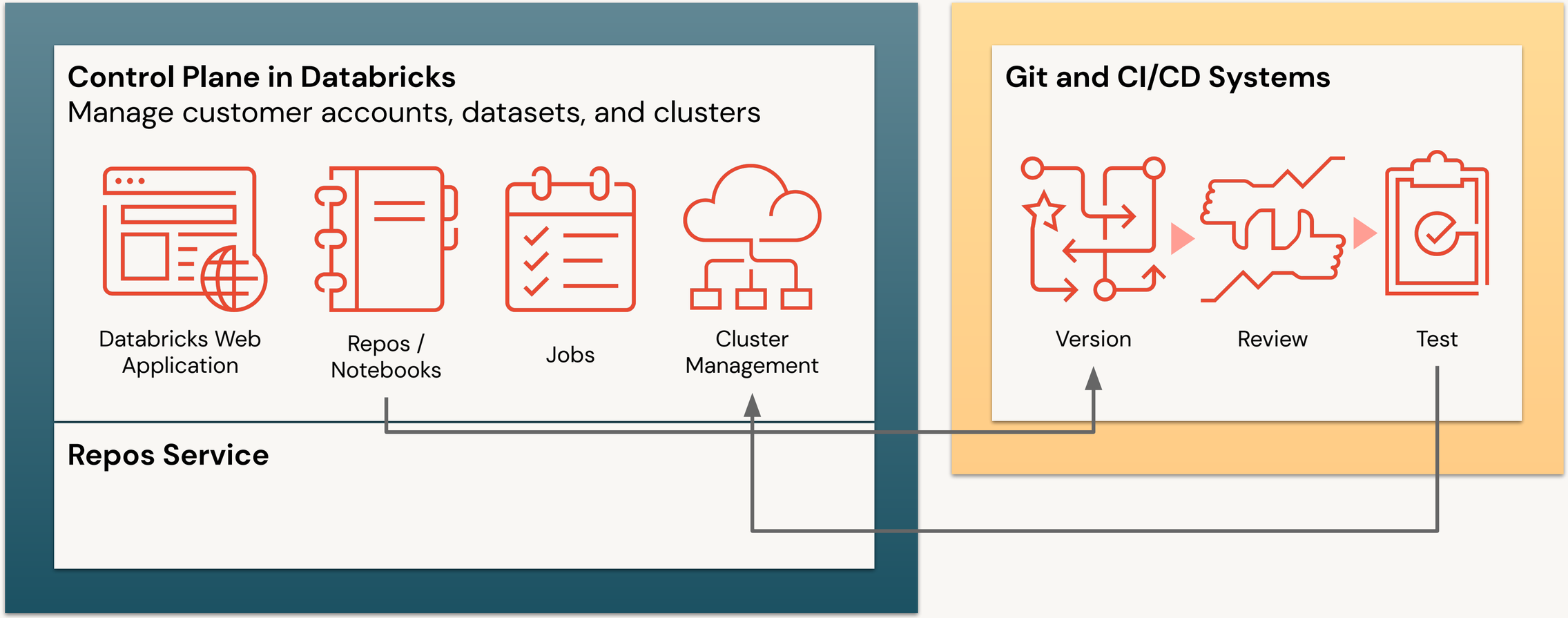
Enterprise ready

Allow lists to avoid exfiltration

Secret detection to avoid leaking keys

Databricks Repos

CI/CD Integration



Databricks Repos

Best practices for CI/CD workflows