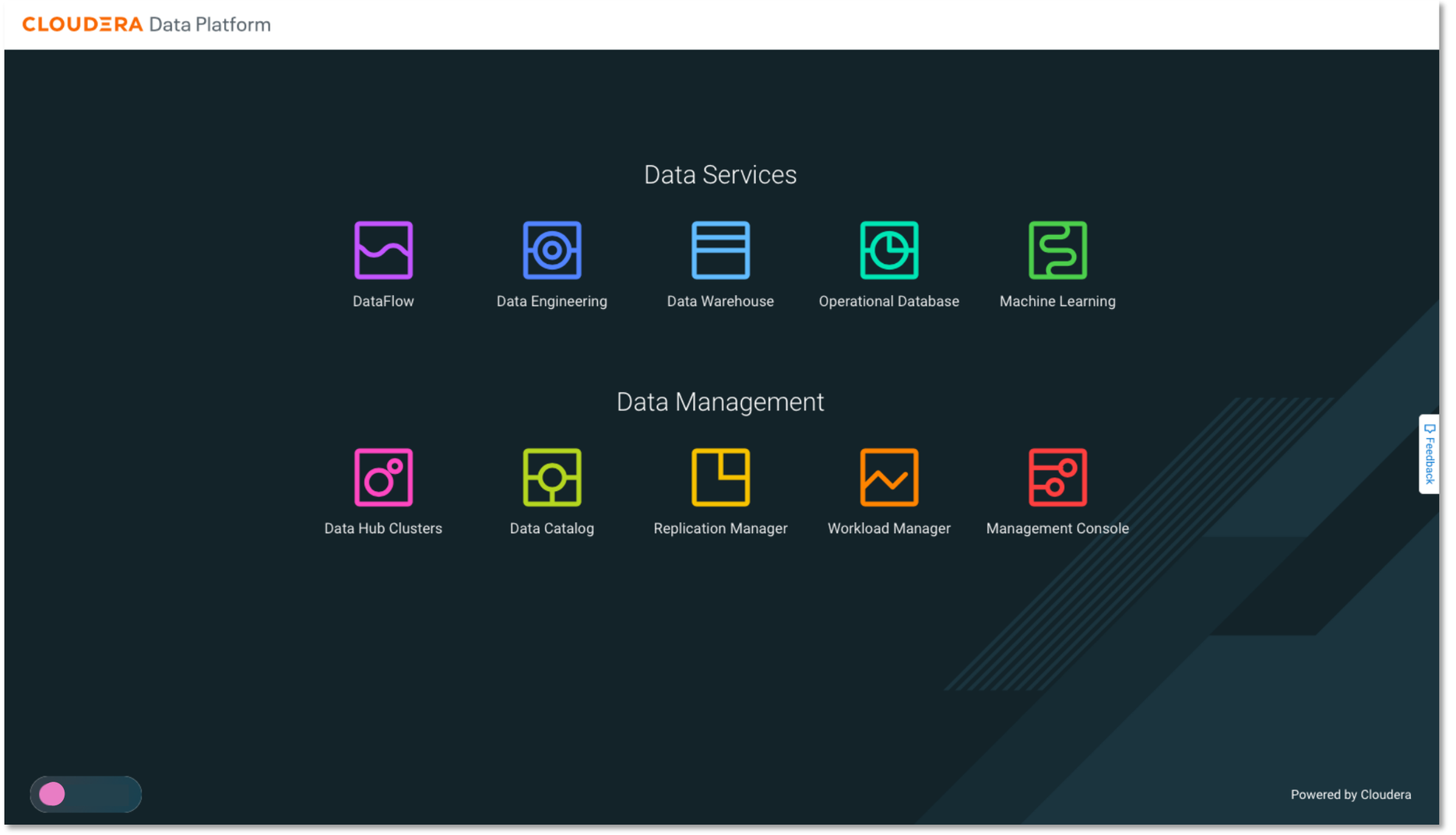
CLOUDERA
Educational Services

Cloudera Data Services

CLOUDERA

Introduction to Data Services

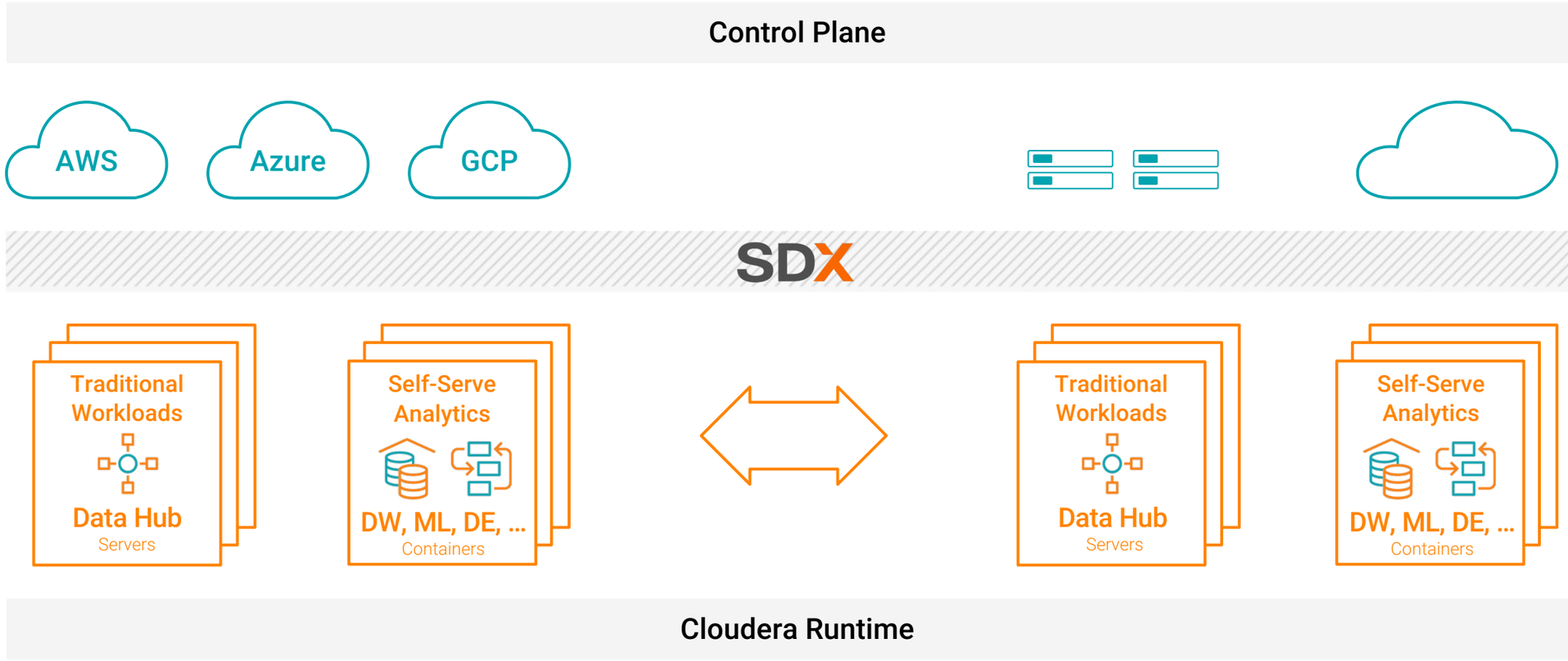
Home



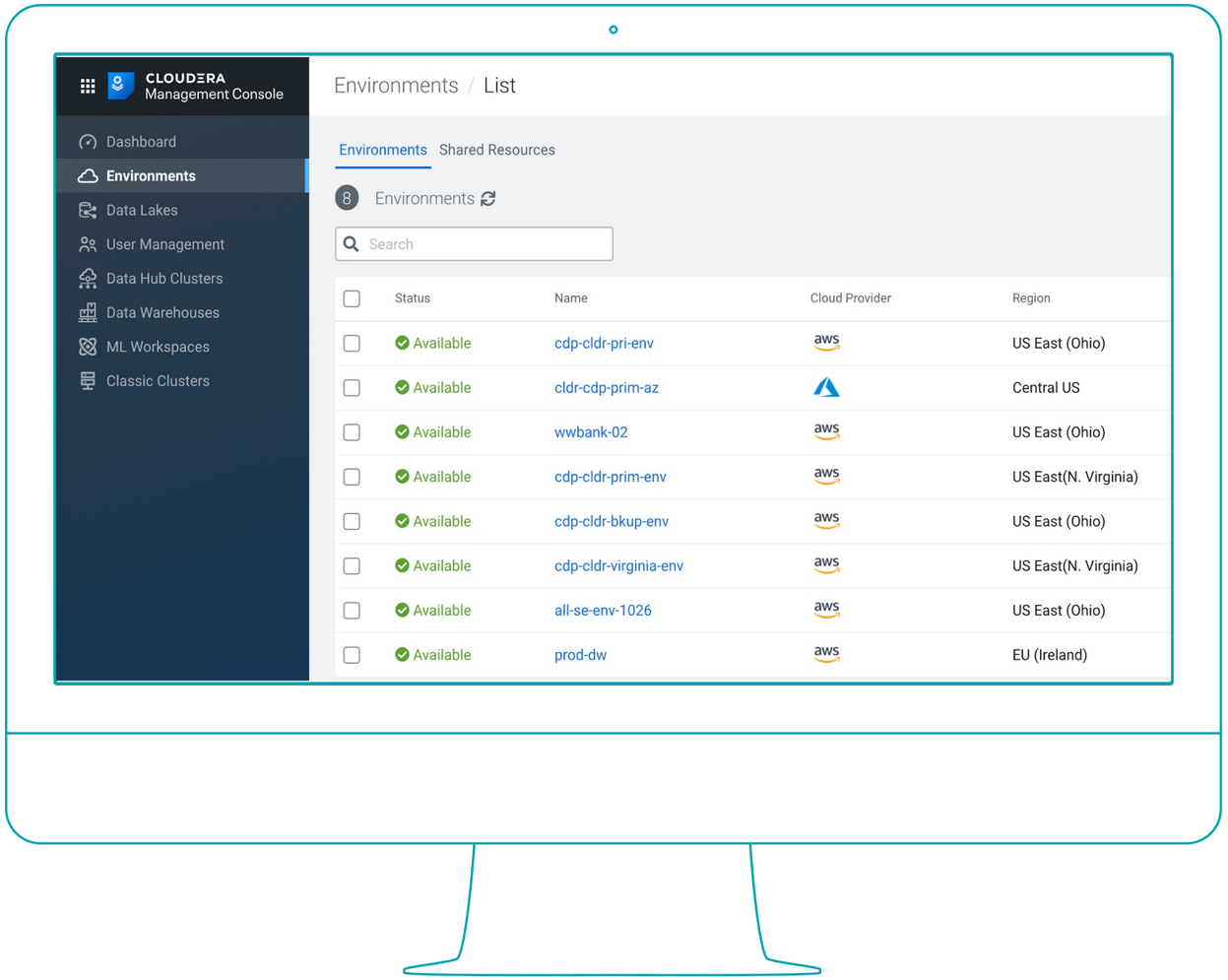
One Platform – Two Form Factors

CDP Public Cloud
(platform-as-a-service)

CDP Private Cloud
(platform as installable software)



Key Concepts & Components



ENVIRONMENTS

Environment

- 1 Template
- 1 Region
- 1 VPC
- Multiple Roles/Buckets

1:1

Data Lake

- SDX: Atlas, Ranger, Knox, IdBroker, CM
- Associated with groups/users

1:N

Data Hub Clusters / Data Services

- DH templates
- ML Env
- DW Database Catalogs/Virtual Compute

Data Hub Clusters and Data Services

What are the consumption options?

A **Data Hub Cluster** is a customizable environment that runs like a traditional Hadoop cluster, but is designed to leverage Cloud Storage.

A **Data Service** is a container-based compute environment for specific purposes:

ML, DW, DE, OD, DF

Data Hub Clusters

Data Warehouse

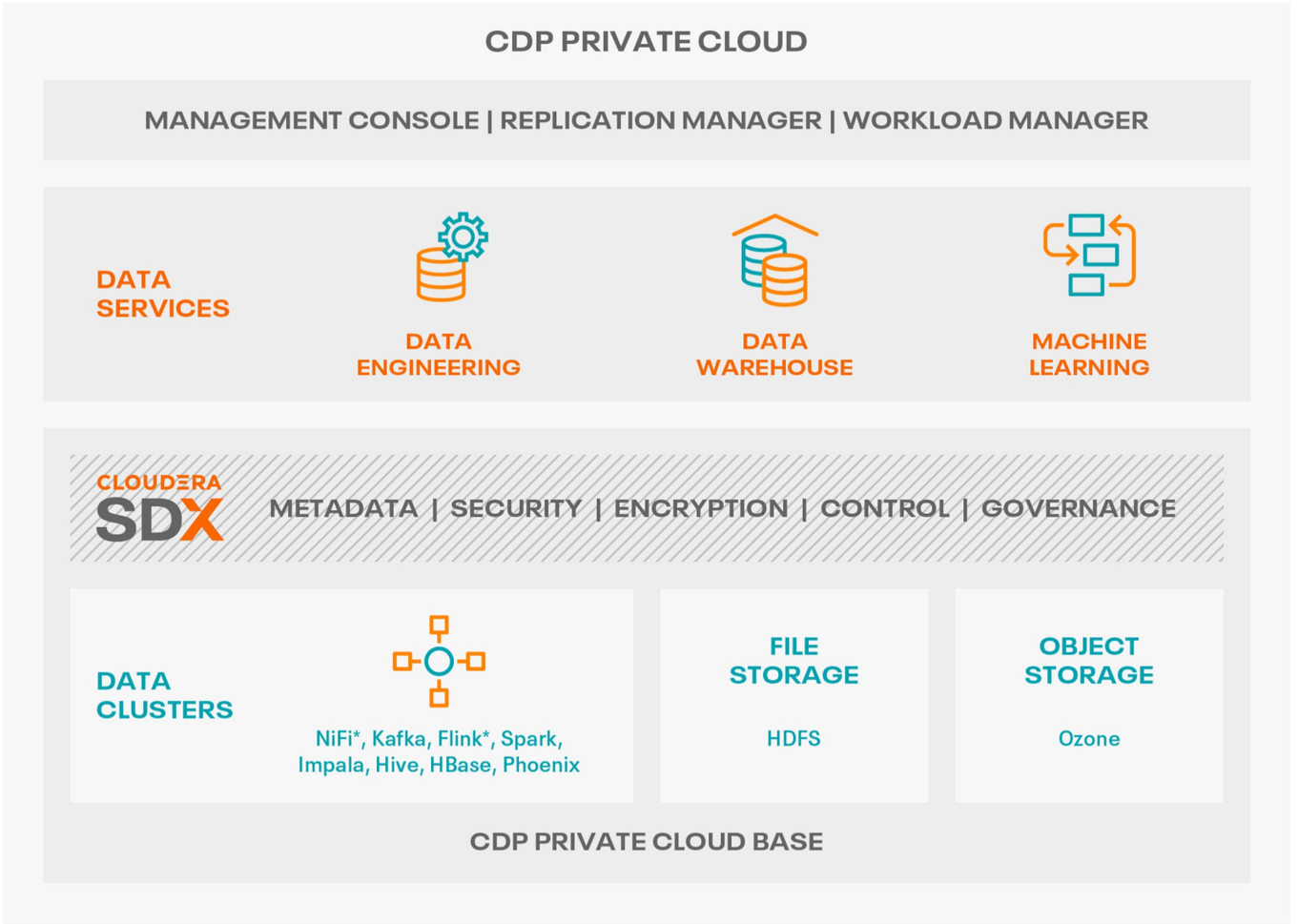
Machine Learning

What is CDP Private Cloud?

New set of data analytics applications
Featuring use-case optimized interfaces

Running on its own container cloud
Simplified deployment, with fast provisioning and efficient scaling

With access to a shared data lake
That is secured and governed



*Standalone SKU

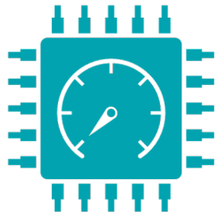
Impact of Challenges

Noisy neighbors

"The Marketing team is stalling the cluster, again!"

Complex upgrades

"We are still on 2.x, you'll have to wait to be able to use 3.x"



Lack of elasticity

"We don't use more than 40% of our infra, and yet some services often lack resources"

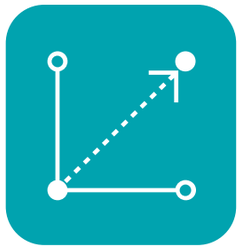
Extended onboarding cycles

"Your environment should be ready in 4-6 weeks, hopefully…"

Co-location of Storage & Compute

"We can't grow compute independently of storage"

Advantages of CDP Private Cloud



Meet SLAs with Predictable Performance

Workload isolation and better managed multi-tenancy minimizes spikes impacting critical workloads

Improve Cost Efficiency

Optimized resource utilization from disaggregated storage affords cost efficiencies across the cluster

Rapid Time-to-Value

Simplified provisioning and onboarding of new use-cases

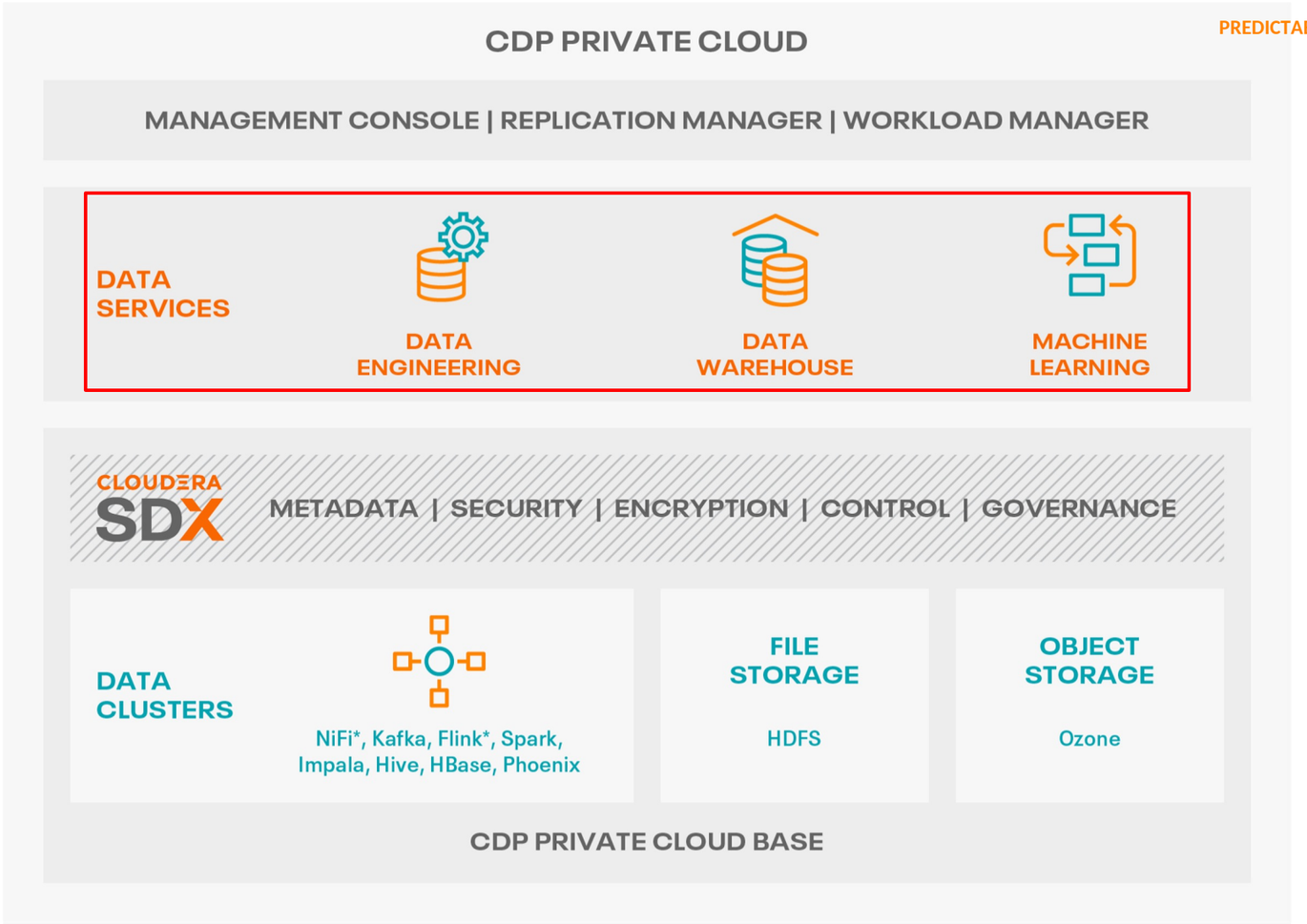
A Deeper Look at CDP Private Cloud

Predictable workload performance with tenant isolation

PREDICTABLE

Avoid Noisy Neighbors

- Tenant isolation provided by the container cloud, enabling dedicated compute
- Predictable performance ensures you meet your application SLAs



*Standalone SKU

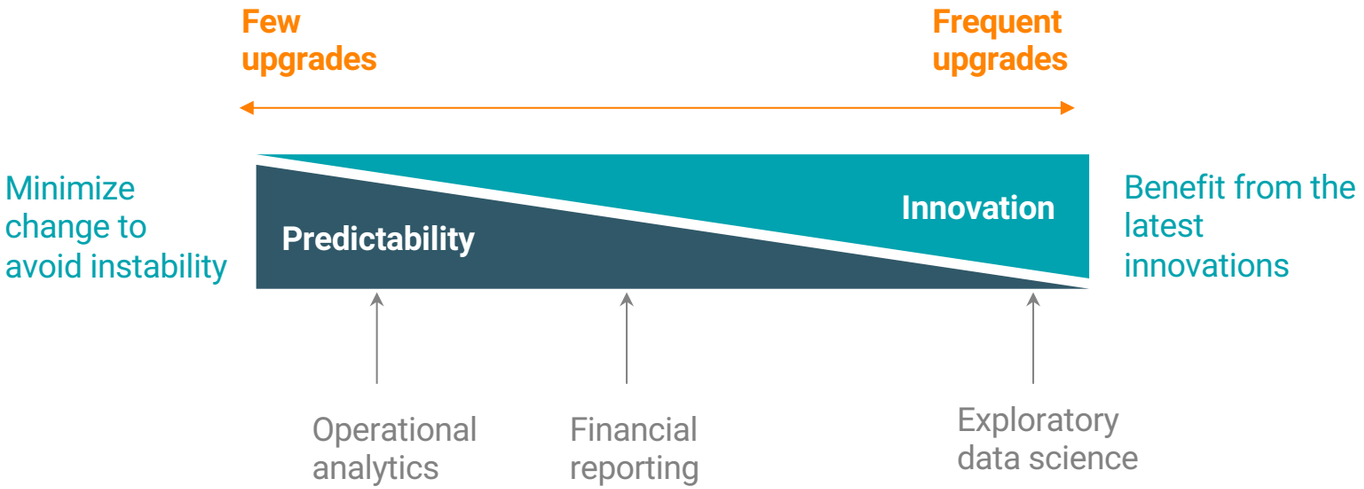
A Deeper Look at CDP Private Cloud

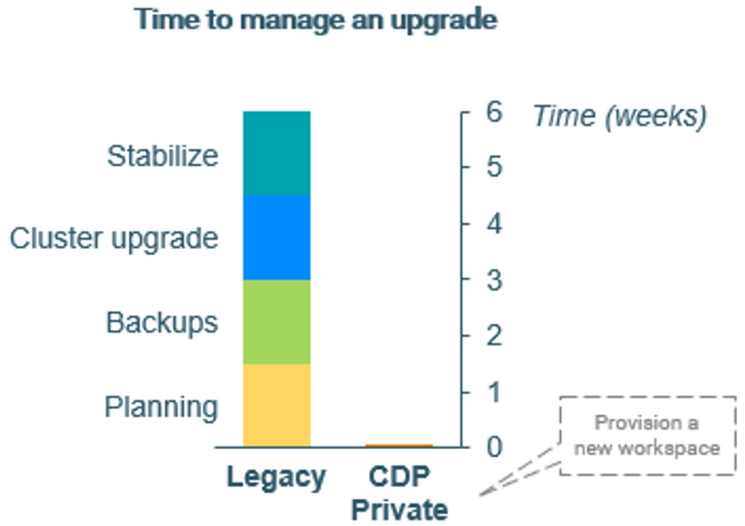
Better multi-tenancy for ‘upgrade agility’

PREDICTABLE

Independent Upgrades

- Upgrade each tenant when needed, without impacting others
- Teams favoring stability vs innovation are no longer at odds



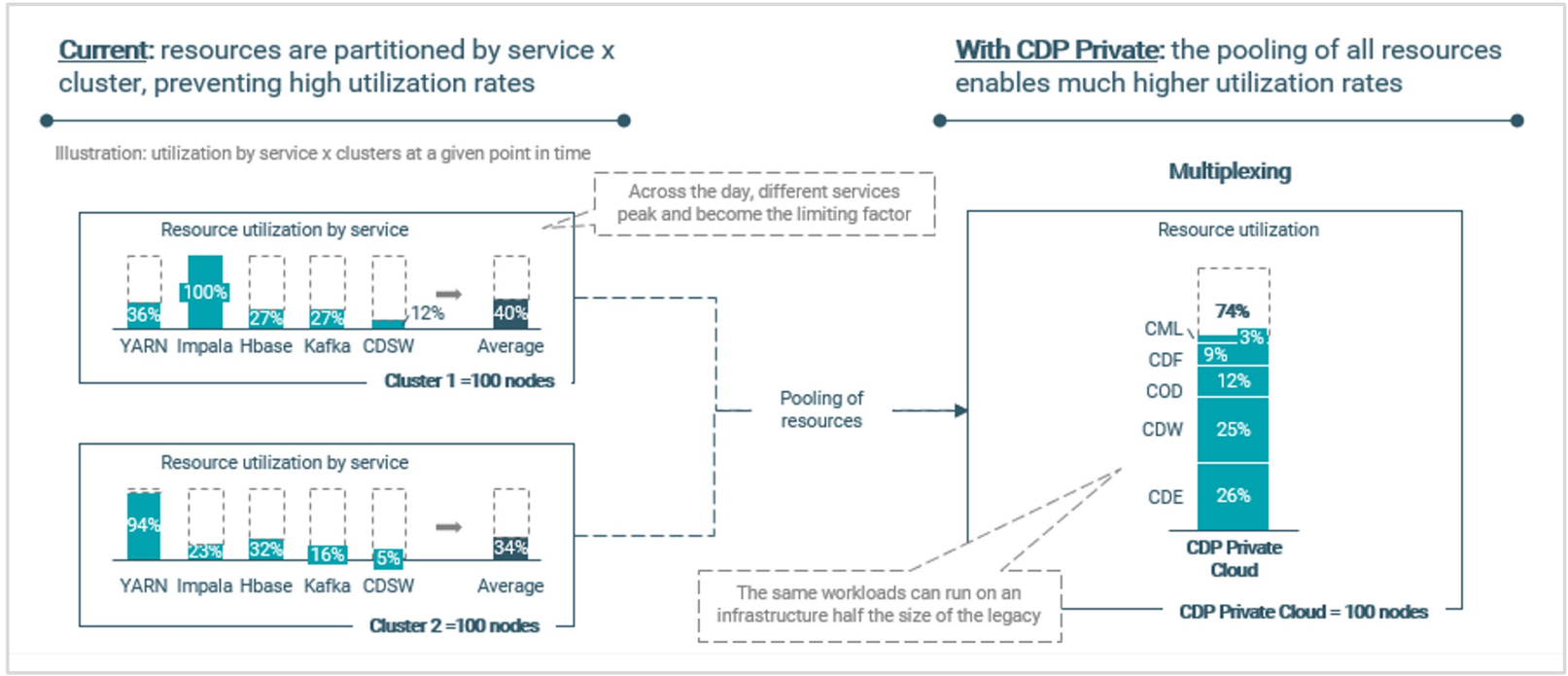


A Deeper Look at CDP Private Cloud

Performance without replicating data or creating silos

Consolidate Your Clusters

- Consolidate clusters for higher utilization and better ROI
- Shared data lake with single source of data and metadata
- Consistent schema, security & governance. Set policies once, apply everywhere
- Simplify multi-function job creation
  [Ingest → ETL → DW → ML → ODB]



A Deeper Look at CDP Private Cloud

Improve cost efficiency with better infrastructure utilization

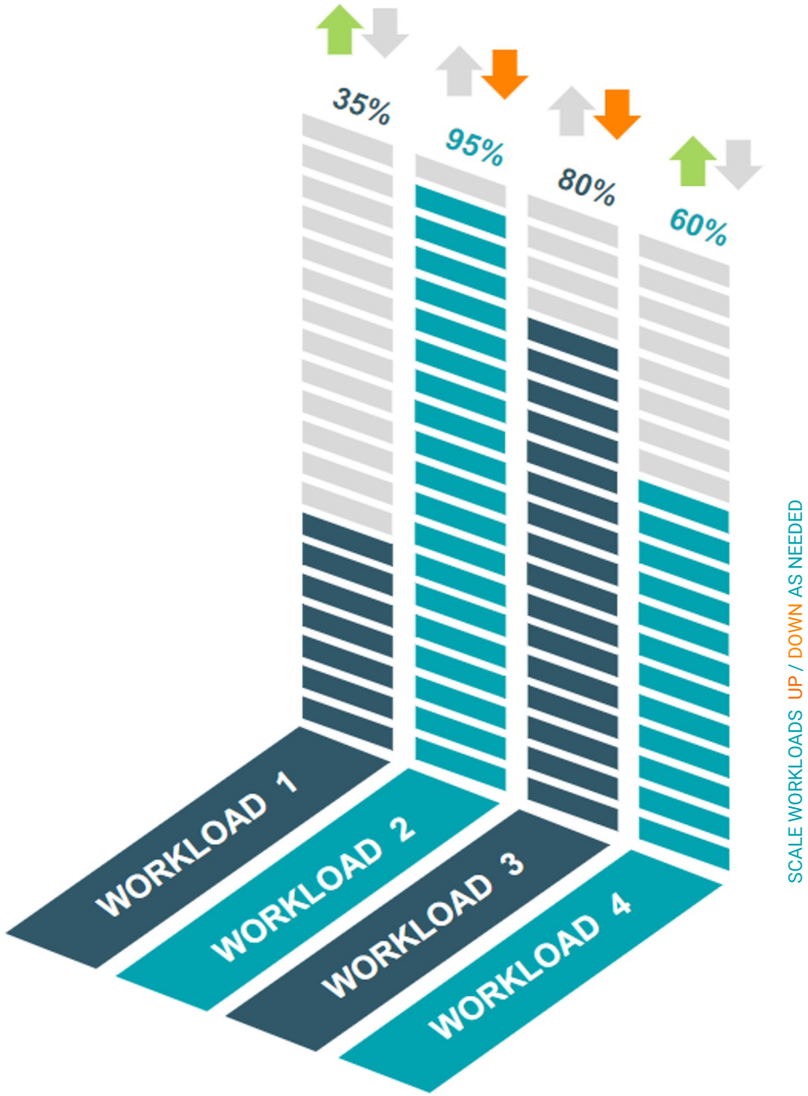
COST EFFICIENT

Elasticity to Auto-Scale & Auto-Suspend

- Use what you need, when you need it
- Shift excess capacity according to demand

Containerized Compute Platform

- Workloads abstracted from infrastructure
- Fully decoupled compute and storage
- Embedded Container Service or bring your own Red Hat OpenShift license



** See product roadmap for when these capabilities will be available*

A Deeper Look at CDP Private Cloud

Faster time-to-value with simplified onboarding

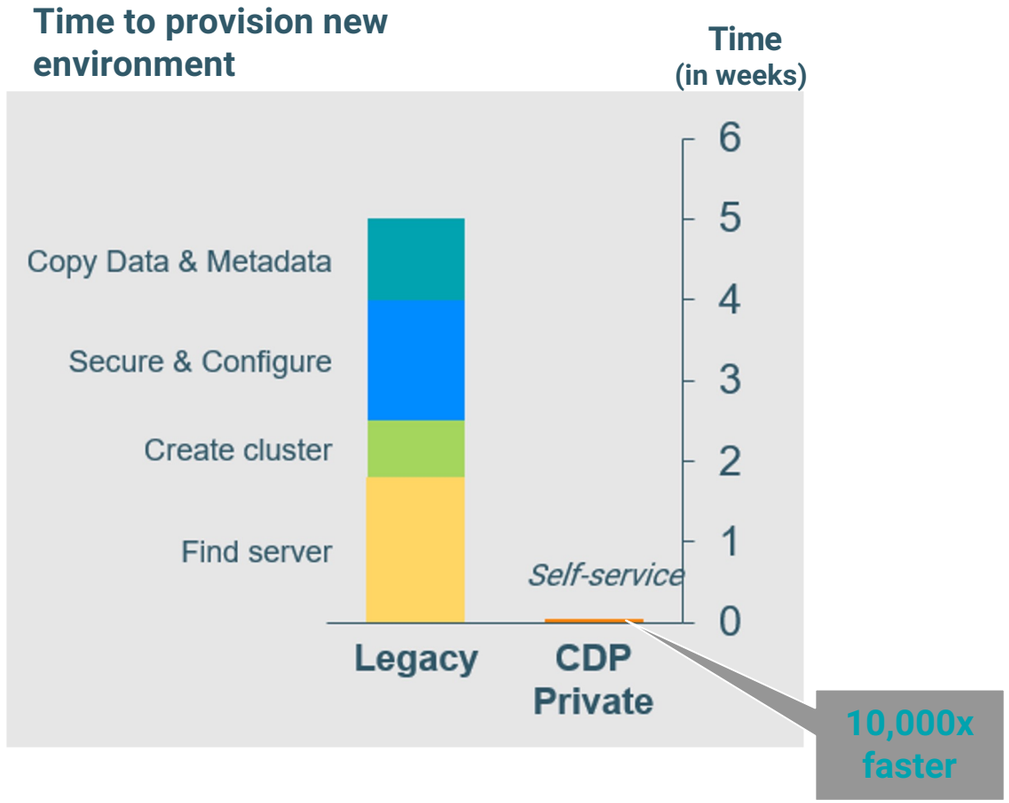
RAPID

Push-button Provisioning

- Near instantaneous provisioning - reduce weeks of work down to minutes

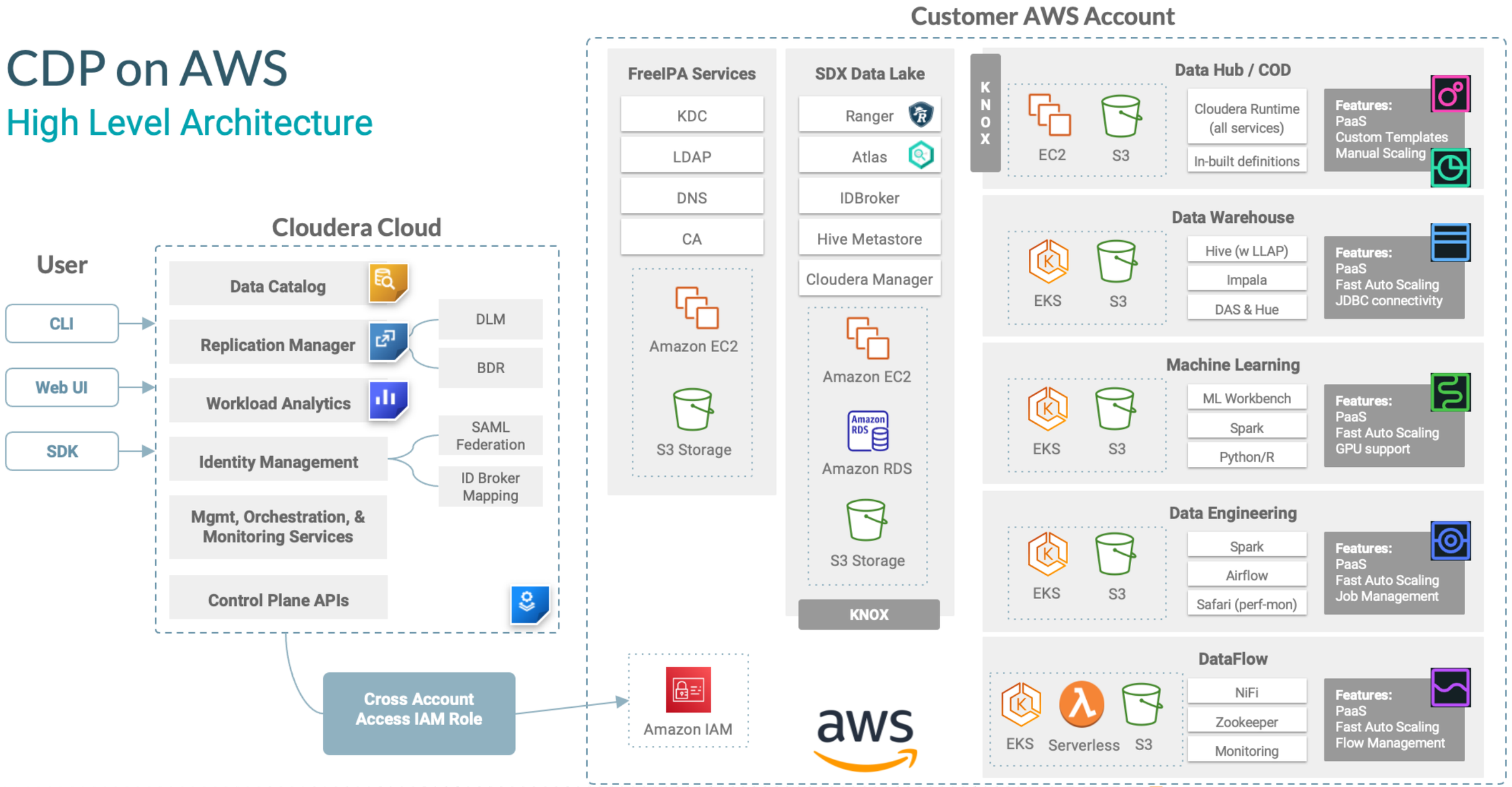
Redesigned User Interfaces

- Workflows optimized for self-service analytical experiences



CDP on AWS

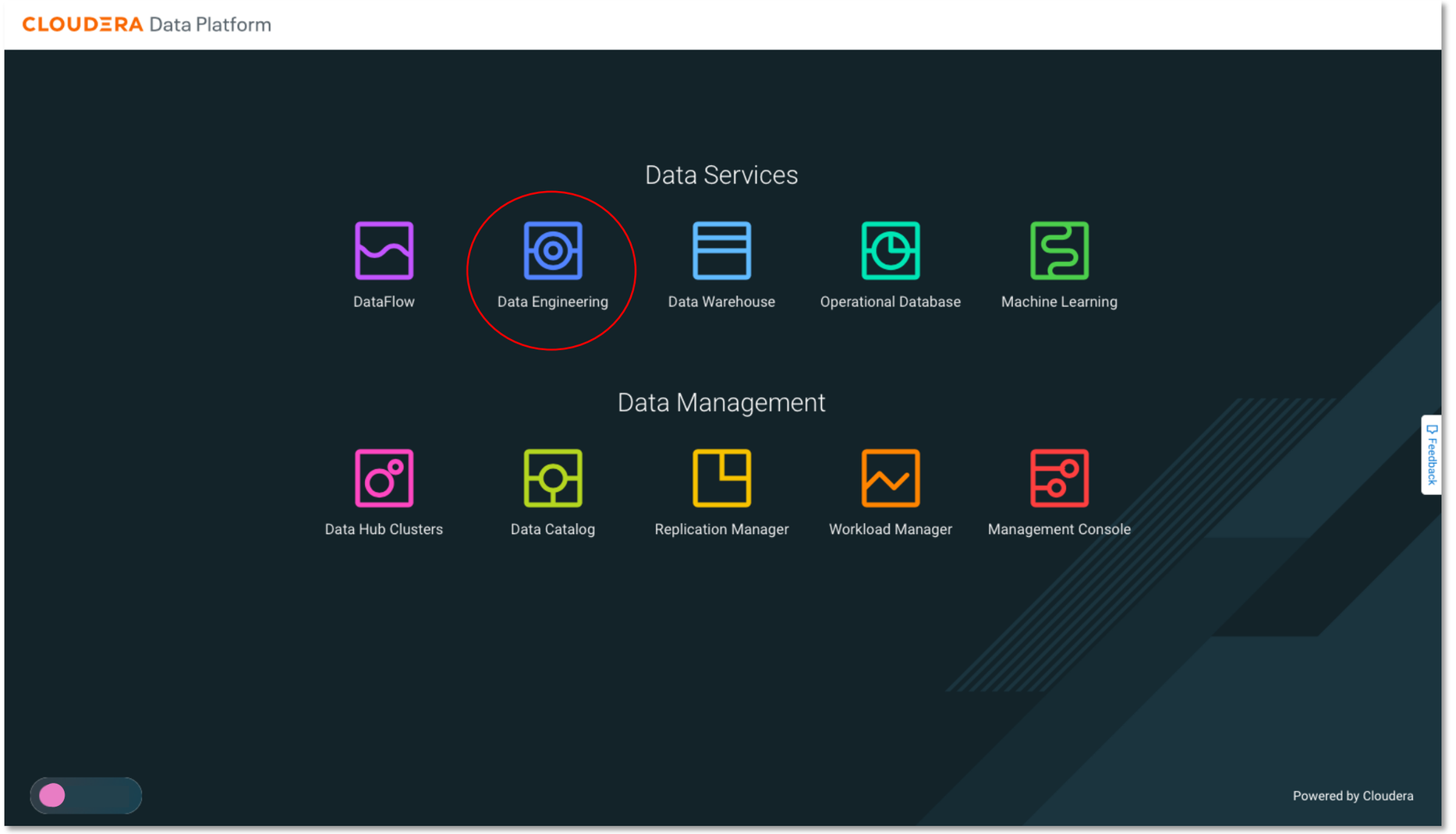
High Level Architecture



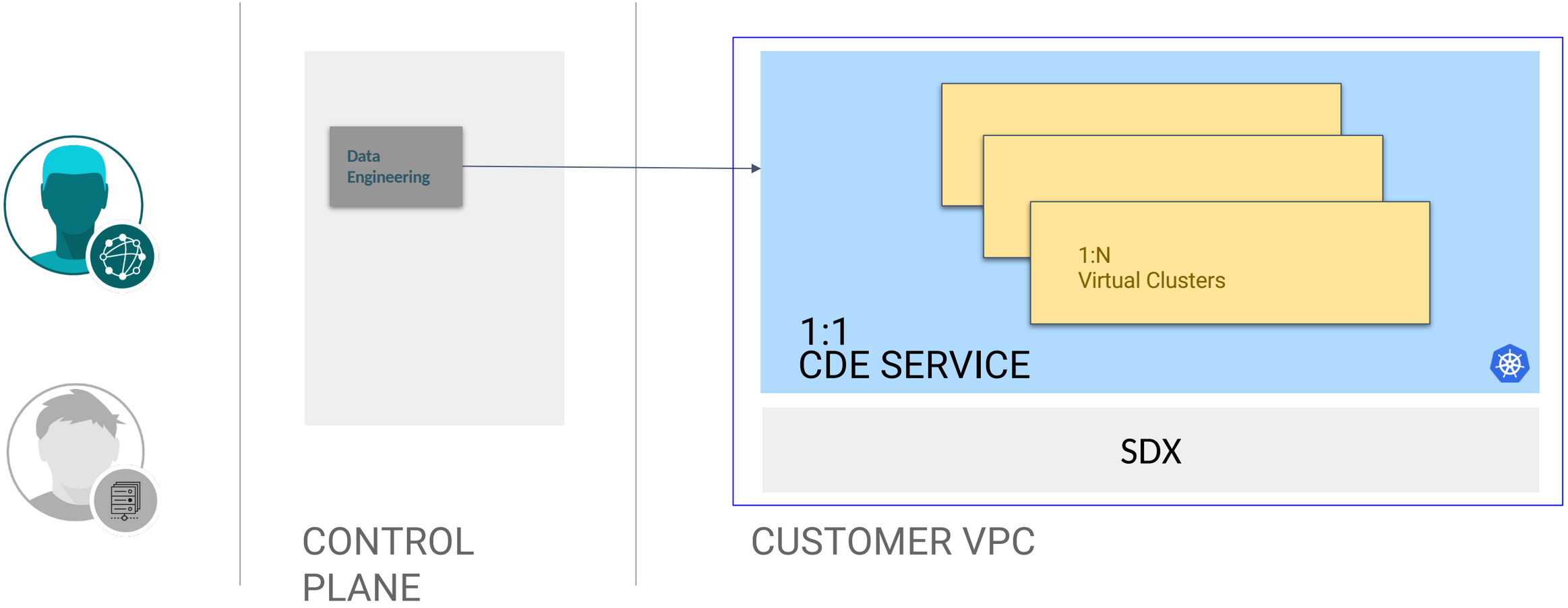
CLOUDERA

Data Engineering Service

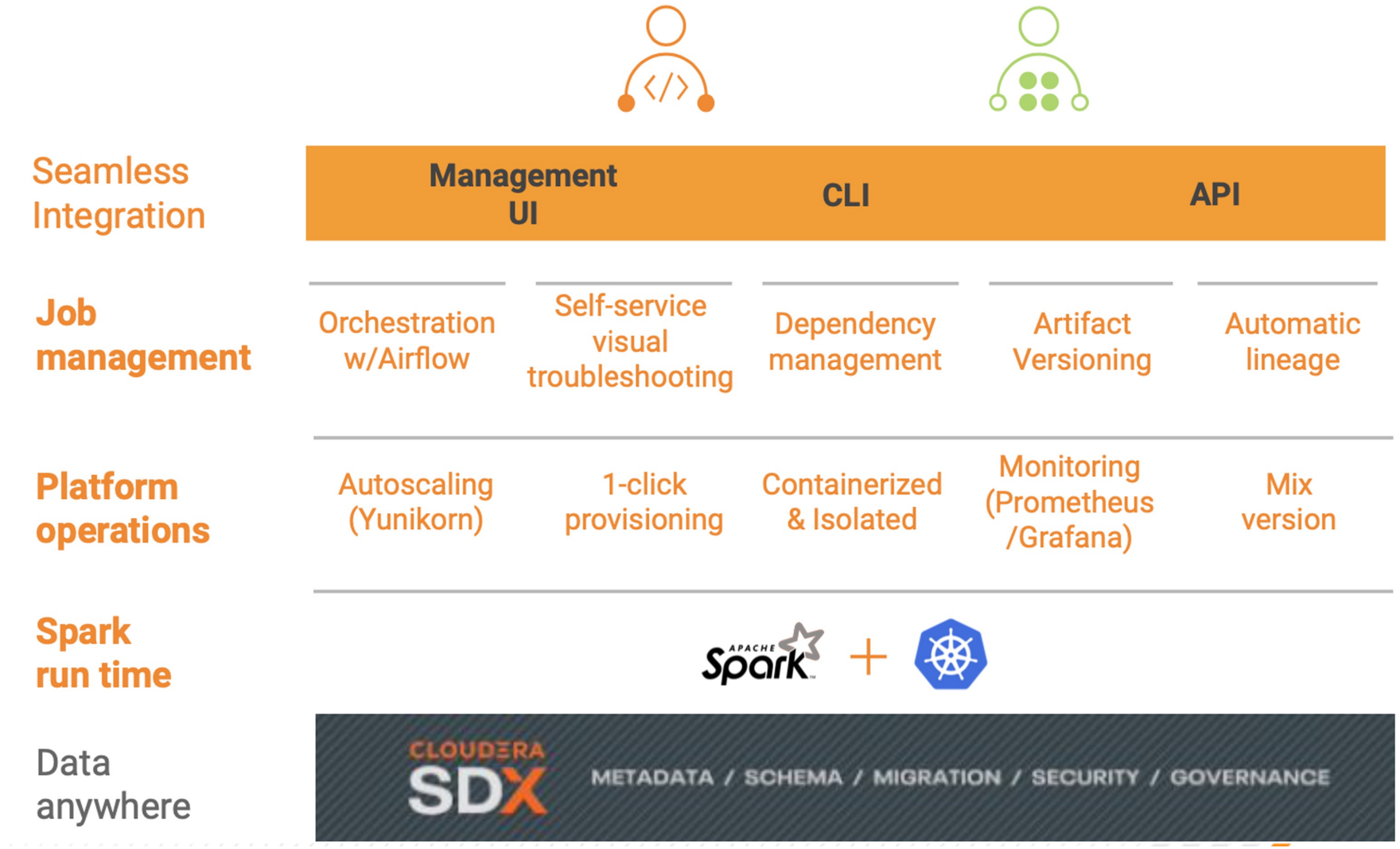
Data Engineering Service



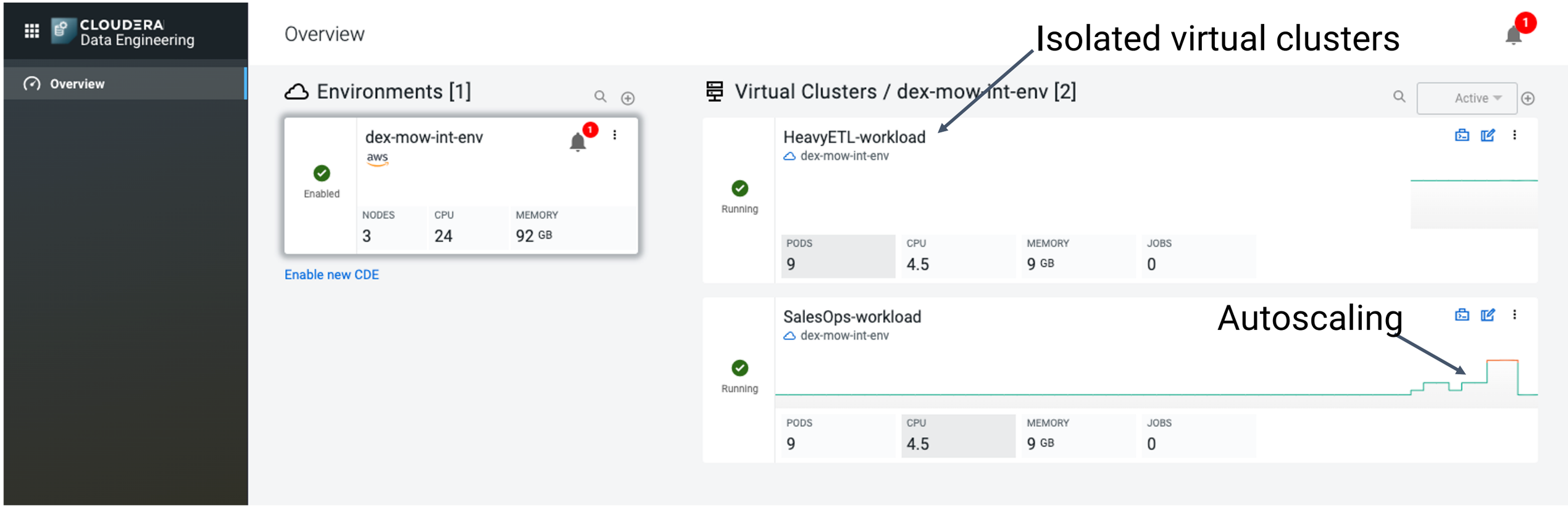
CDE Service Components and Relationships



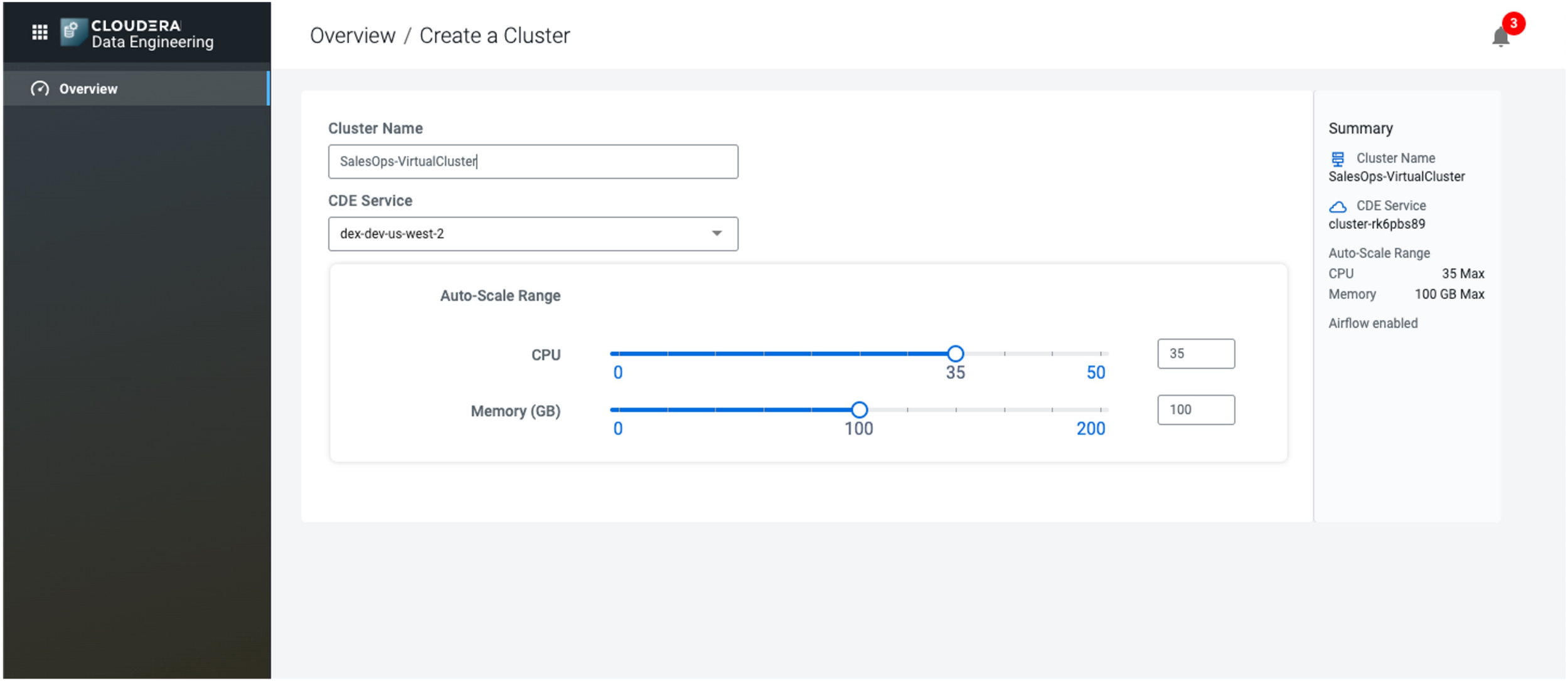
Cloudera Data Engineering



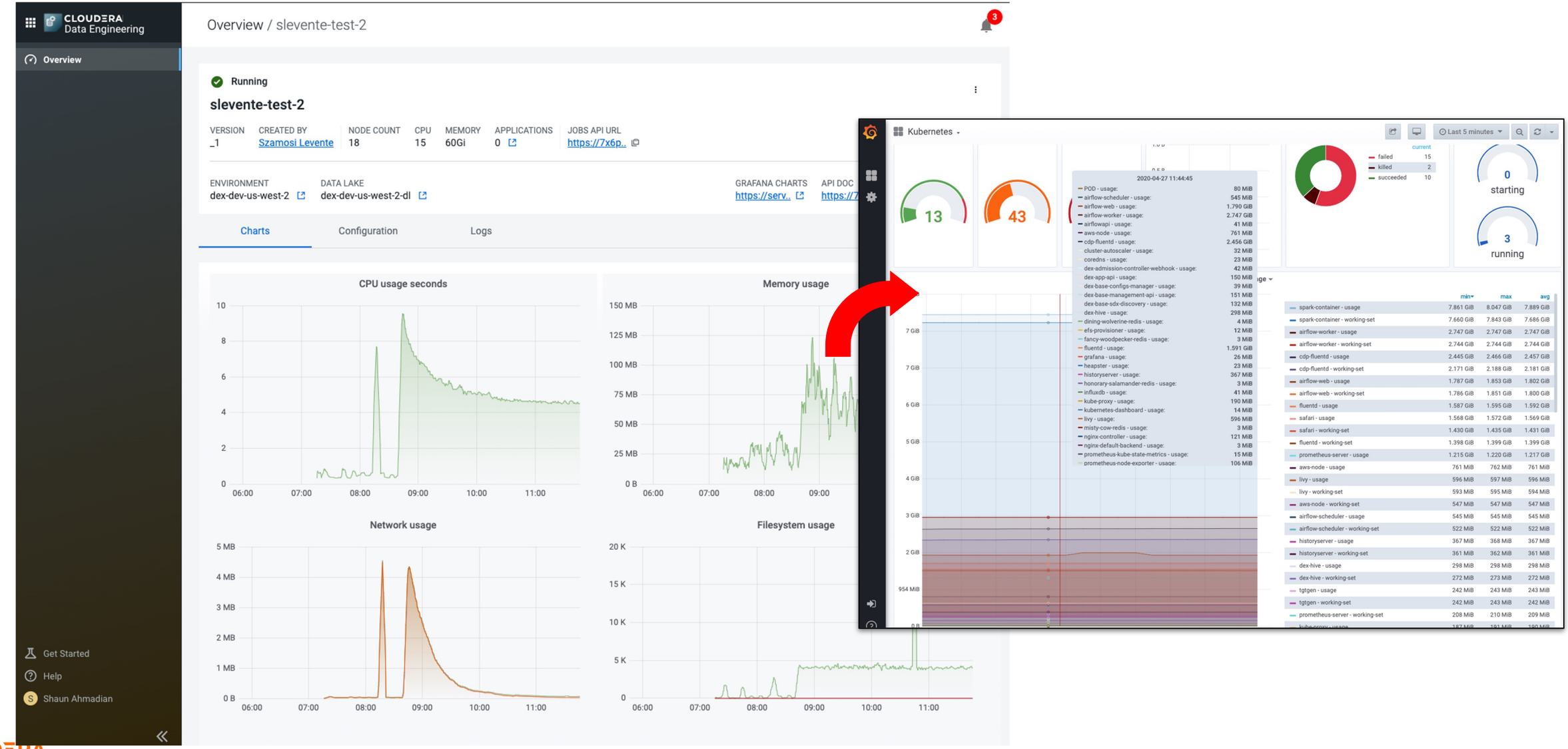
Autoscaling



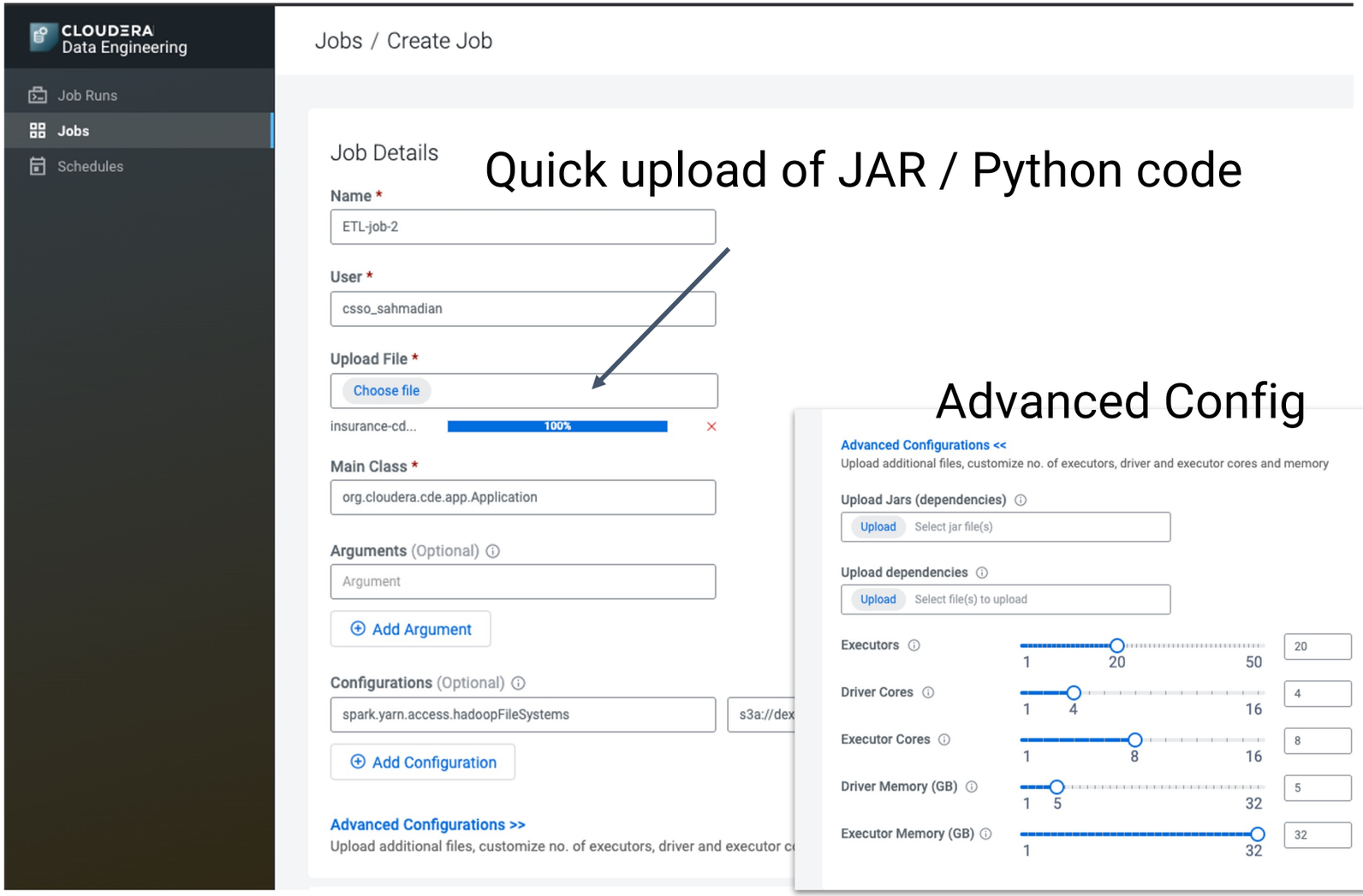
Provision Isolated Workloads with Quotas

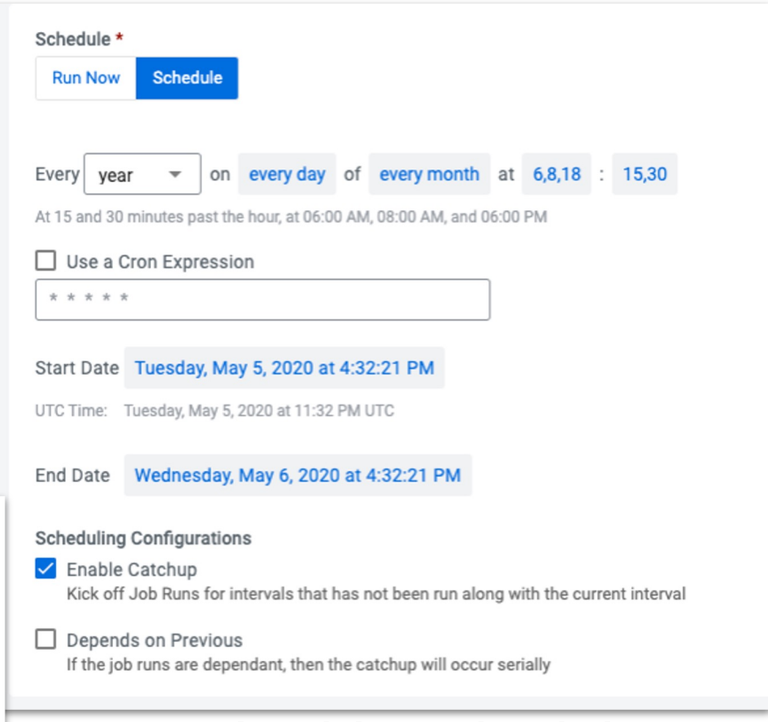


Monitor Capacity & Resource Usage Thru Grafana



Easy Job Deployment

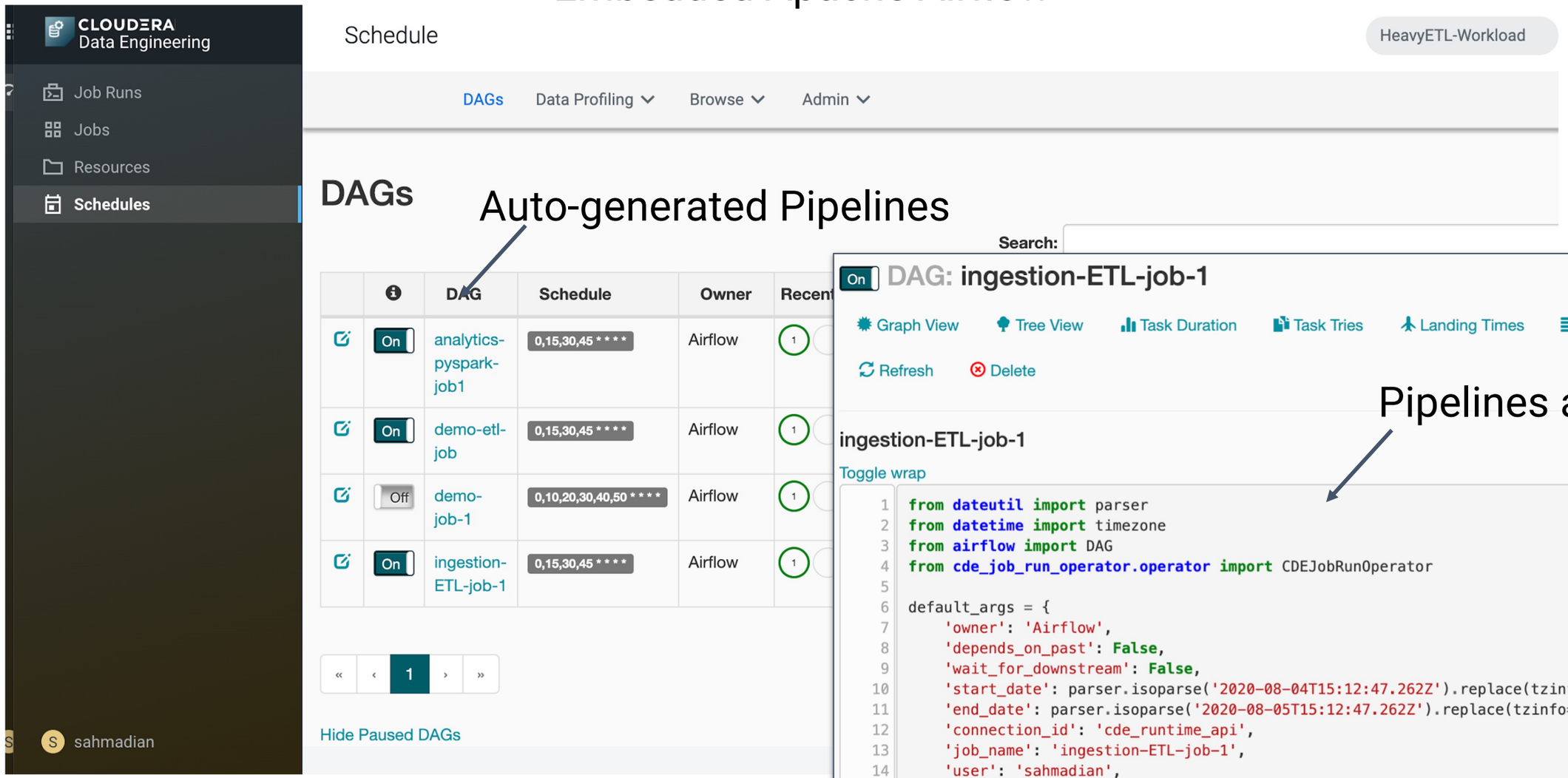




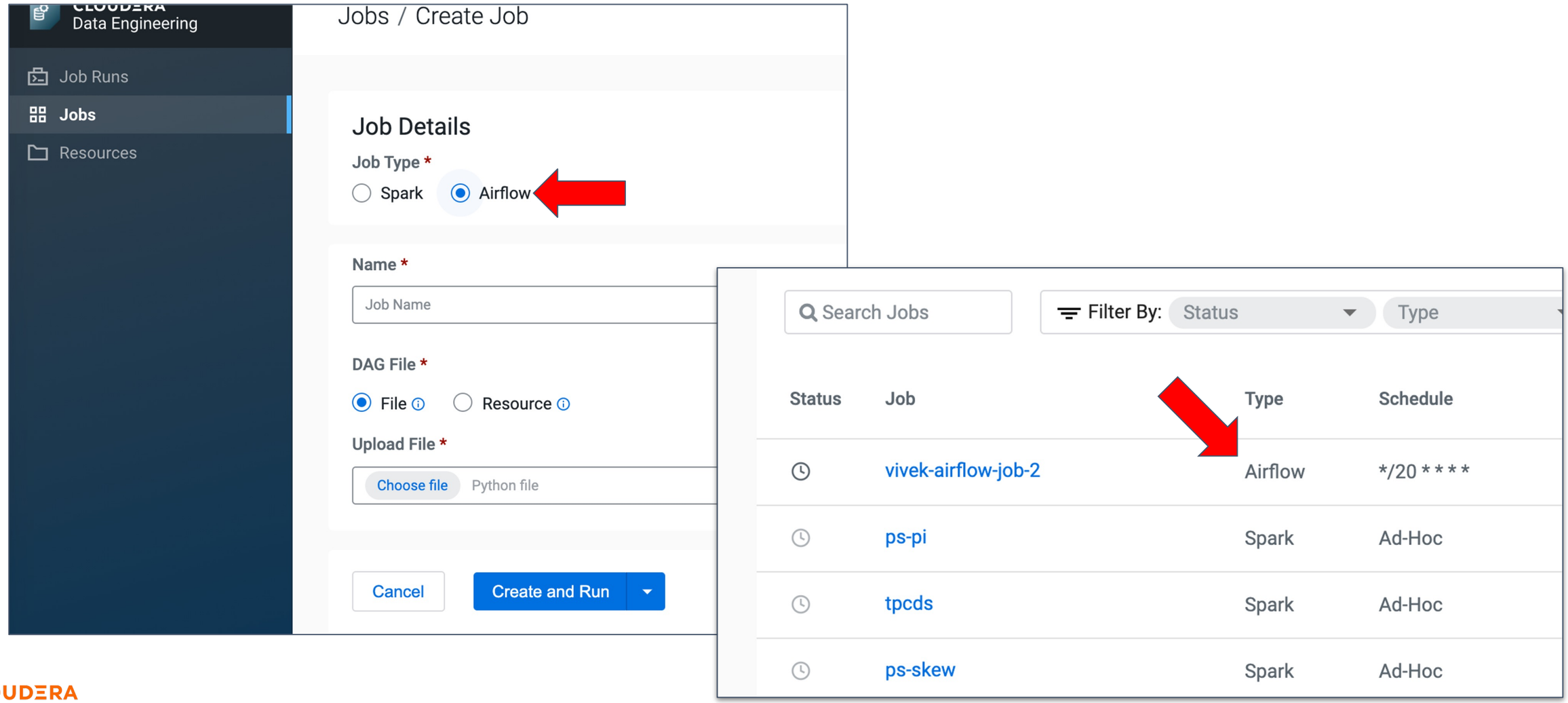
Flexible Scheduler

Flexible Orchestration Backed By Airflow

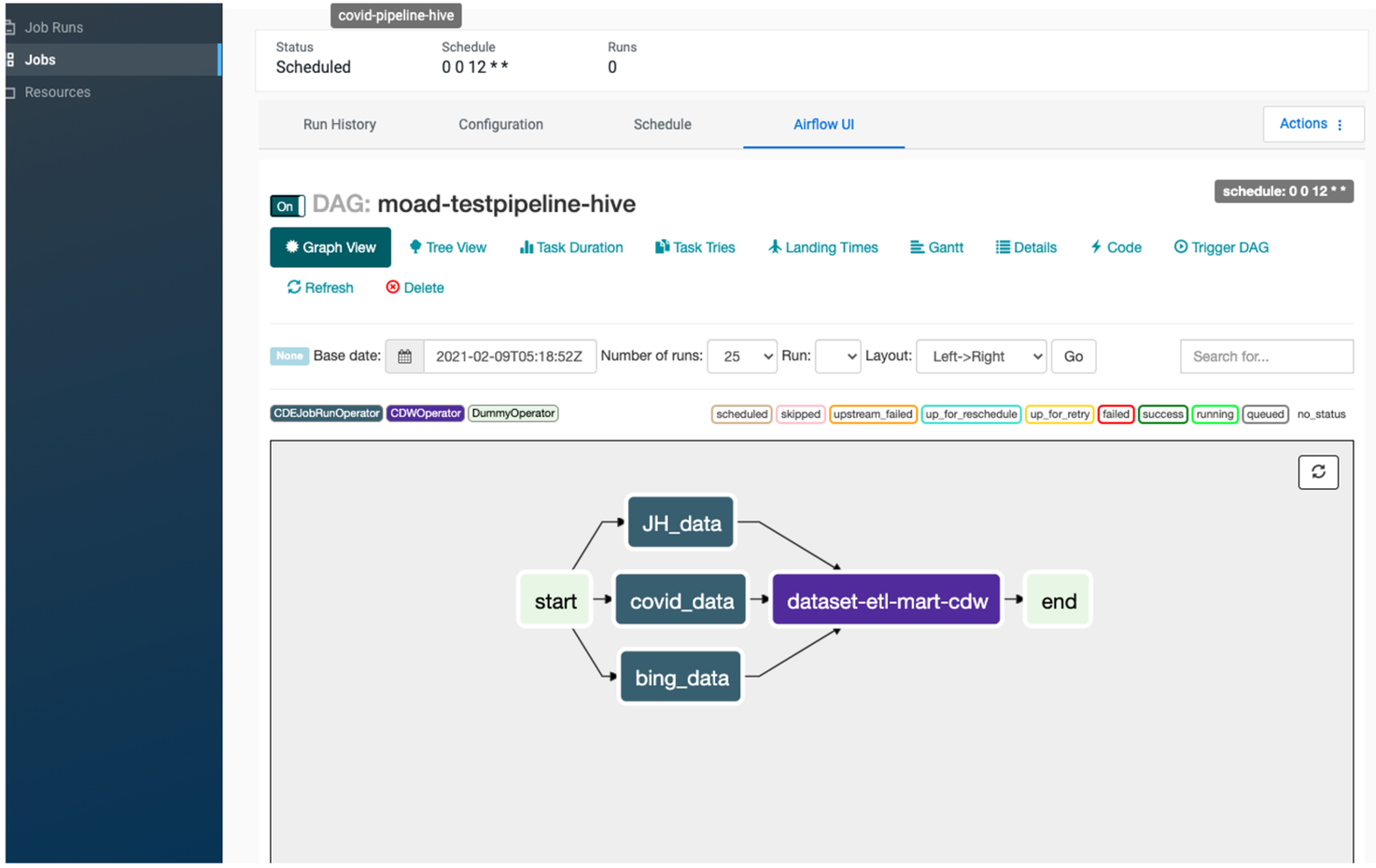
Embedded Apache Airflow



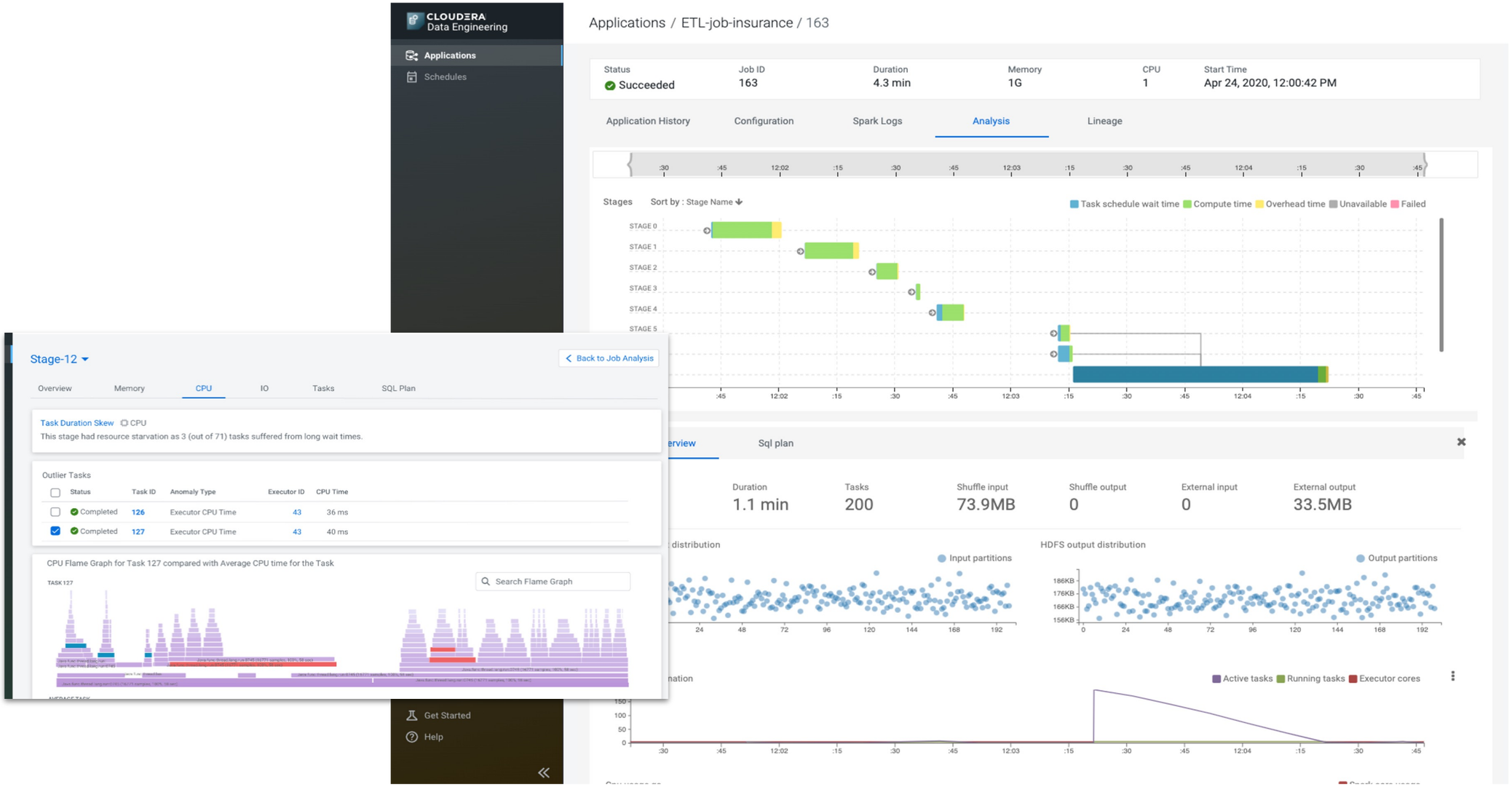
Deploy Airflow Jobs



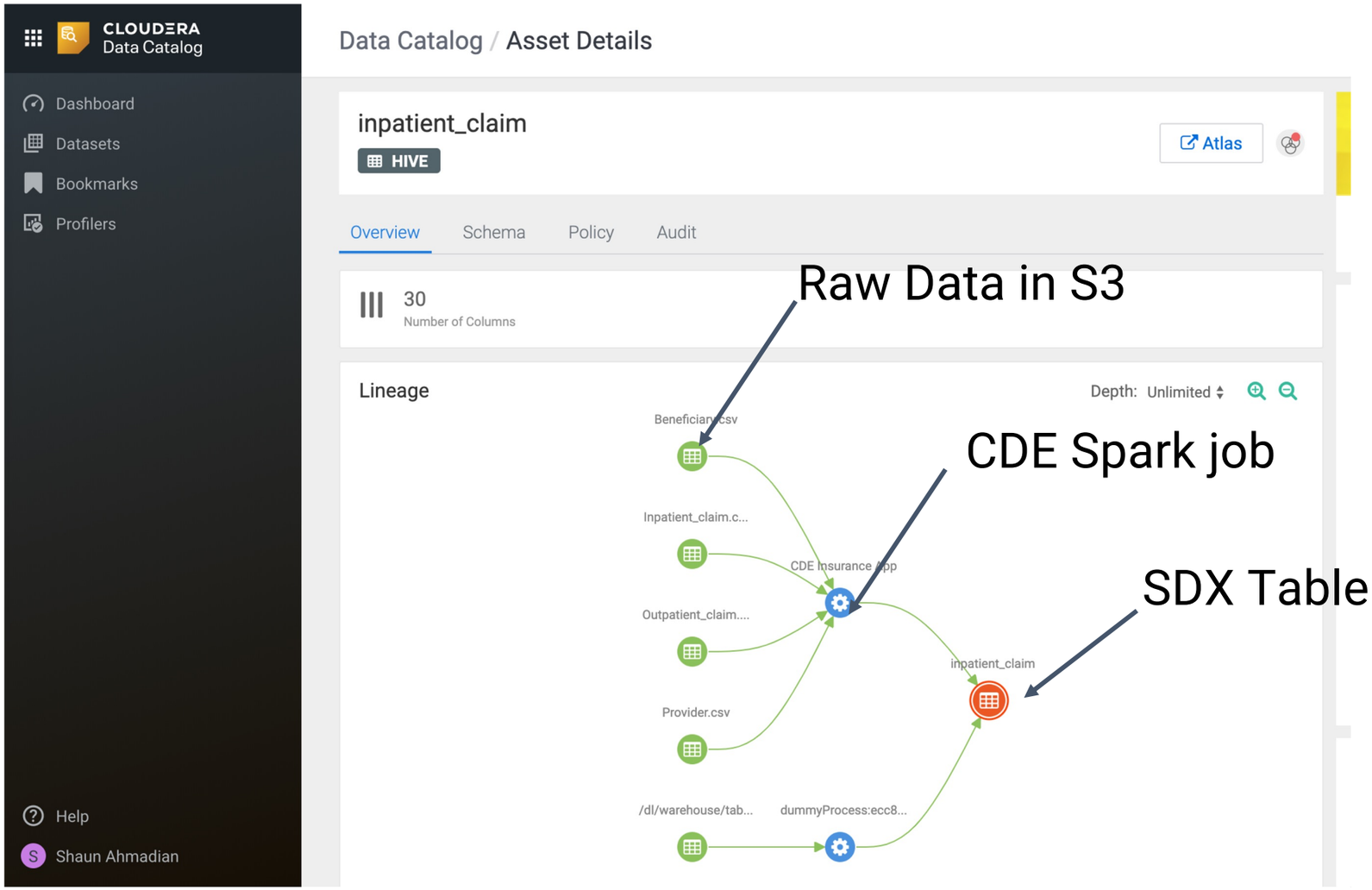
Build Spark And Hive Pipelines With Dependencies



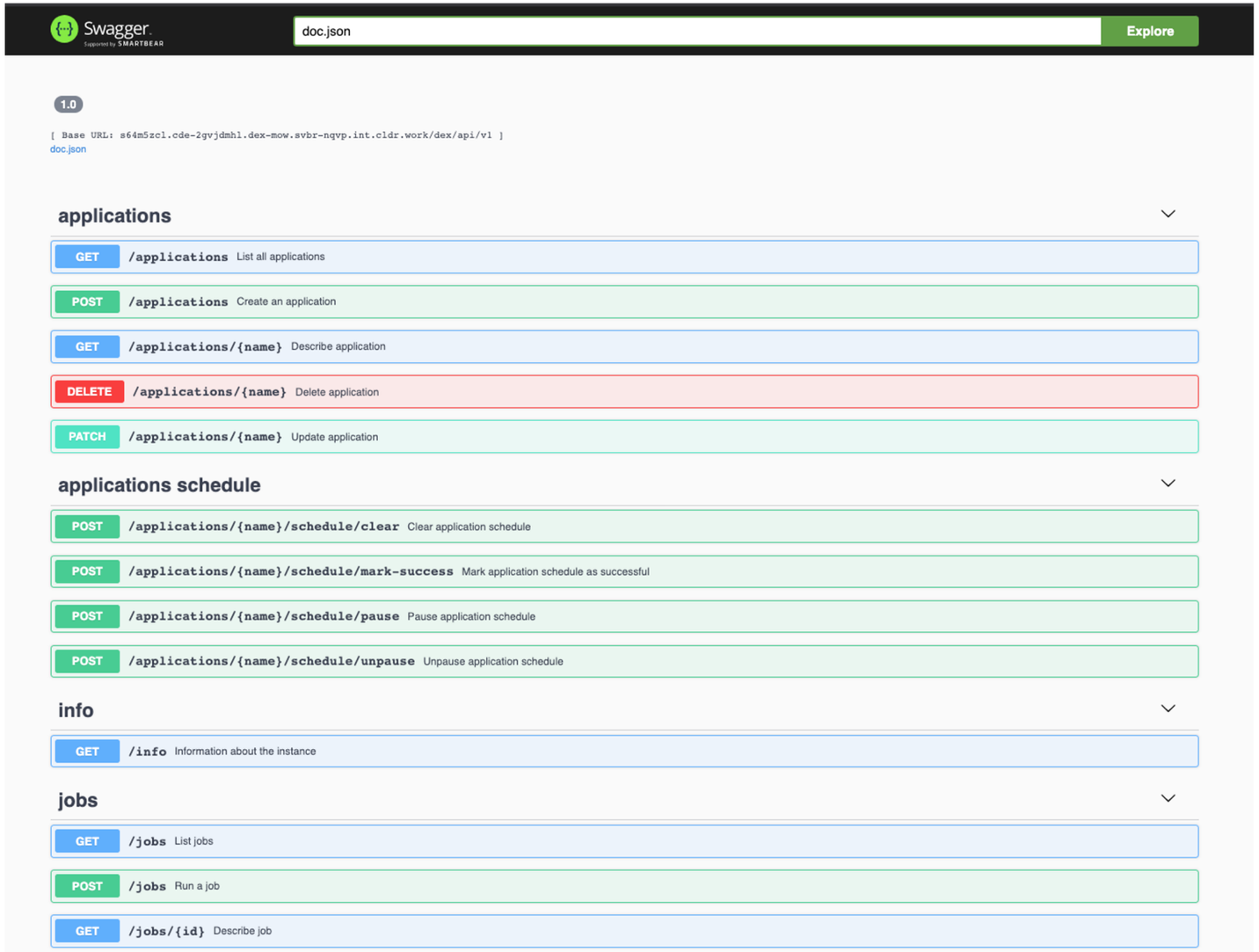
Visual Troubleshooting & Performance Tuning



Automatic Capture Of Data Lineage in SDX



Rich API



Knowledge Check

1. CDE can run Spark on YARN, Spark on Kubernetes and Spark on MESOS.
2. CDE is only available in CDP Public Cloud.
3. CDE allows you to chain multiple Spark jobs using Oozie or Airflow.
4. CDE allows you to run Flink jobs.
5. CDE updates the data lineage in Atlas.
6. CDE comes with a basic API.

Exercises

- Data Engineering Data Service Walkthrough
- Create and Trigger Ad Hoc Spark Jobs
- Add Schedule to Ad Hoc Spark Jobs
- Spark Job Data Lineage Using Atlas

Note: These exercises are Zeppelin notebooks. Go to the Data Hub, select Zeppelin, and then select DE/Labs in list of notebooks.

CLOUDERA

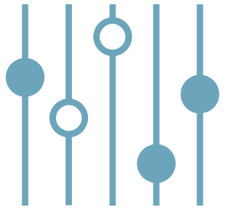
Machine Learning Service

Cloudera Machine Learning

Cloudera Machine Learning is a comprehensive platform to collaboratively build and deploy machine learning capabilities at scale

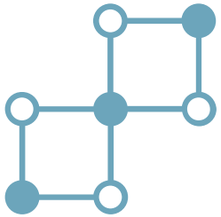
- **Cloudera Machine Learning (CML) is part of the Cloudera Data Platform (CDP)**
  - Enables enterprise data science teams to collaborate across the full data lifecycle
  - Provides immediate access to enterprise data pipelines, scalable compute resources, and access to preferred tools.
  - Optimizes ML workflows across your business with native and robust tools for deploying, serving, and monitoring models
- **To use CML, open a web browser and sign into CDP**
  - Use your organizations single sign-on (SSO) system

Accelerate Machine Learning from Research to Production



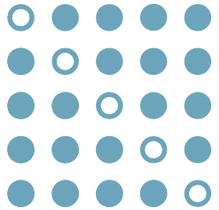
ANALYZE DATA

Explore data securely and share data insights with the team



TRAIN MODELS

Run, track, and compare reproducible experiments



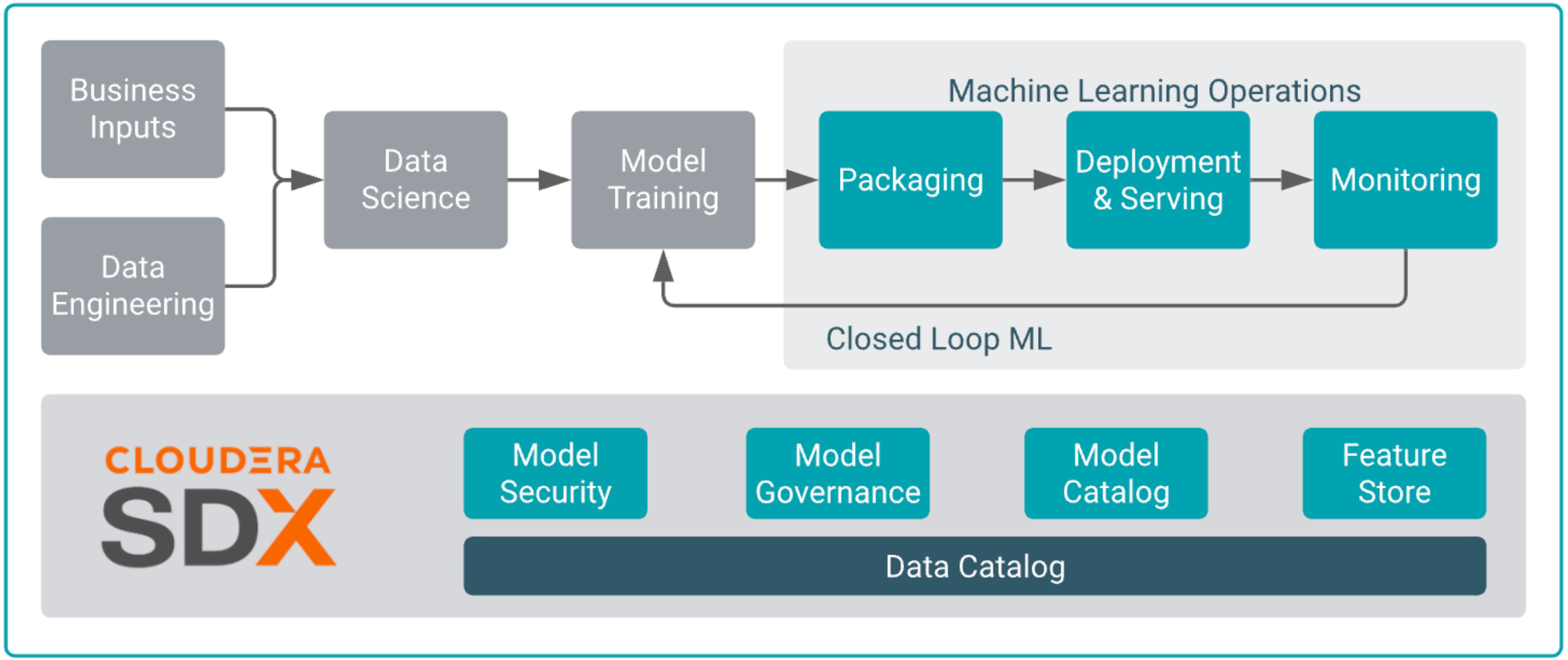
DEPLOY APIs

Deploy and monitor models as APIs to serve predictions

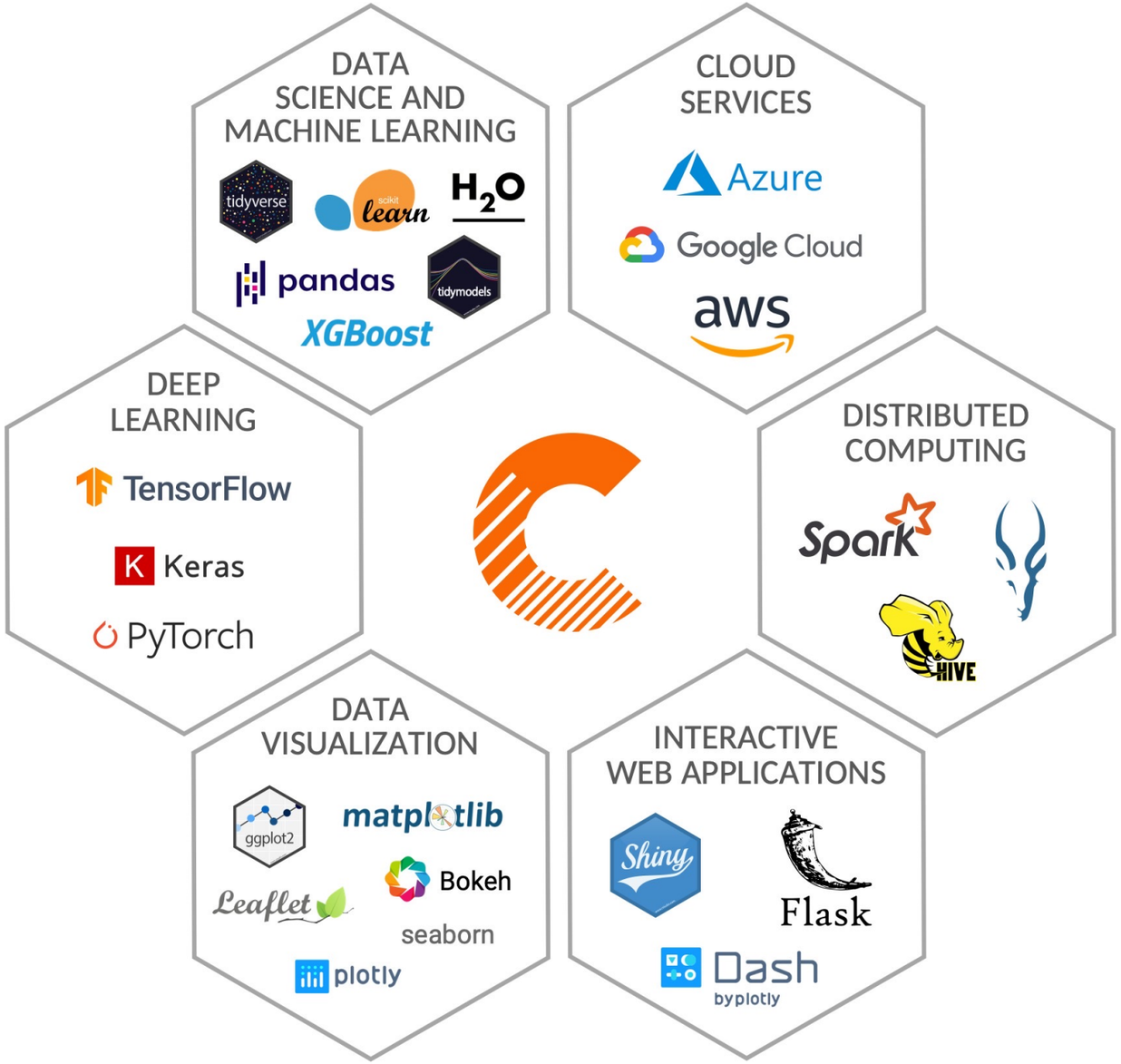
MANAGE SHARED RESOURCES

Provide a secure, collaborative, self-service platform for data science teams

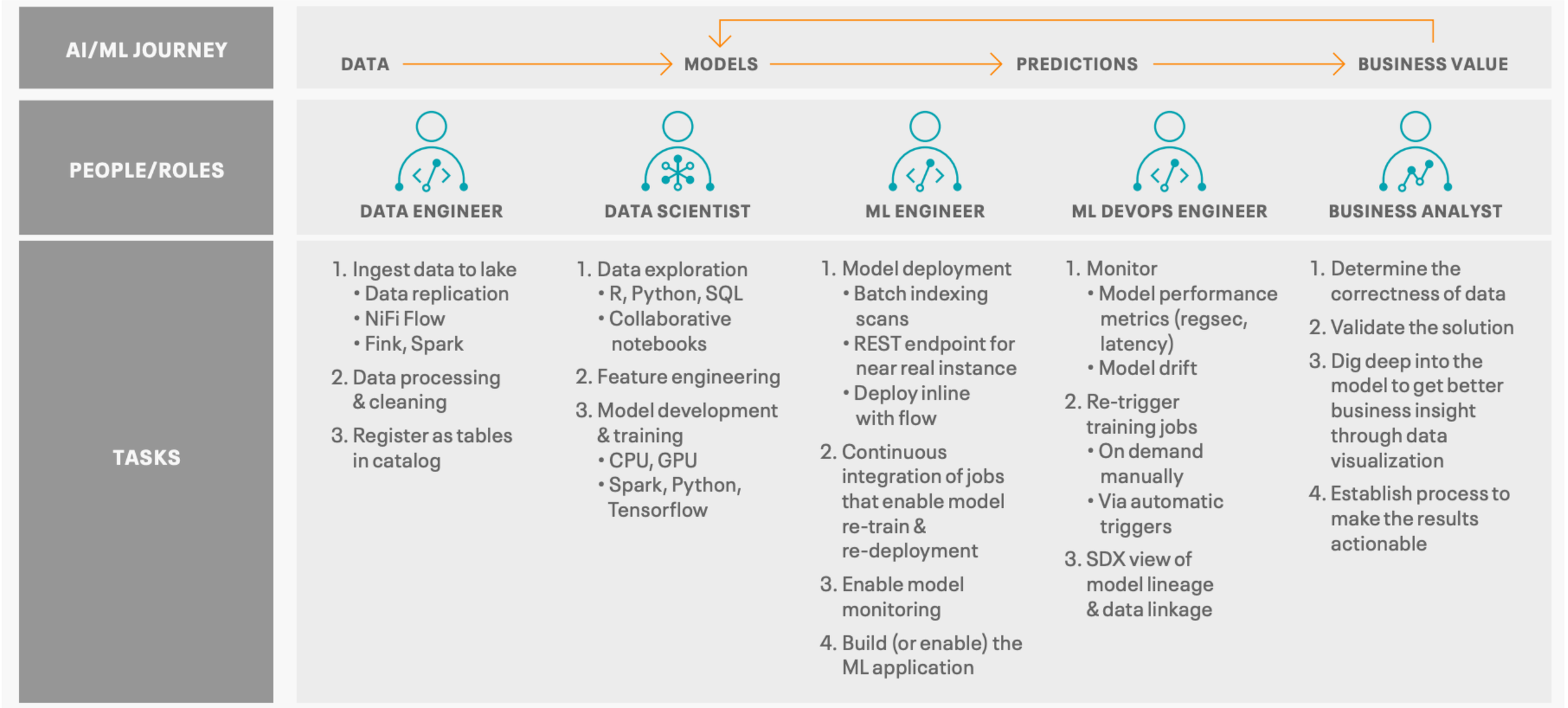
ML Workflow and Operations



What CML Can Help You Do



AI/ML Journey



CML and CDSW

How is Cloudera Machine Learning (CML) related to Cloudera Data Science Workbench (CDSW)?

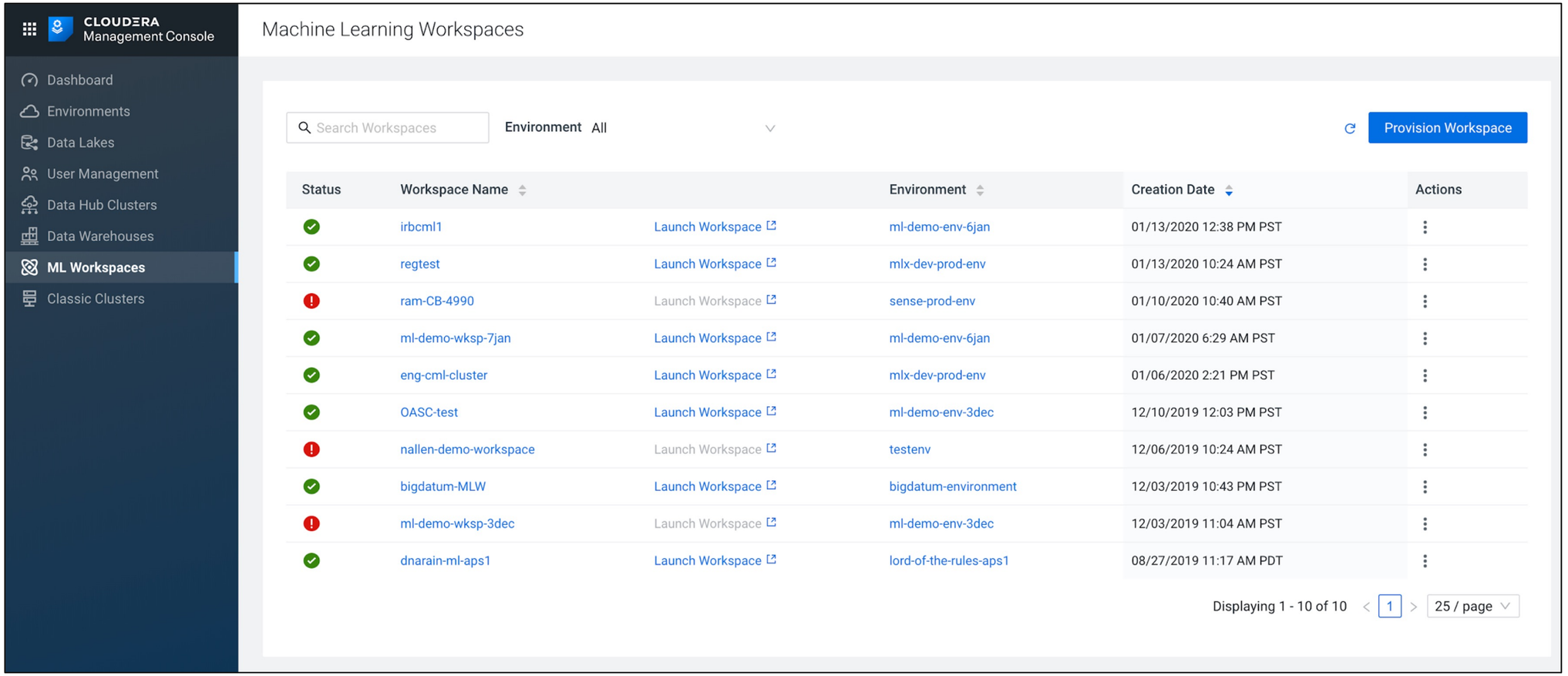
- **CML expands the end-to-end workflow of Cloudera Data Science Workbench (CDSW) with cloud-native benefits such as**
  - Rapid provisioning
  - Elastic autoscaling
  - Distributed dependency isolation
  - Distributed GPU training
- **Both products help data engineers and data science teams be more productive on shared data and compute, with strong security and governance**

Differences Between CML and CDSW

CDSW	CML Private Cloud	CML Public Cloud
Operates in a Kubernetes sidecar connected to CDH/HDP	Operates in the customer's own Kubernetes-based private cloud (e.g. OpenShift)	Operates in the customer's public cloud, using EKS in AWS and AKS in Azure
No autoscaling; resources are dedicated to CDSW	Shared resource pool; resources limited by the private cloud environment	Cloud-based autoscaling
Distributed compute workloads are pushed to the CDH/HDP cluster (Spark-on-YARN)	Workloads run on the dedicated Kubernetes cluster separate from CDP-CD (Spark-on-K8S)	Workloads run on the dedicated Kubernetes cluster separate from CDP (Spark-on-K8S)
Close to *End-of-Life*	Where all the R&D goes	Where all the R&D goes

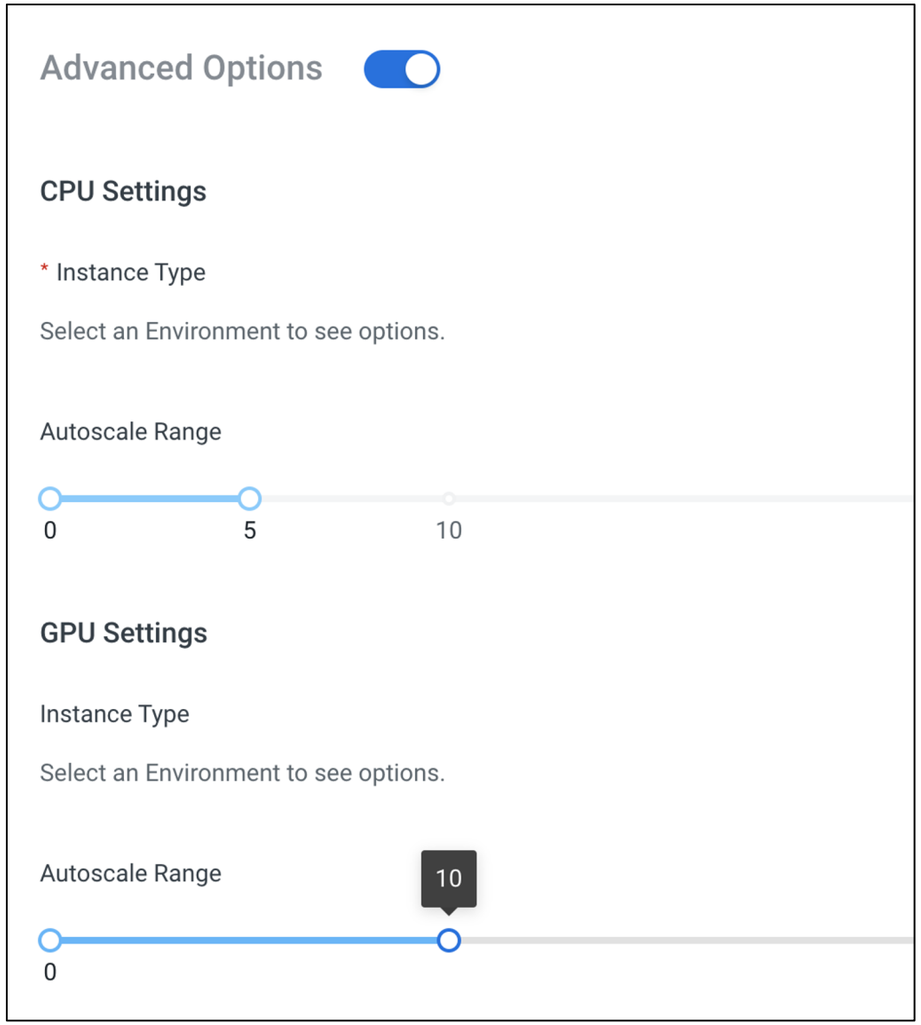
ML Workspaces

- **Each ML workspace enables collaborative teams of data scientists to**
  - Develop, test, train, and deploy machine learning models
  - Build predictive applications that can be used withing the organization

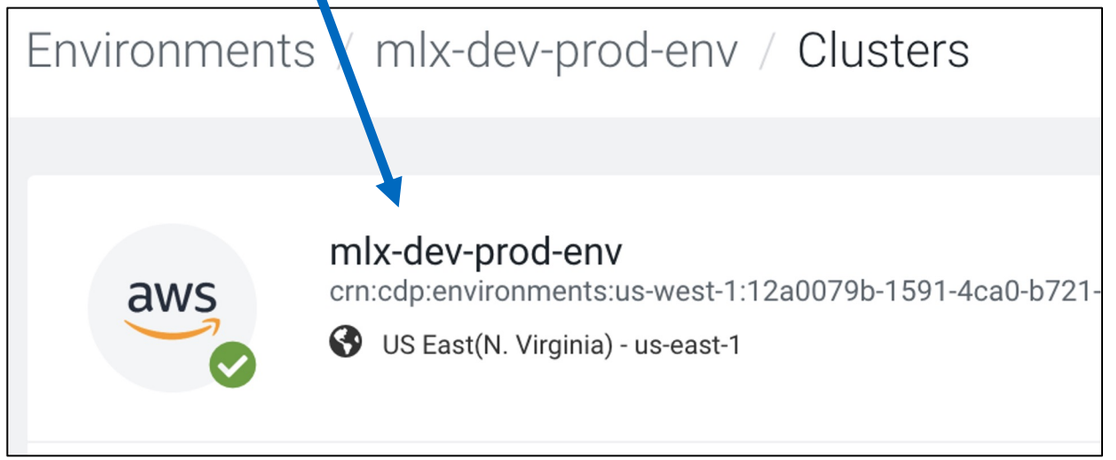


Elastic Kubernetes Cluster

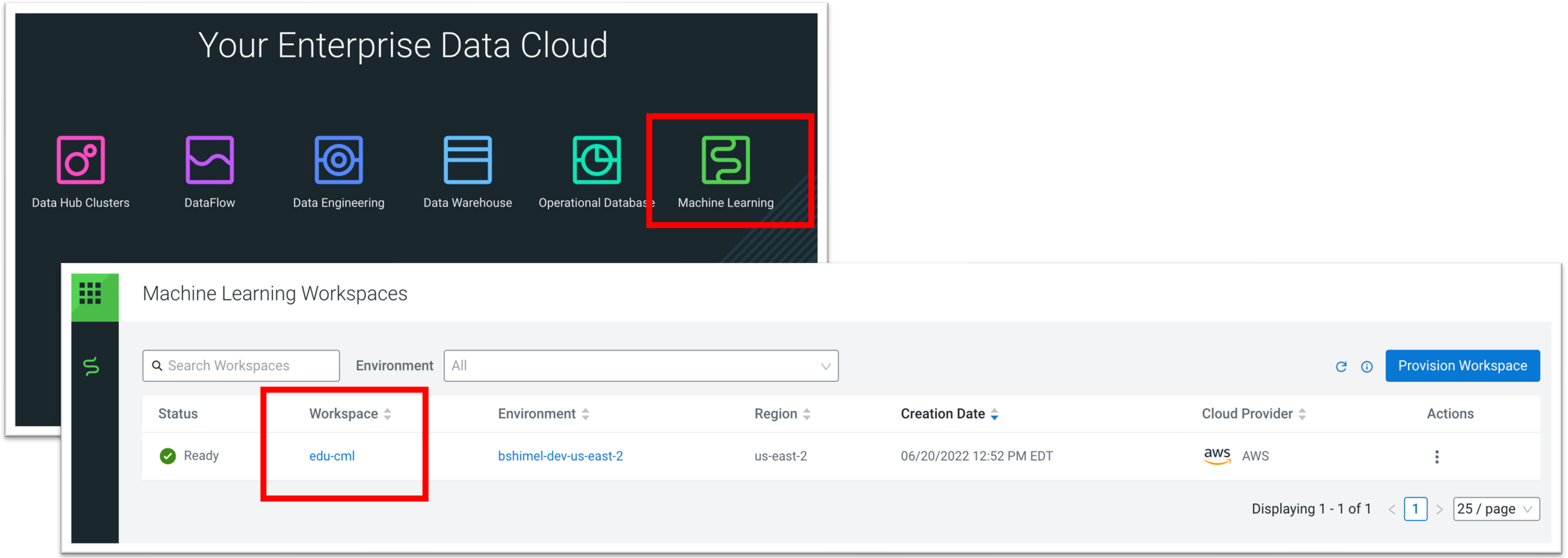
- **Each ML Workspace is an elastic Kubernetes cluster**



Created within cloud environments



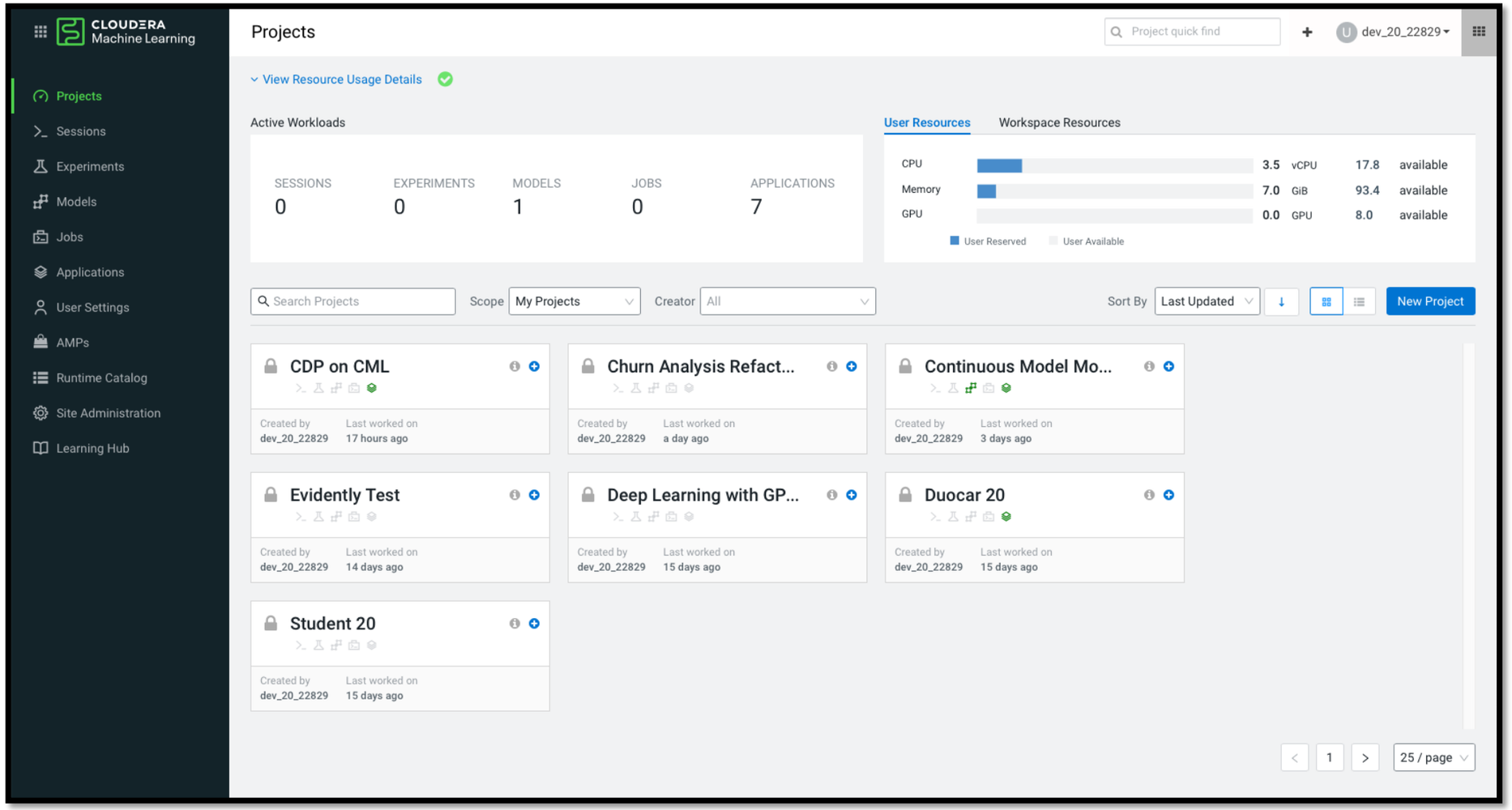
Accessing a Workspace



- Log in to the CDP web interface https://console.cdp.cloudera.com
- Click **Machine Learning**
- Click the link for desired workspace

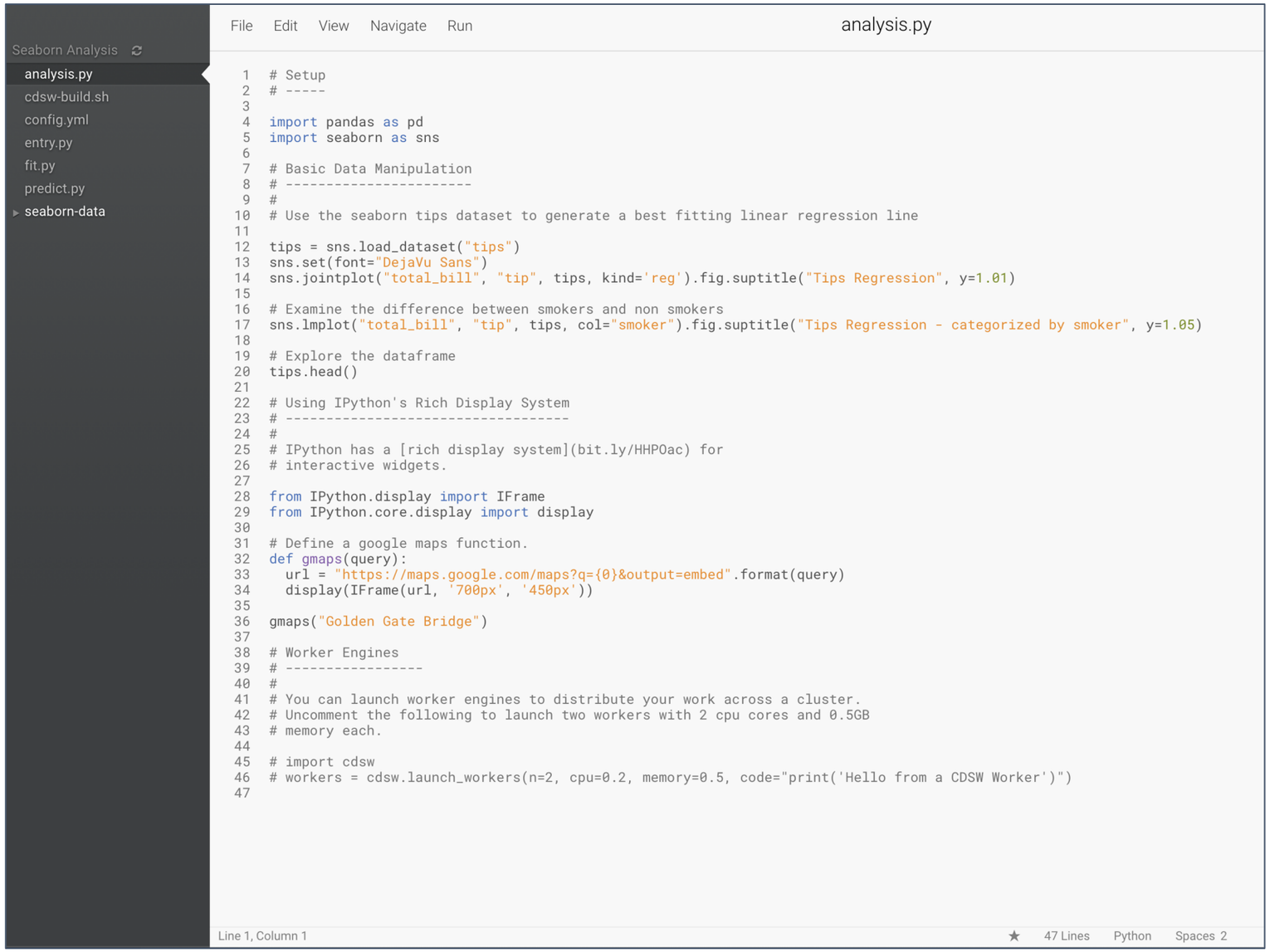
Projects

- **One or more projects are created within an ML Workspace**

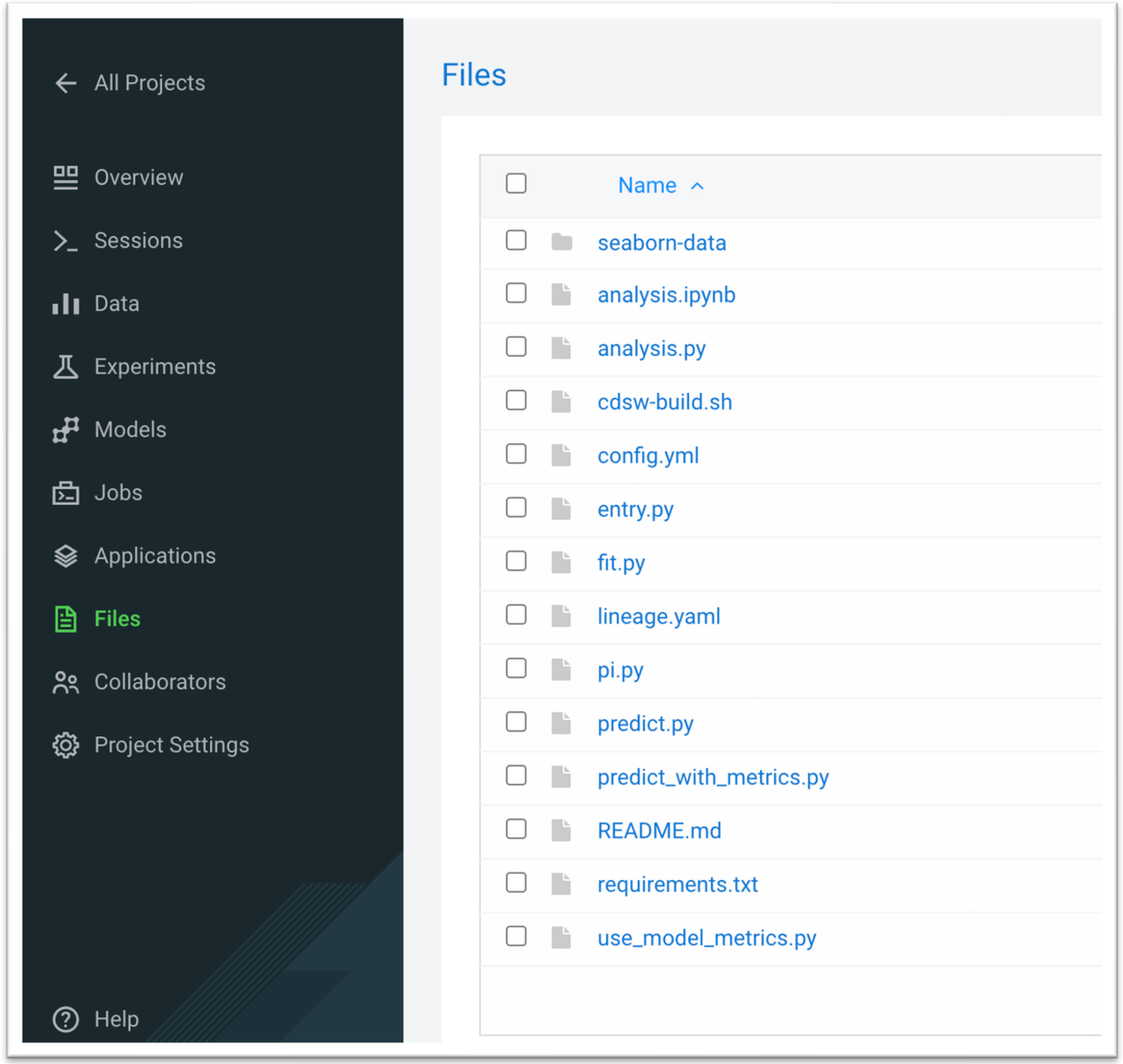


A Project Contains...

- **A project is a directory structure containing code, data, and assets**
- **Project code is commonly written in Python, R, or Java**



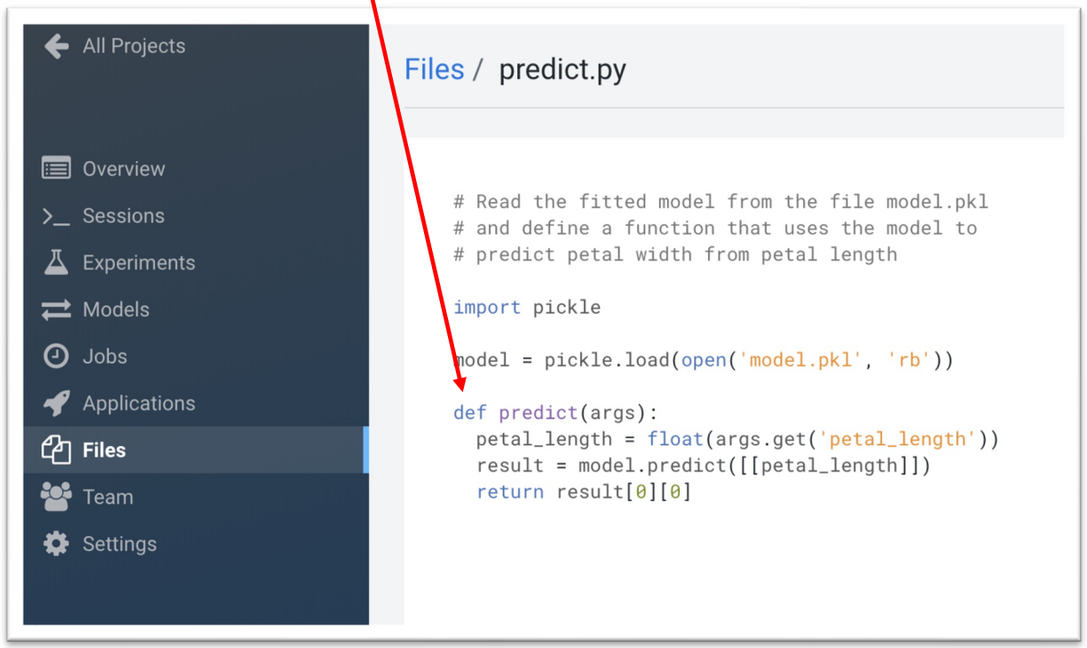
Managing Project Files



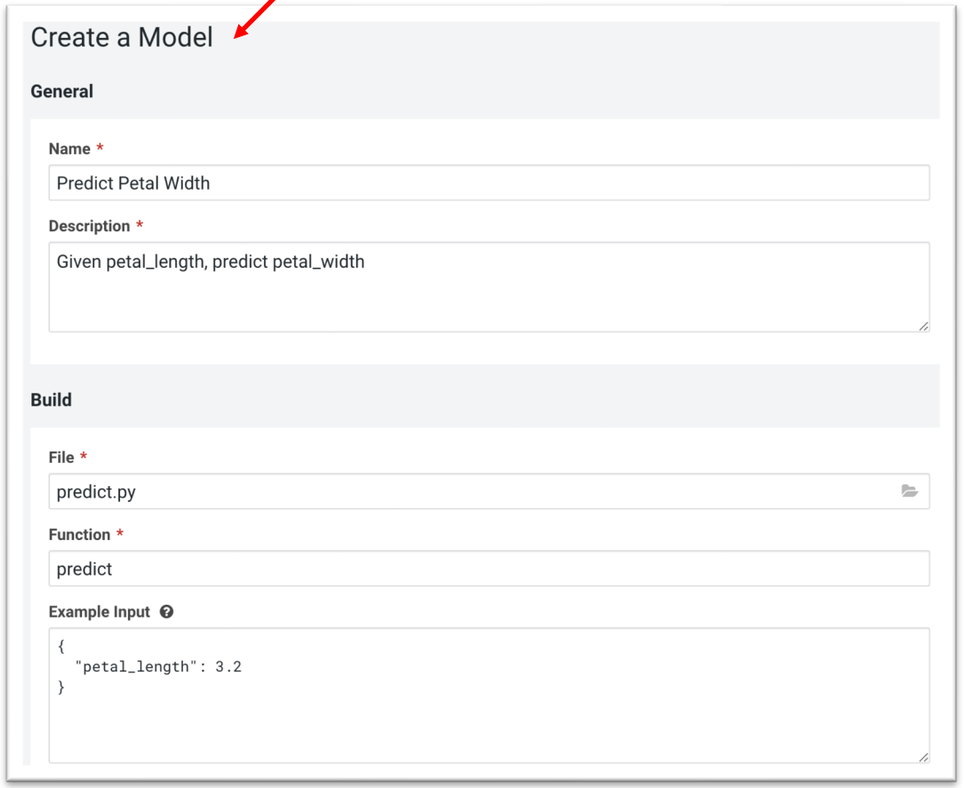
- **Move, rename, copy, and delete files within the scope of the project**
  - Upload new files to a project
  - Download project files
- **Be careful about deleting files and directories in CML!**
  - There is no way to undo the deletion

Models Are Deployed Within a Project

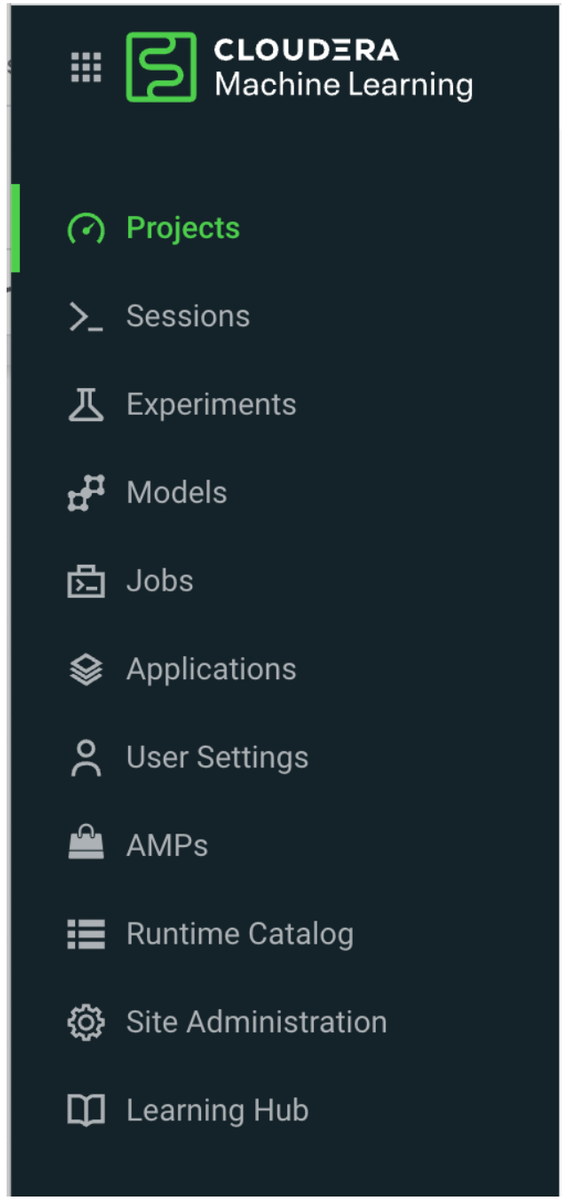
Write the function that makes the prediction…



…and then use CML to serve it up



Project Components



- **Sessions**
  - Directly leverage the CPU, memory, and GPU compute available across the workspace
- **Experiments**
  - Run multiple variations of model training workloads in order to train the best possible Model
- **Models**
  - Implementation of machine learning algorithms based on sample data
- **Jobs**
  - Orchestrate an entire end-to-end automated pipeline
- **Applications**
  - Deliver interactive experiences for business users

User Roles

- **If a user has more than one role**
  - Role with highest level of permissions takes precedence
- **If a user is a member of more than one group**
  - Can gain additional roles

Environment Roles

- **Access to CML workspaces within a given CDP environment**

MLAdmin

- Create and delete ML workspaces
- Administrator-level access to all workspaces in the environment
  - Run workloads
  - Monitor and manage user activity
  - Also needs the account-level role of IAMViewer in order to list users or assign resource roles

MLUser

- List ML workspaces provisioned
- Run workloads on all workspaces provisioned

MLBusinessUser

- View applications deployed under the projects they have been added to

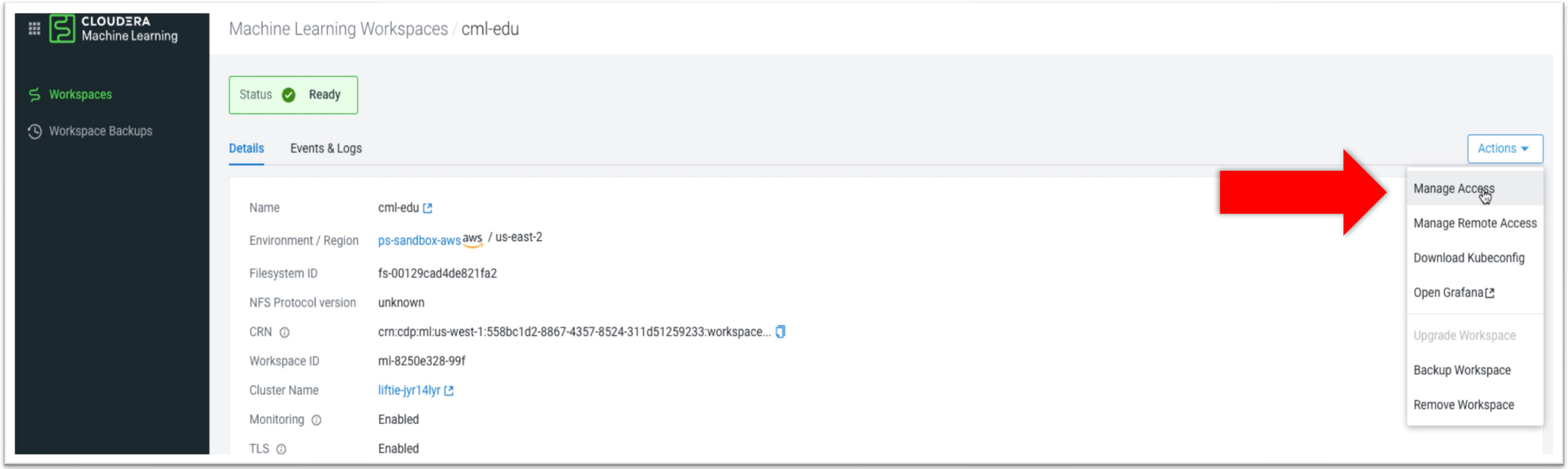
Workspace Roles

- **Access to a <u>specific</u> CML workspace within a given CDP environment**

MLWorkspaceAdmin	• Manage all machine learning workloads and settings inside a specific workspace. • Also needs the account-level role of IAMViewer in order to list users or assign resource roles.
MLWorkspace BusinessUser	• View shared machine learning applications insides a specific workspace
MLWorkspaceUser	• Run machine learning workloads inside a specific workspace

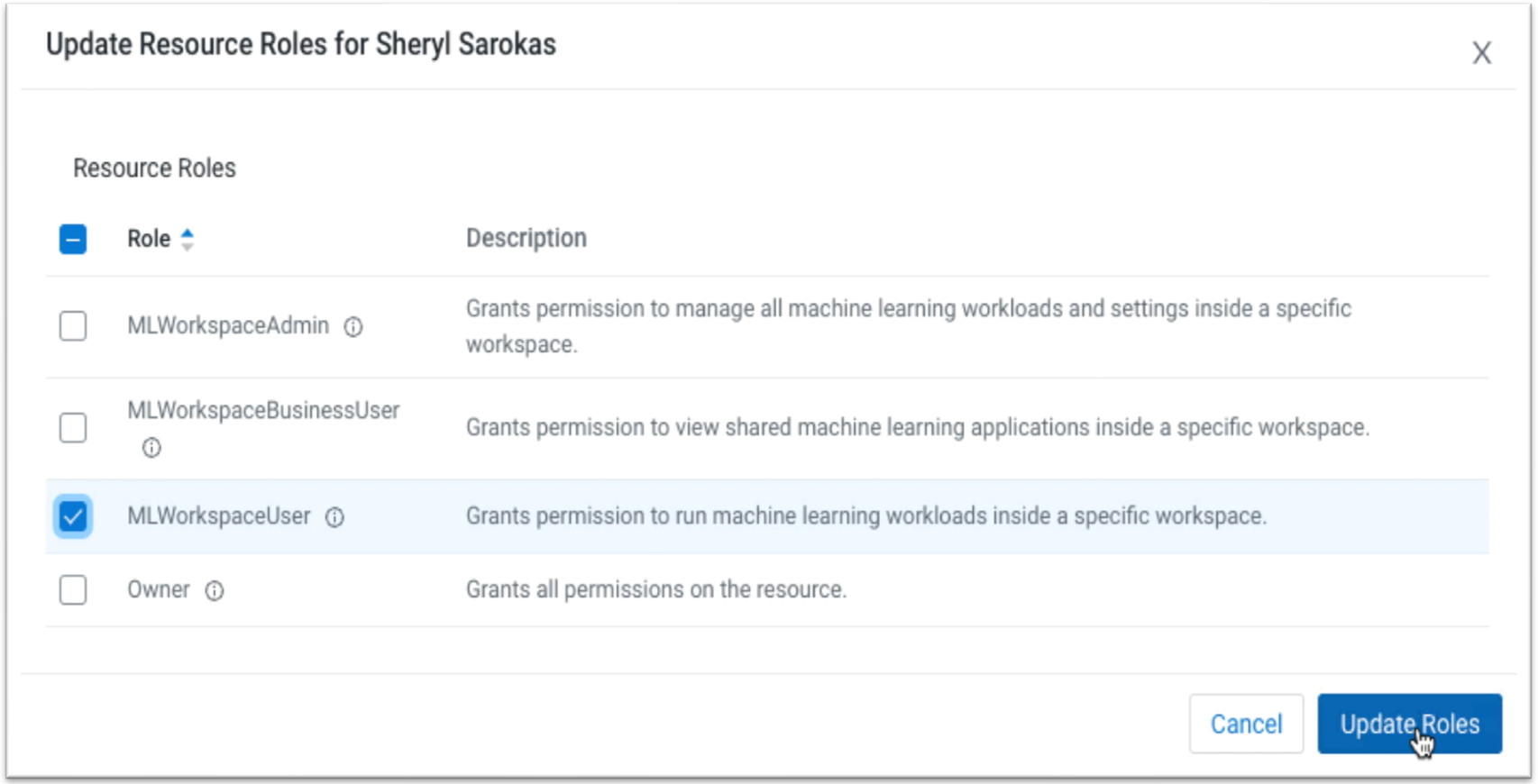
Configuring User Access

- **In the workspace, select Manage Access from the Actions menu**



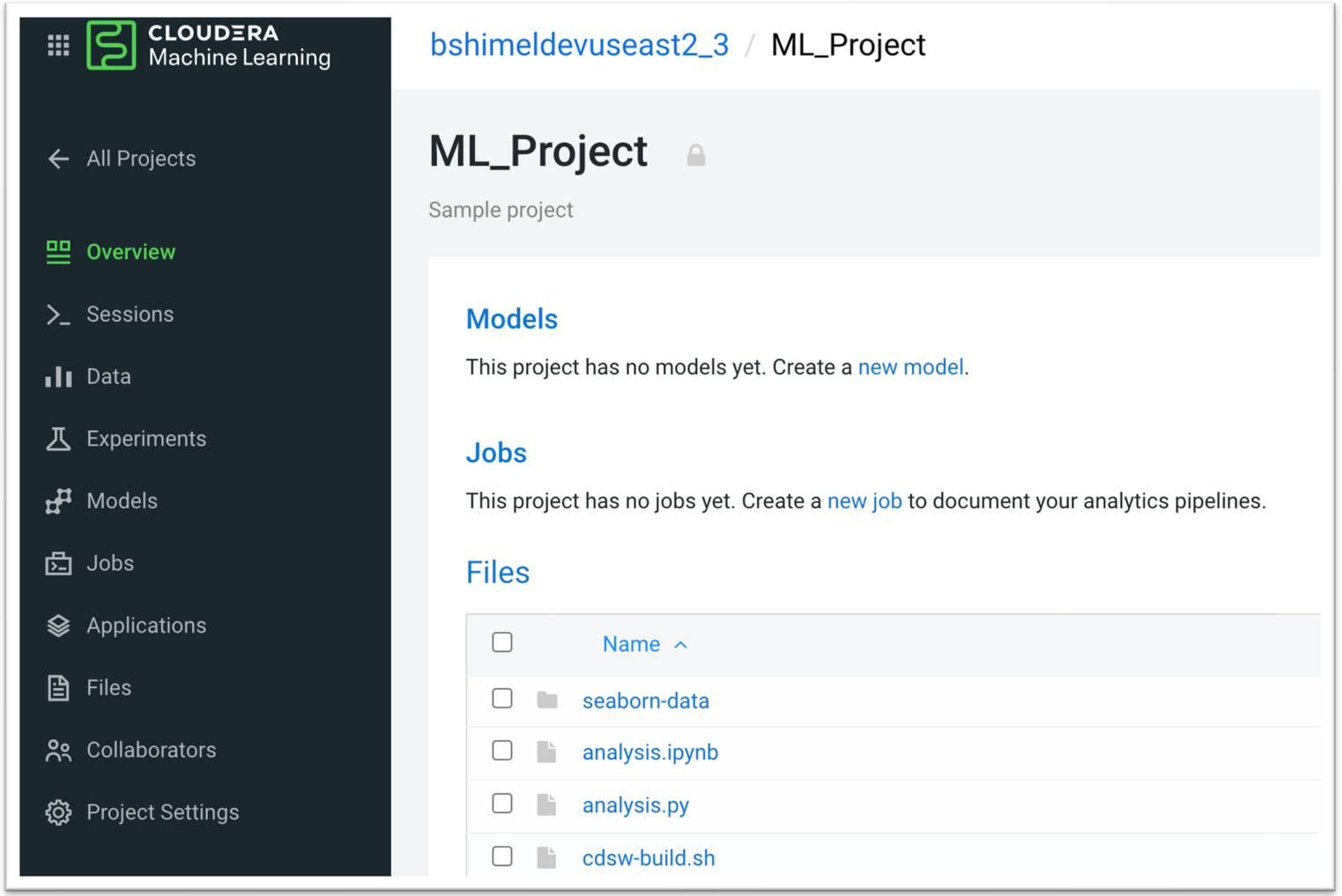
Updating a User's Roles

- **Select the user to add or remove a role**



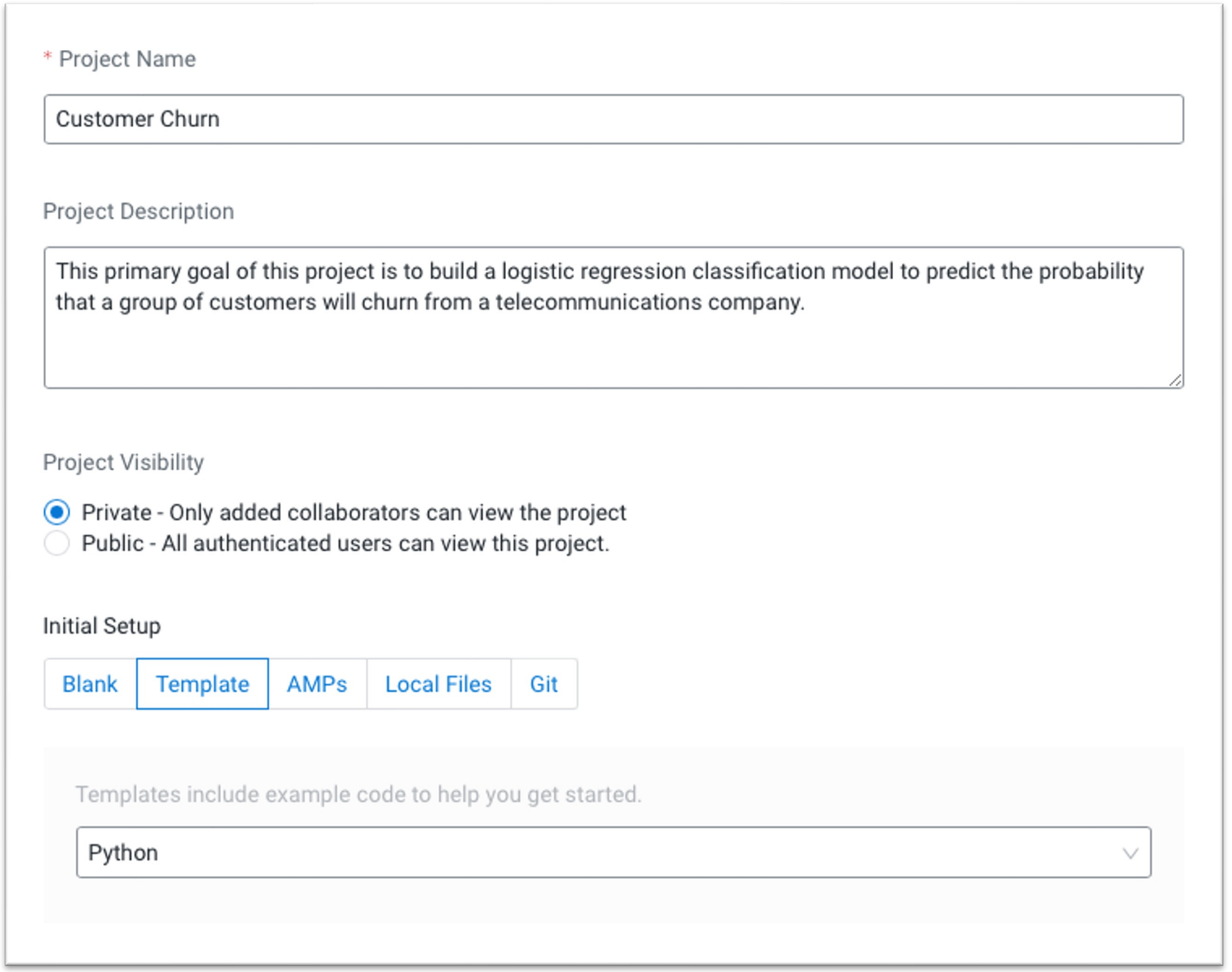
Projects Are Independent

- **Users can work freely without interfering with one another or breaking existing workloads**
- **Allows for collaboration with other users**
  - Teams can be created to access the project
  - Individual users can be added to private projects

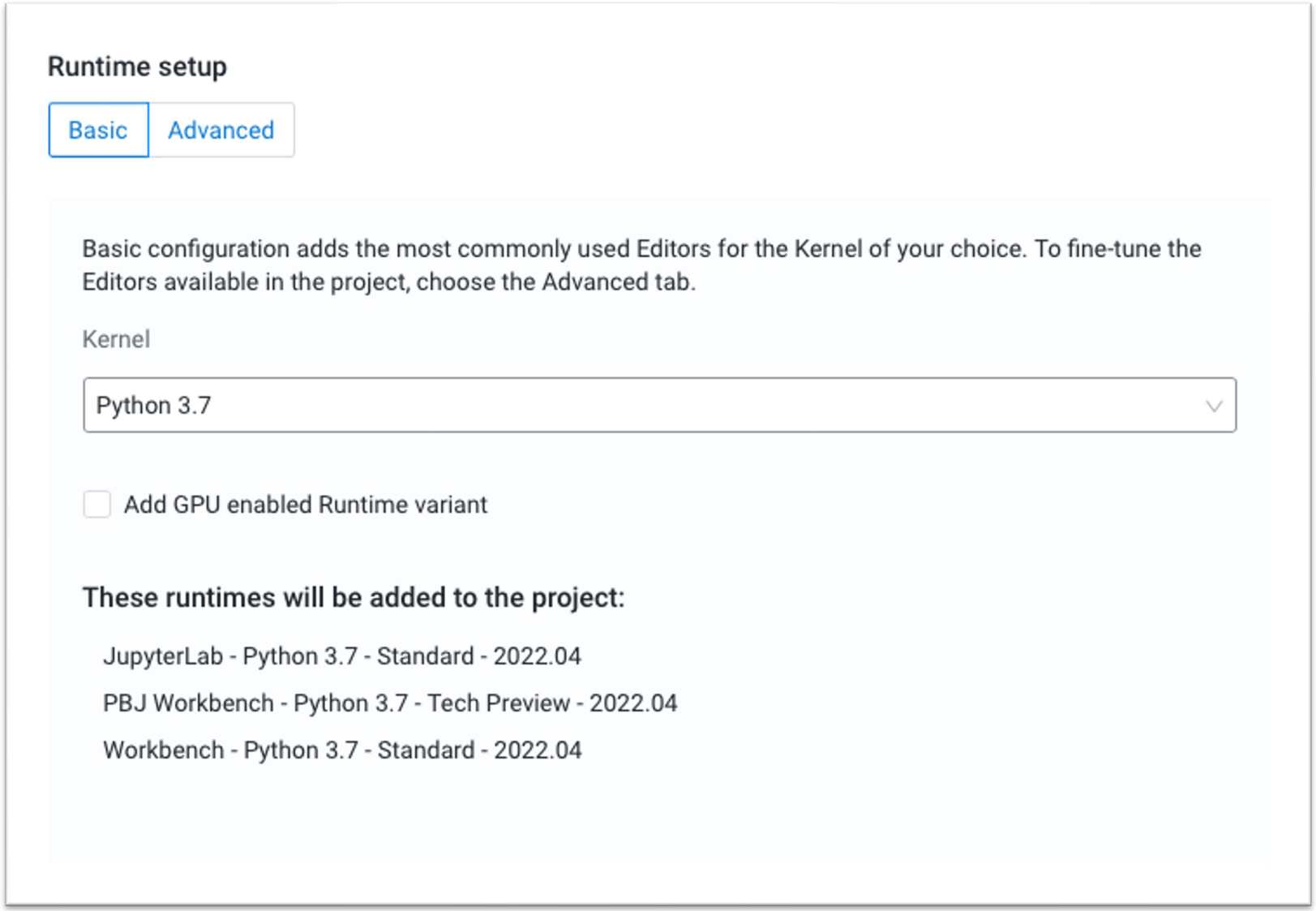


New Project

- **Name**
- **Description**
- **Visibility**
  - Private
  - Public
- **Initial Setup**
  - Blank
  - Template
  - AMPs
  - Local file
  - Git

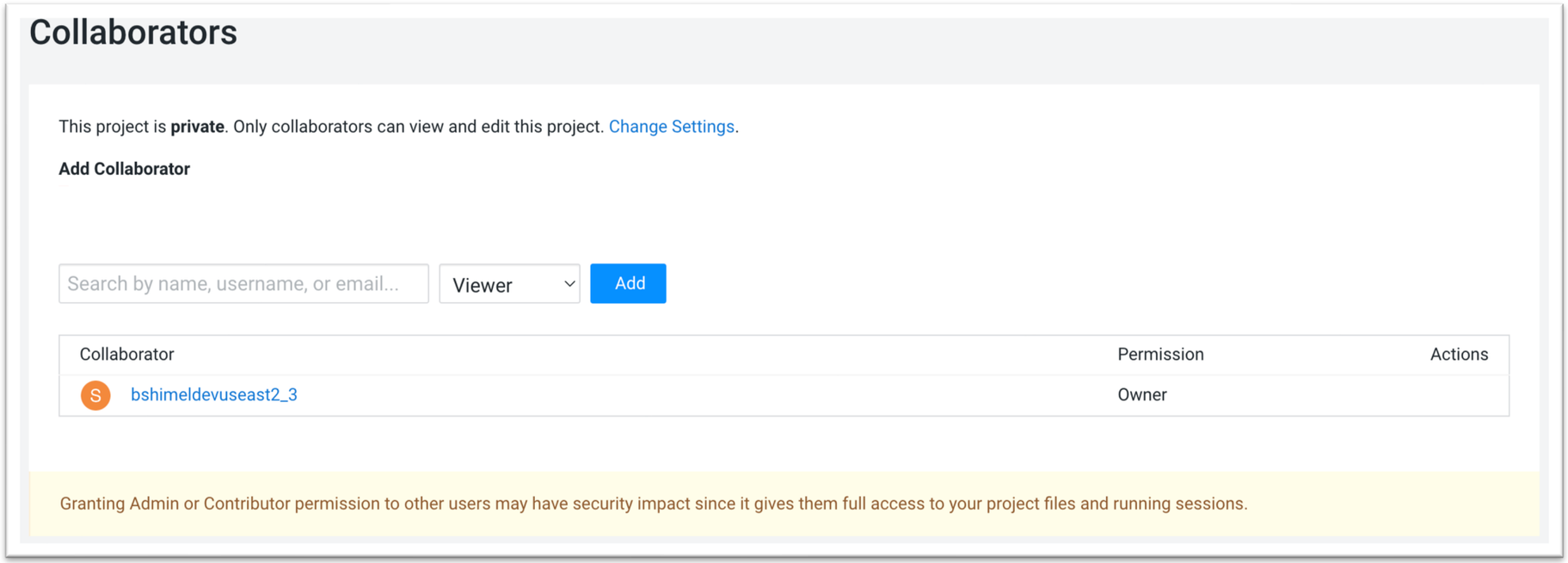


New Project Runtime Setup



Project Visibility

- **When you create a project in your personal context, specify one of the following visibility levels to the project**
  - **Public**: grant read-level access to *everyone* with access to the Cloudera Machine Learning application
  - **Private**: you must explicitly add someone as a project collaborator to grant them access
- **To work with colleagues on a project, add them to the project as a collaborator**



Chapter Topics

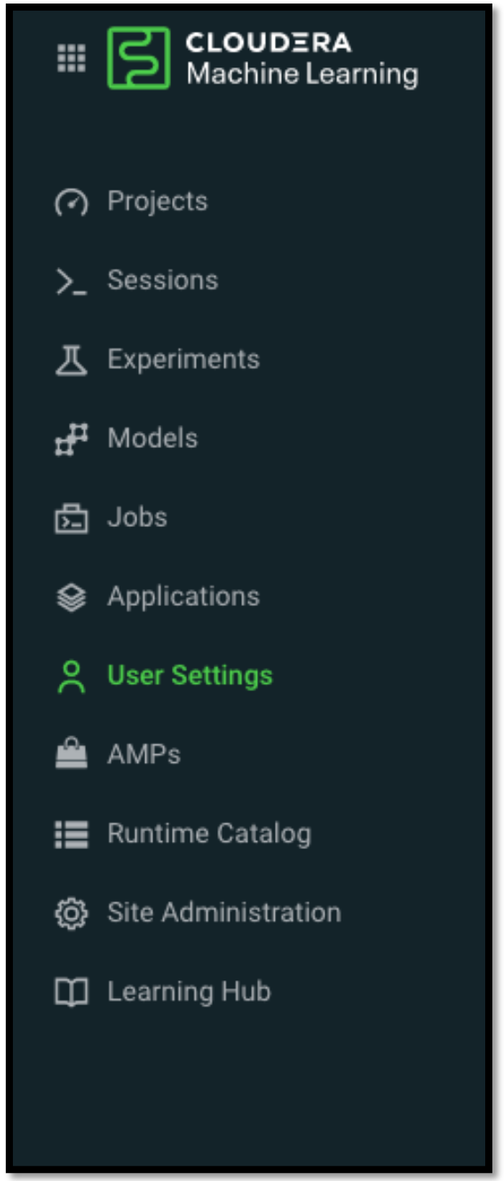
Introduction to CML

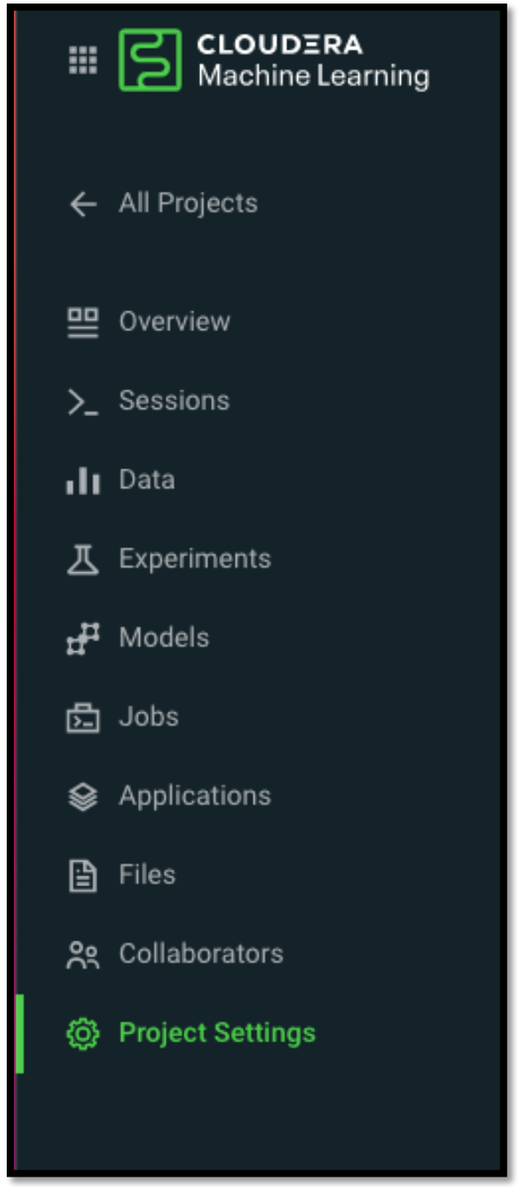
- Overview
- CML Versus CDSW
- ML Workspaces
- Workspace Roles
- Projects and Teams
- **Settings**
- Runtimes/Legacy Engines
- Exercise Overview
- Essential Points
- Hands-On Exercise: Getting Started with CML

Settings

- **Located in CML menu**
- **Are context sensitive**

Workspace

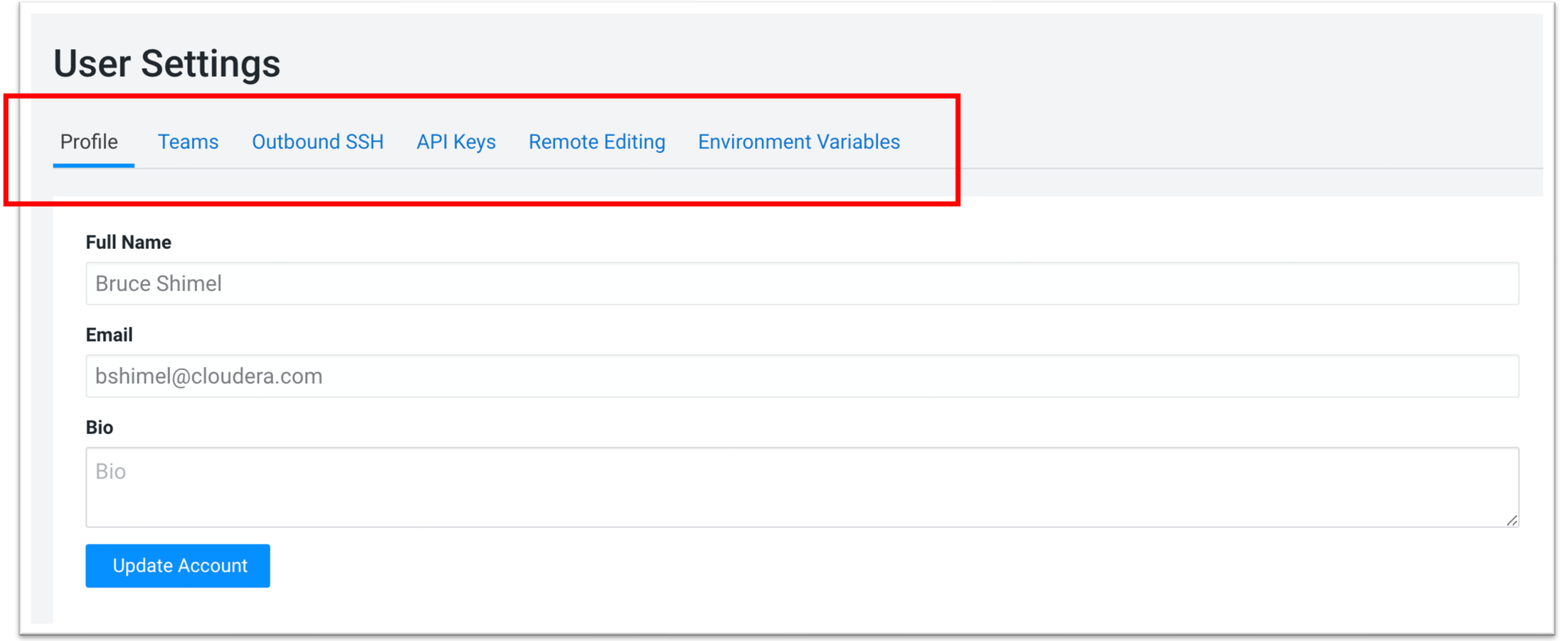




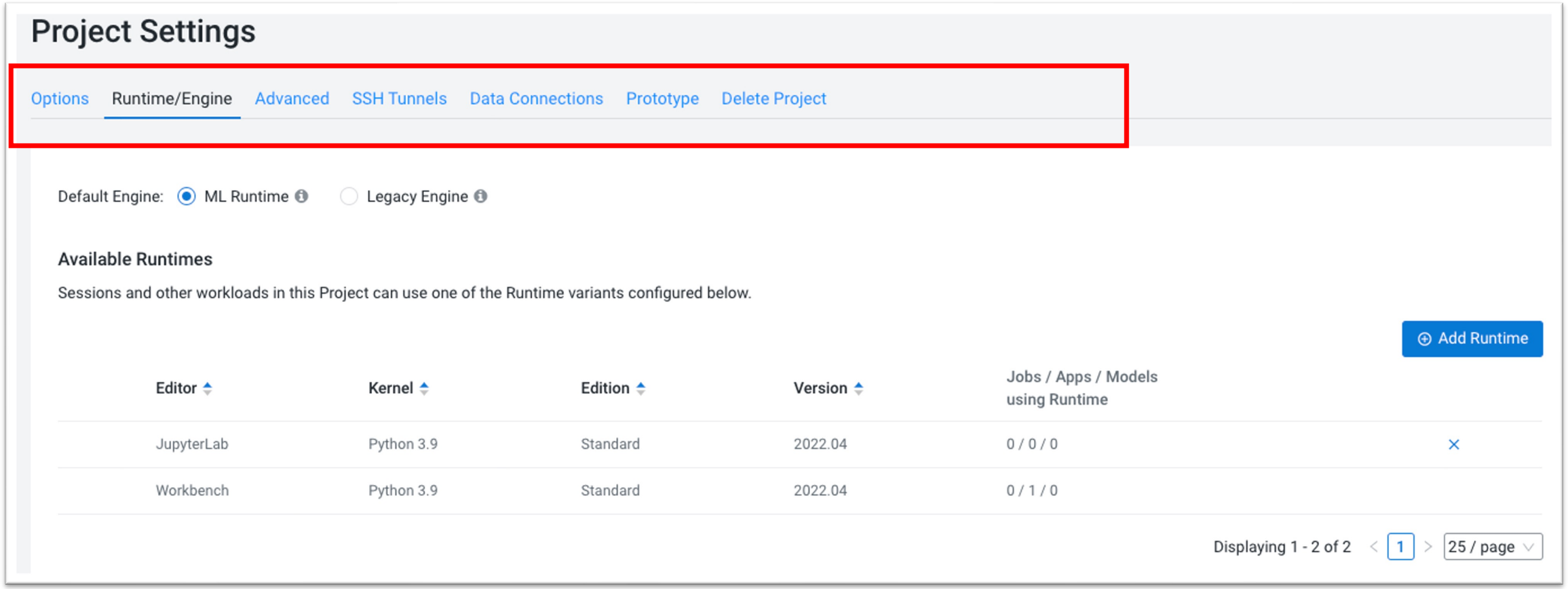
Project

User Settings

- **Specific to the User within the Workspace**

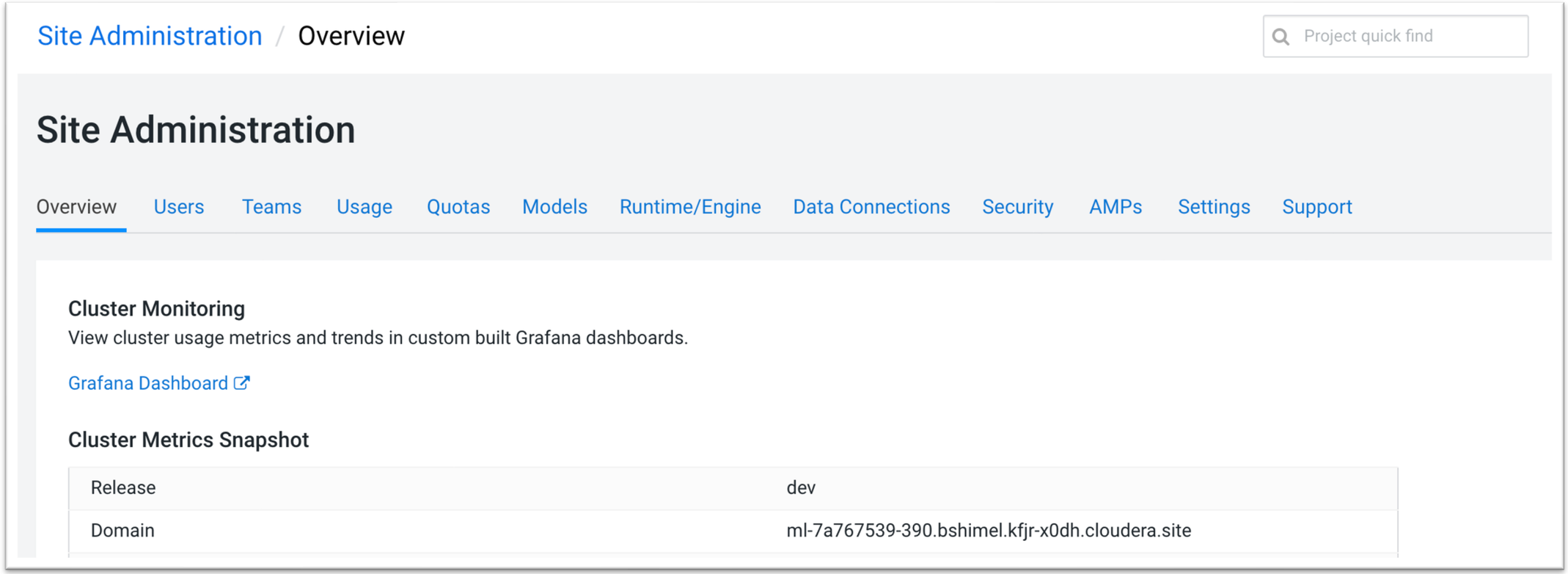


Project Settings in CML



Site Administration

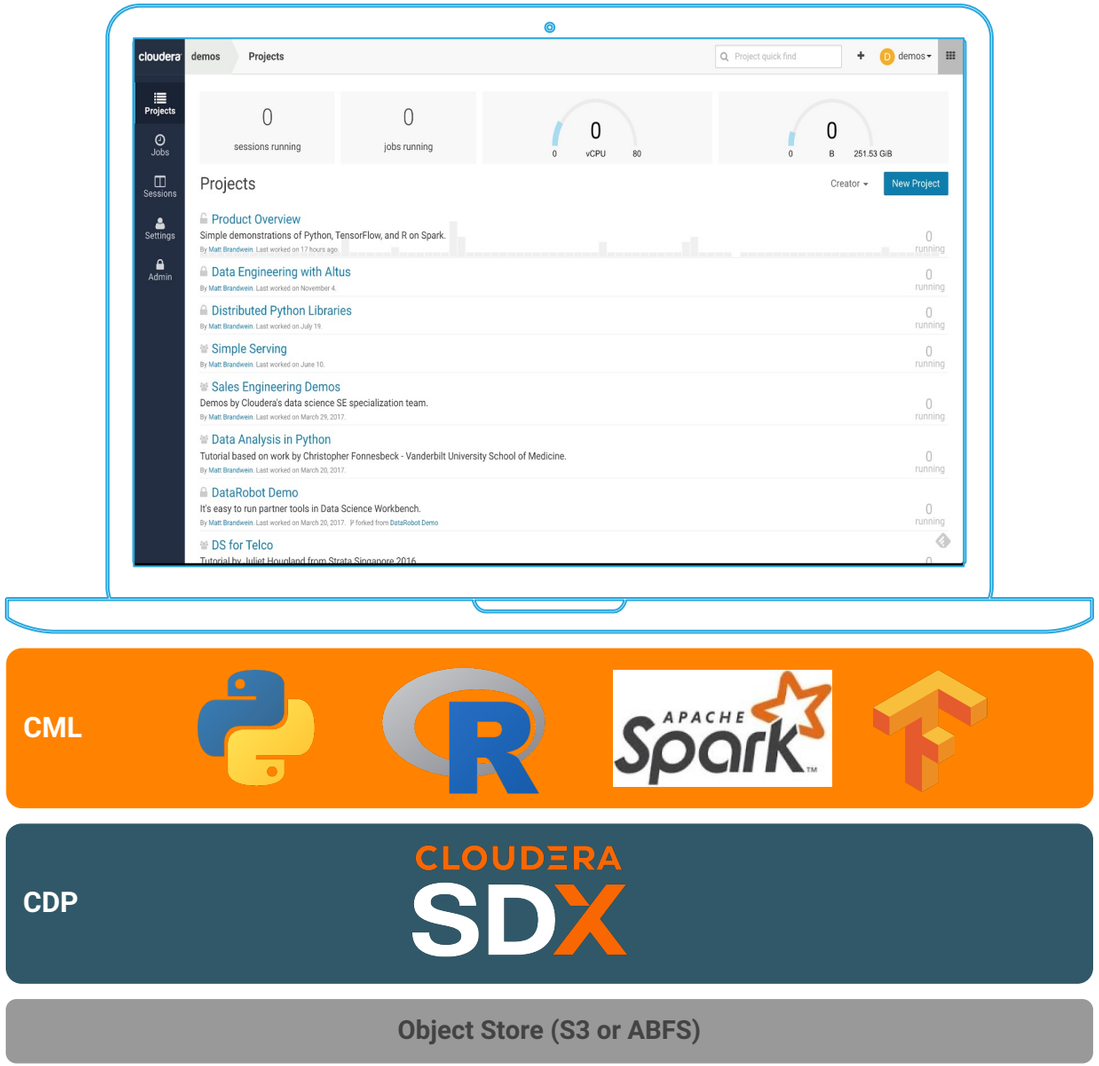
- Site administrators can manage the entire workspace
  - Monitor and manage all user activity across a workspace
  - Add new custom engines
  - Configure certain security settings



ML Runtimes

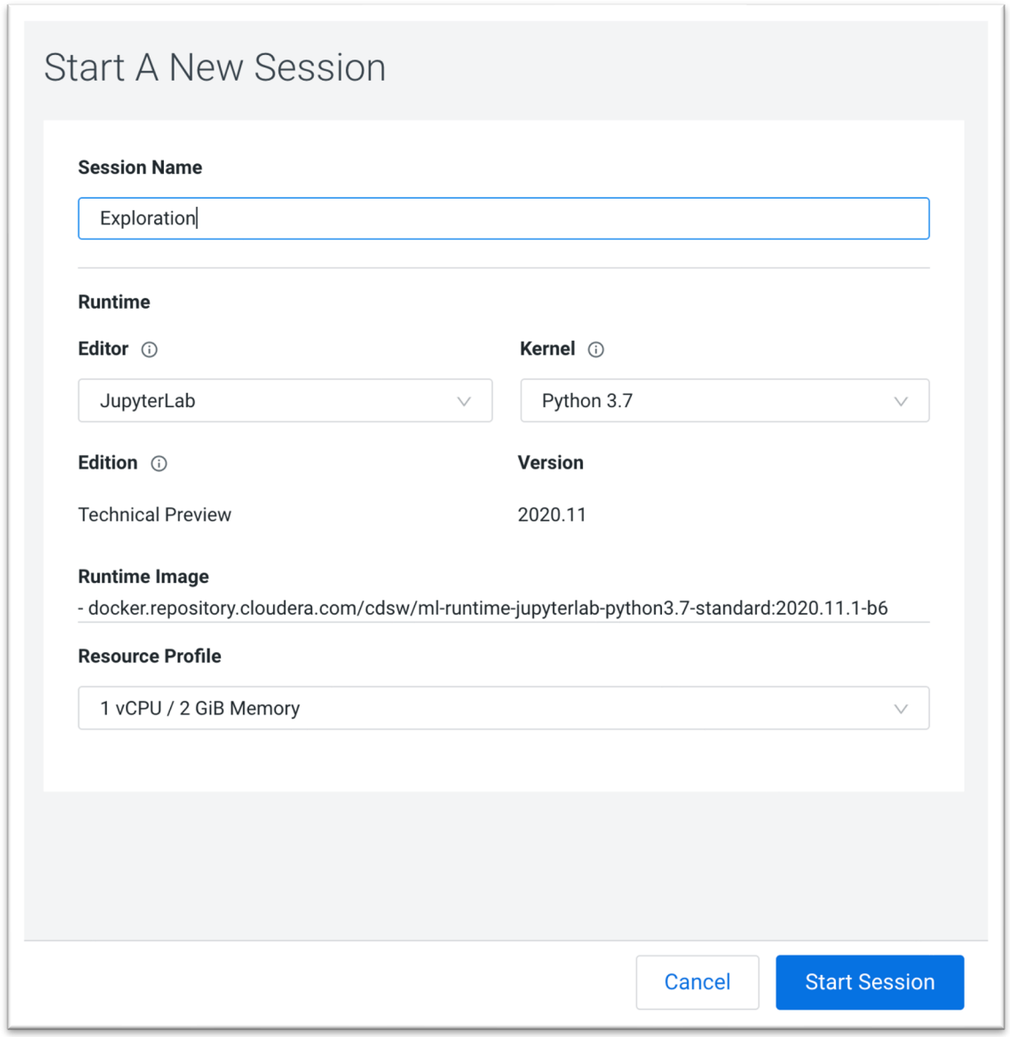
- **Responsible for running the code written by users and intermediating access to the data**
- **Keeps the images small and improves performance, maintenance, and security**
- **Similar to a virtual machine**
  - Container images that contain the Linux OS, interpreter(s), and libraries
  - Customized to have all the necessary dependencies to access the computing cluster
  - Keeps each project's environment entirely isolated

Cloudera ML Runtimes



- **<u>Isolated</u>**, **<u>containerized</u>** working environment for our end users
- Core feature to enable self-service data science
  - Direct access to data
  - On-demand resources
  - Ability to install and use any library, framework without IT assistance.
- Out of the box support
  - Editors: Workbench, JupyterLab
  - Kernels: Python 3.7-3.9, R 3.6, 4.0-4.1, Scala 2.11
  - Editions: Standard, NVIDIA GPU

Runtime Environments Enable Flexibility



Editor

Development interface to write and execute code

Examples: Workbench, JupyterLab

Kernel

Execution engine for the session of work

Examples: Python 3.7, Python 3.8, R3.6

Edition

Set of capabilities (tools/libraries) available for the run

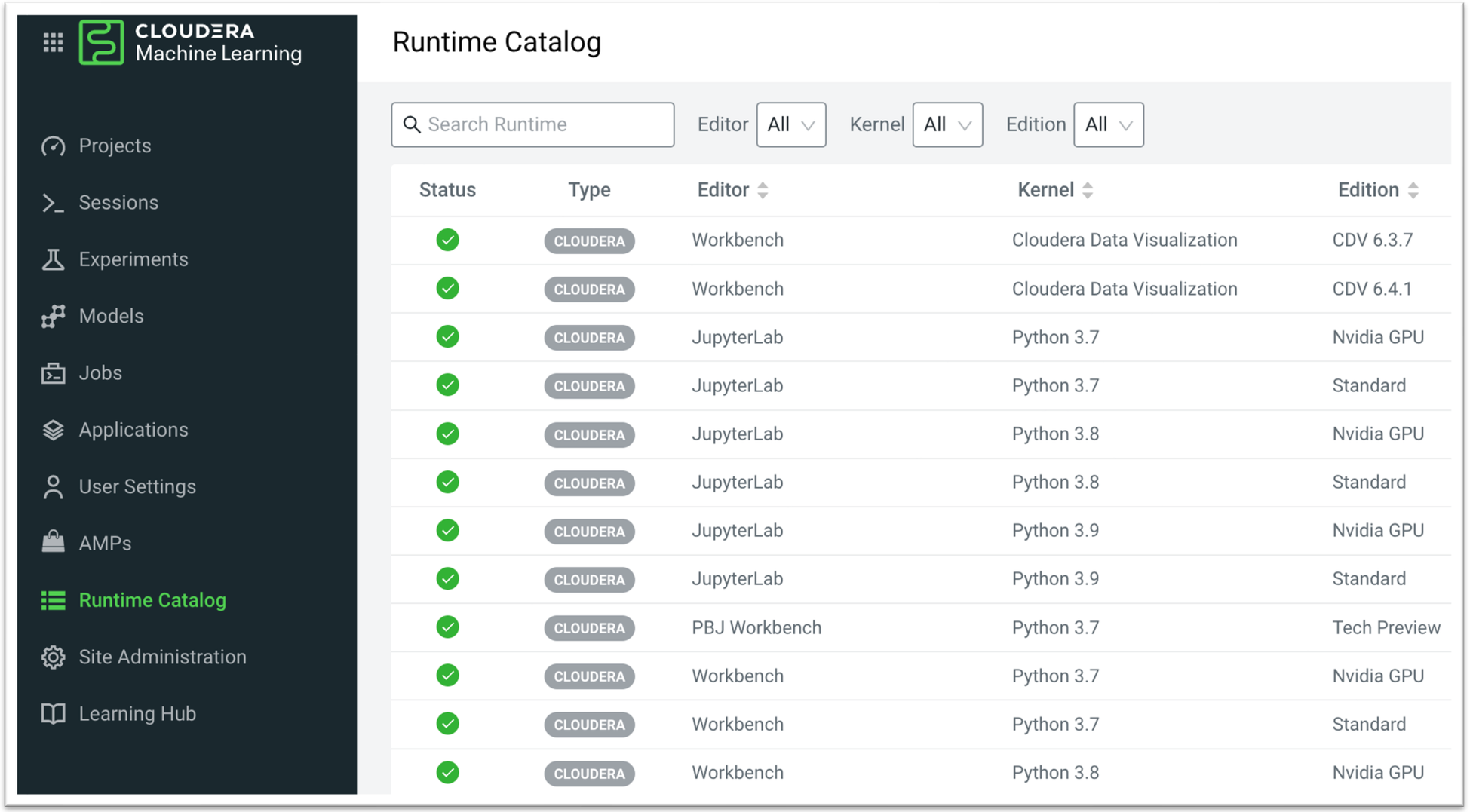
Examples: Standard, RAPIDS

Version

Major version of the runtime

Example: 2022-04

ML Runtime Catalog



ML Runtimes Versus Legacy Engines

- **Runtimes and the Legacy Engine have the same purpose**
  - Container images that contain the Linux OS, interpreter(s), and libraries
  - The environment in which your code runs
- **ML Runtimes**
  - Small image
  - Contains a single interpreter and UNIX tools
  - Improved performance, maintenance, and security
  - Recommended to use for all new projects
- **Legacy Engine**
  - Huge image
  - Contains four Engine interpreters (Python 2, Python 3, R, Scala), and UNIX tools
  - Existing Engine-based projects can be migrated to ML Runtimes

Essential Points

- **Cloudera Machine Learning (CML) is part of the Cloudera Data Platform (CDP)**
  - Analyze data
  - Train models
  - Deploy APIs
- **An ML Project**
  - Contains all the code, configuration, and libraries needed to reproducibly run analyses
  - Is independent, ensuring users can work freely without interfering with one another or breaking existing workloads
  - Allows for collaboration with other users
- **The Settings link in the left menu is context-sensitive and provides configuration for a user, a project, or a team**
- **ML Runtimes are responsible for running the code and intermediating access to the data**

Hands-On Exercise: Getting Started with CML

- **In this exercise, you will**
  - Login to the Cloudera Data Platform exercise environment,
  - View the Cloudera Machine Learning workspace
  - Create a new CML project
  - Create a new session
  - Delete a CML project
- **Please refer to the Hands-On Exercise Manual for instructions**

CLOUDERA

Introduction to AMPs and the Workbench

Introduction to AMPs and the Workbench

By the end of this chapter, you will be able to

- **Run code from within a session**
- **Use Git for version control**
- **Create ML web applications/dashboards and easily share them with other business stakeholders**
- **Install and run AMPs that provide reference example machine learning projects in CML**

Native Workbench Console and Editor

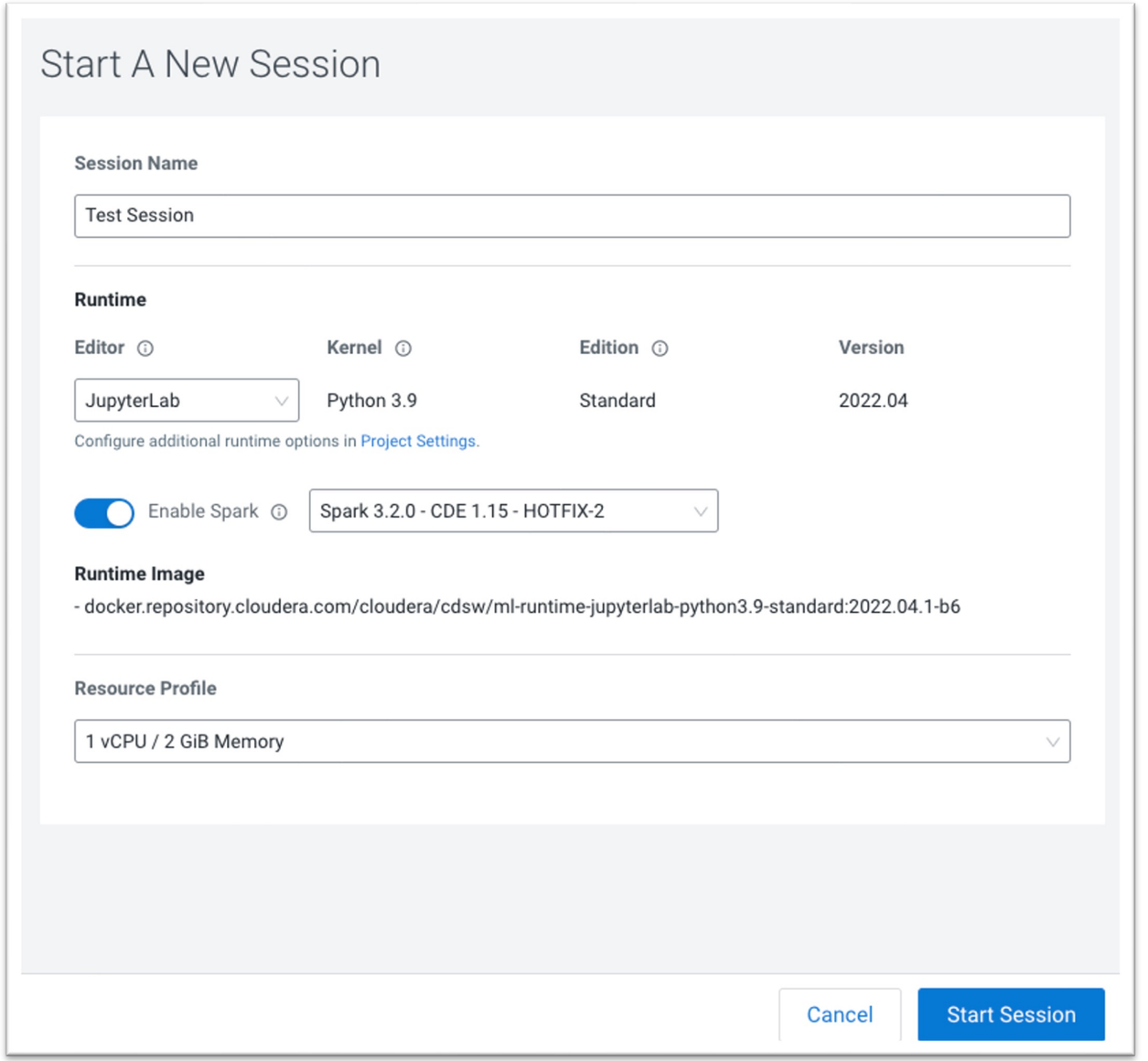
- **Interactive environment tailored specifically for data science**
- **UI includes**
  - An editor where you can edit your scripts.
  - A console where you can track the results of your analysis
  - A command prompt where you can enter commands interactively
  - A terminal where you can use a shell

Working in Workbench

1. **Add code to your project**
   - Create and edit your code
   - Upload existing code files
   - Specify what compute resources you need
2. **Launch a session using Python, R, or Scala**
3. **Run code**
   - From your code files
   - At the session prompt
   - Or run commands in the terminal
4. **Stop a session**
   - Or have it timeout

New Session

- **Editor**
  - Selects the Editor, for example JupyterLab, Workbench
- **Kernel**
  - Selects the Kernel, for example Python 3.7, R4.0
- **Edition**
  - Selects the Runtime Edition
  - Initially only Standard variants are supported
- **Version**
  - Selects the ML Runtimes version



ML Editor

- **To run code from the editor**
  - Select a script from the project files on the left sidebar
  - Click the arrow to run the entire script or

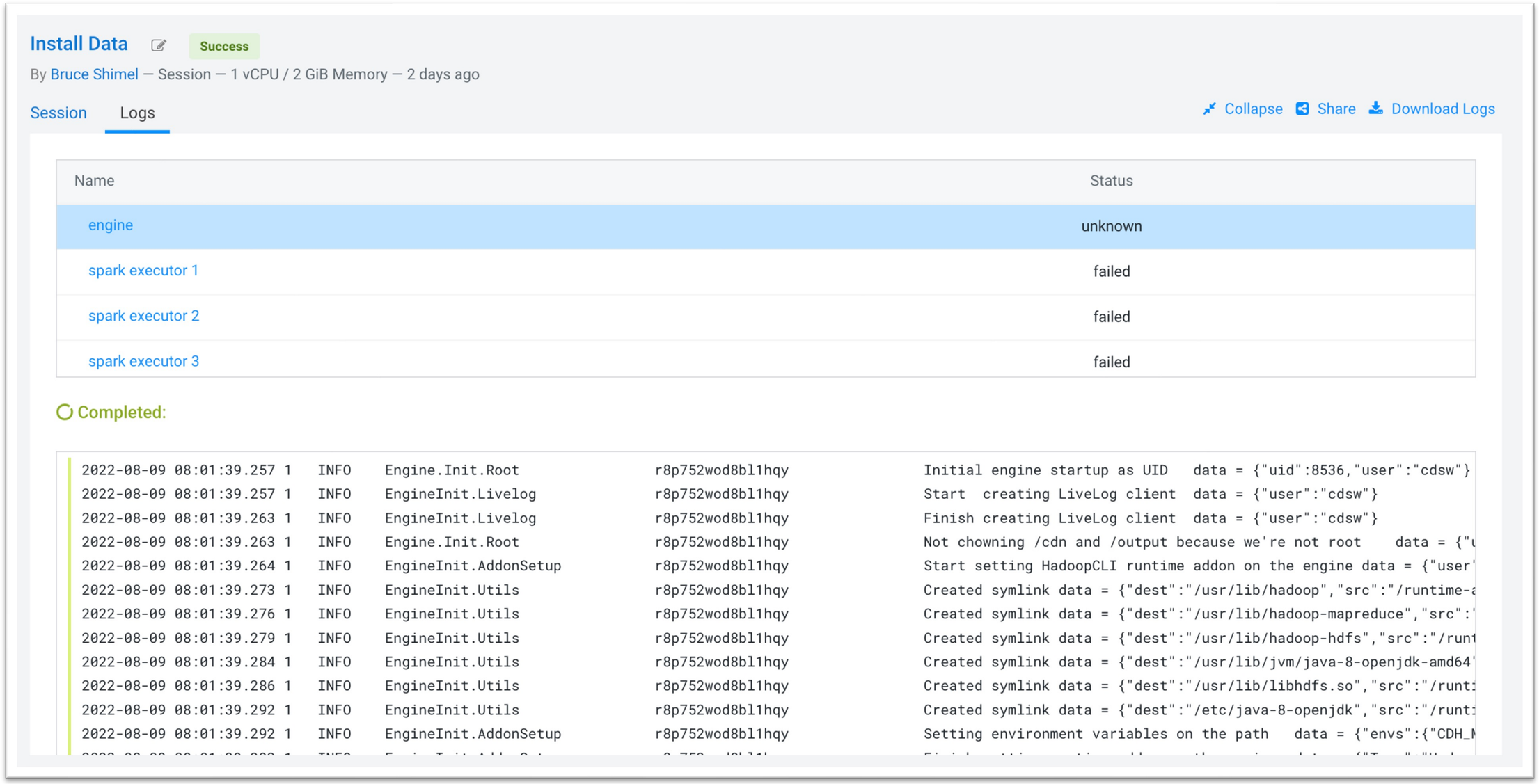
    Highlight the code you want to run and Ctrl+Enter (Windows/Linux) or cmd+Enter (macOS)
- **Code autocomplete**
  - Python and R kernels include support for automatic code completion
  - Single tab to display suggestions
  - Double tab for autocomplete
- **Project code files**
  - All project files are stored to persistent storage at `/var/lib/cdsw/current/projects/<project name>`

Third-Party Editors

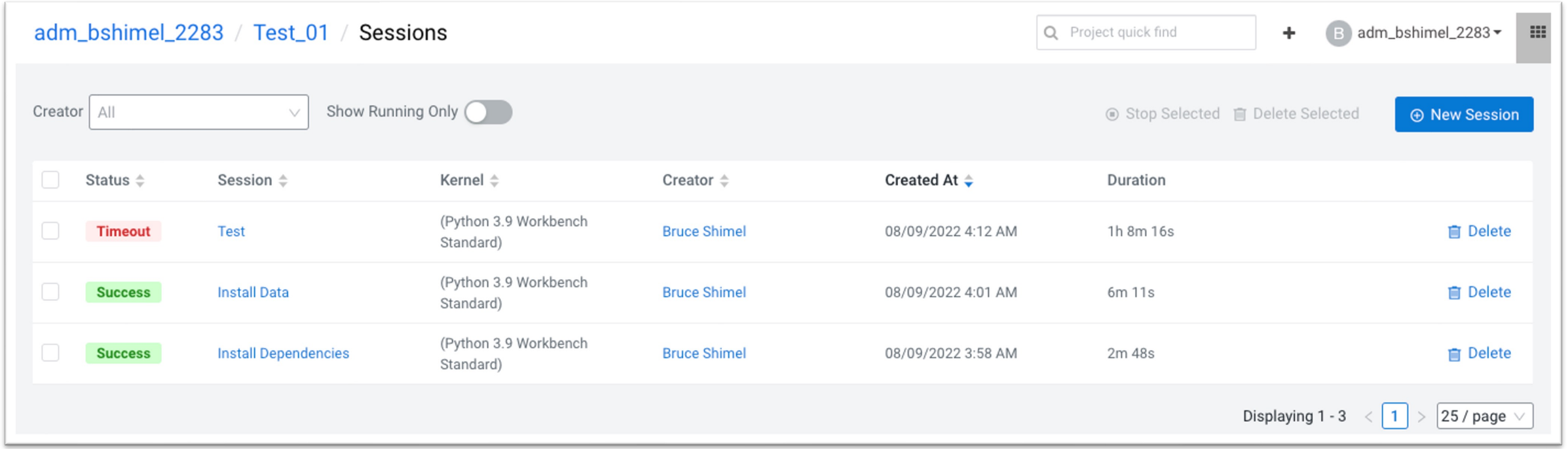
- **CML can be configured to work with third-party editors**
  - Browser-based IDEs such as Jupyter
  - Local IDEs that run on your machine
- **The configuration for an IDE depends on which type of editor you want to use**
- **You can only edit and run code interactively with the IDEs**
- **Tasks such as creating a project or deploying a model require the**
  - CML web UI
  - API v2

Session Output

Session Logs



Session List



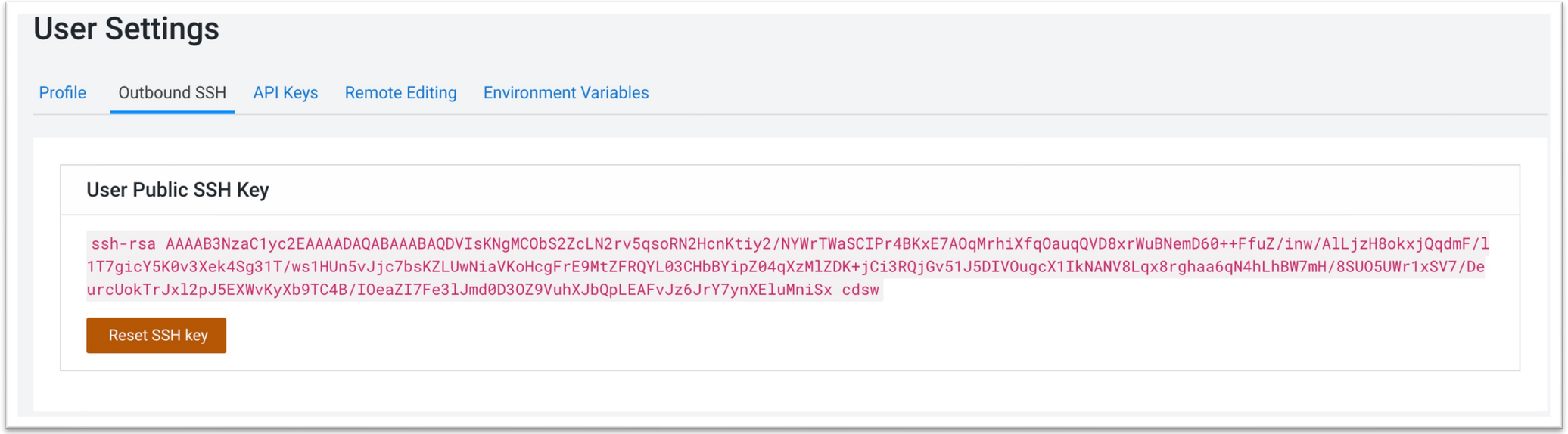
Using Git with CML

- **CML is designed to be used with Git**
  - Create a project by cloning a Git repository
  - Perform Git operations to sync a project with a remote repository
- **CML provides full access to Git on the command line**
  - It does not provide a graphical user interface for Git
- **CML can be used with repositories on any Git server or service, including:**
  - Cloud-based Git hosting services such as GitHub, GitLab, Bitbucket Private Git servers (if accessible)
  - Other version control systems with Git interfaces

Authorizing CML to Access Git

- **In CML, you can create a project by cloning a Git repository**
  - For read-only access to public repos, no extra setup is required
  - For write access or private repo access, authorization is required
    - Add the public SSH key from your CML account to your Git provider account
      - This is a self-service task in the CML user interface

Adding SSH Key to GitHub



1. Sign in to CML
2. Go to User Settings
3. Go to the Outbound SSH tab and copy your public SSH key
4. Sign in to your GitHub account and add the CML key copied in the previous step to your GitHub account
5. For instructions, refer the GitHub documentation on adding SSH keys to GitHub

Cloning a Git Repository in CML

- **In CML, you can create a new project by cloning a Git repository**
- **CML accepts two types of Git URLs**

HTTPS URLs

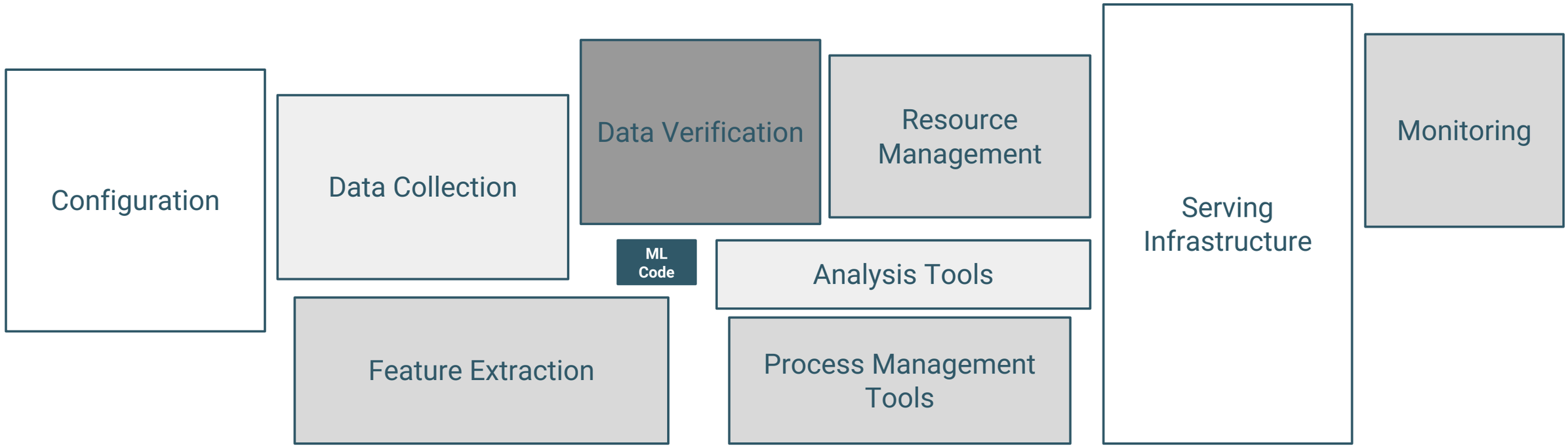
- Begin with https://
- Use for read-only access to public repos
- No extra setup required before using

SSH URLs

- Begin with username@hostname:
- Use for write access or private repo access
- Before using, authorize your CML account to access the Git repo

Hidden Technical Debt of ML Systems

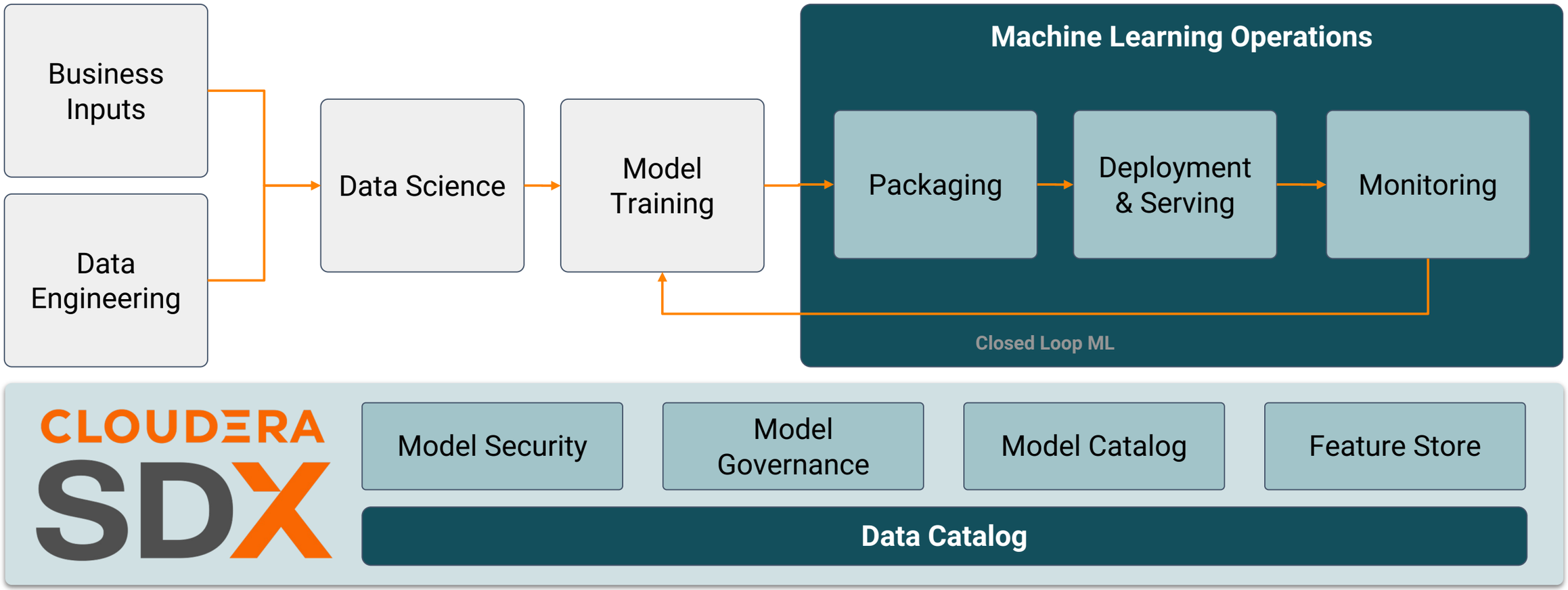
Google Paper



Source: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

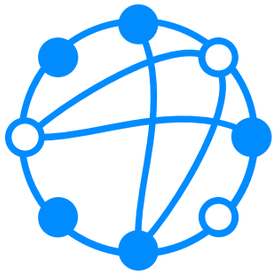
Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle.
The required surrounding infrastructure is vast and complex.

ML Operations Closes the Loop

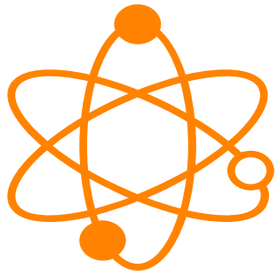


AI Applications

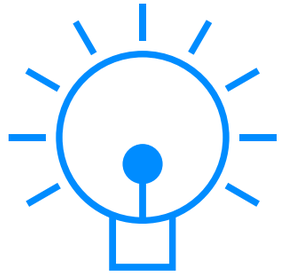
- **Power more data science and AI use cases faster across the enterprise**



Build **predictive data applications**: exploration, model development, training, and deployment, hosted Applications



Tools for the **end-to-end workflow**: SQL, Visuals, Python/R Code, Experiments, Models and Applications



Goals: **Insights** & **Data Products**

At the hand of the **Business**

Applications Defined

- **Applications are long-running web applications that make calls to deployed models that**
  - Give data scientists a way to create ML web applications/dashboards and easily share them with other business stakeholders
  - Can range from single visualizations embedded in reports, to rich dashboard solutions such as Tableau
  - Can be interactive or non-interactive
  - Stand alongside other existing forms of workloads in CML (sessions, jobs, experiments, models)
  - Must be created within the scope of a project like all other workloads
  - Are launched within their own isolated runtime
- **Additionally, like models, engines launched for applications do not time out automatically**
  - Will run as long as the web application needs to be accessible by any users
  - Must be stopped manually when needed

Test Your Application

- **Before you deploy an application, make sure your application has been thoroughly tested**
  - Use sessions to develop, test, and debug your applications
  - Test web apps by embedding them in sessions as described here:

```
https://docs.cloudera.com/machine-learning/cloud/projects/topics/ml-embedded-web-apps.html
```

Create Your Application

Field	Description
Name	*Unique* name for the application
Subdomain	Subdomain that will be used to construct the URL for the web application. For example, if you use test-app as the subdomain, the application will be accessible at test-app.<ml-workspace-domain-name>.
Description	Description for the application
Script	Script that hosts a web application on either CDSW_READONLY_PORT or CDSW_APP_PORT. Applications running on either of these ports are available to any users with at least read access to the project. The Python template project includes an entry.py script to test
Engine Kernel and Resource Profile	Kernel and computing resources needed for this application
Set Environment Variables	Specify the name and value for application variables

Application-level environment variables override the project-level environment variables if there is a conflict

Secure Your Application

- **You can provide access to Applications via either the `CDSW_APP_PORT` or the `CDSW_READONLY_PORT`**
- **Any user with read or higher permissions to the project can access an application served through either port**
  - CML applications are accessible by any user with read-only or higher permissions to the project. The creator of the application is responsible for managing the level of permissions the application users have on the project through the application
  - CML does not actively prevent you from running an application that allows a read-only user (i.e. Viewers) to modify files belonging to the project
  - By default, authentication for applications is enforced on all ports and users cannot create public applications. If desired, the Admin user can allow users to create public applications that can be accessed by unauthenticated users (see next slide)
  - For Transparent Authentication, CML can pass user authentication to an Application, if the Application expects an authorized request. The REMOTE-USER field is used for this task

Public Applications

- **To allow users to create public applications on an ML workspace:**
  1. As an Admin user, turn on the feature flag in Admin > Security by selecting Allow applications to be configured with unauthenticated access
  2. When creating a new application, select Enable Unauthenticated Access
  3. For an existing application, in Settings select Enable Unauthenticated Access
- **To prevent all users from creating public applications**
  - Go to **Admin > Security** and deselect Allow applications to be configured with unauthenticated access
  - After one minute, all existing public applications stop being publicly accessible.

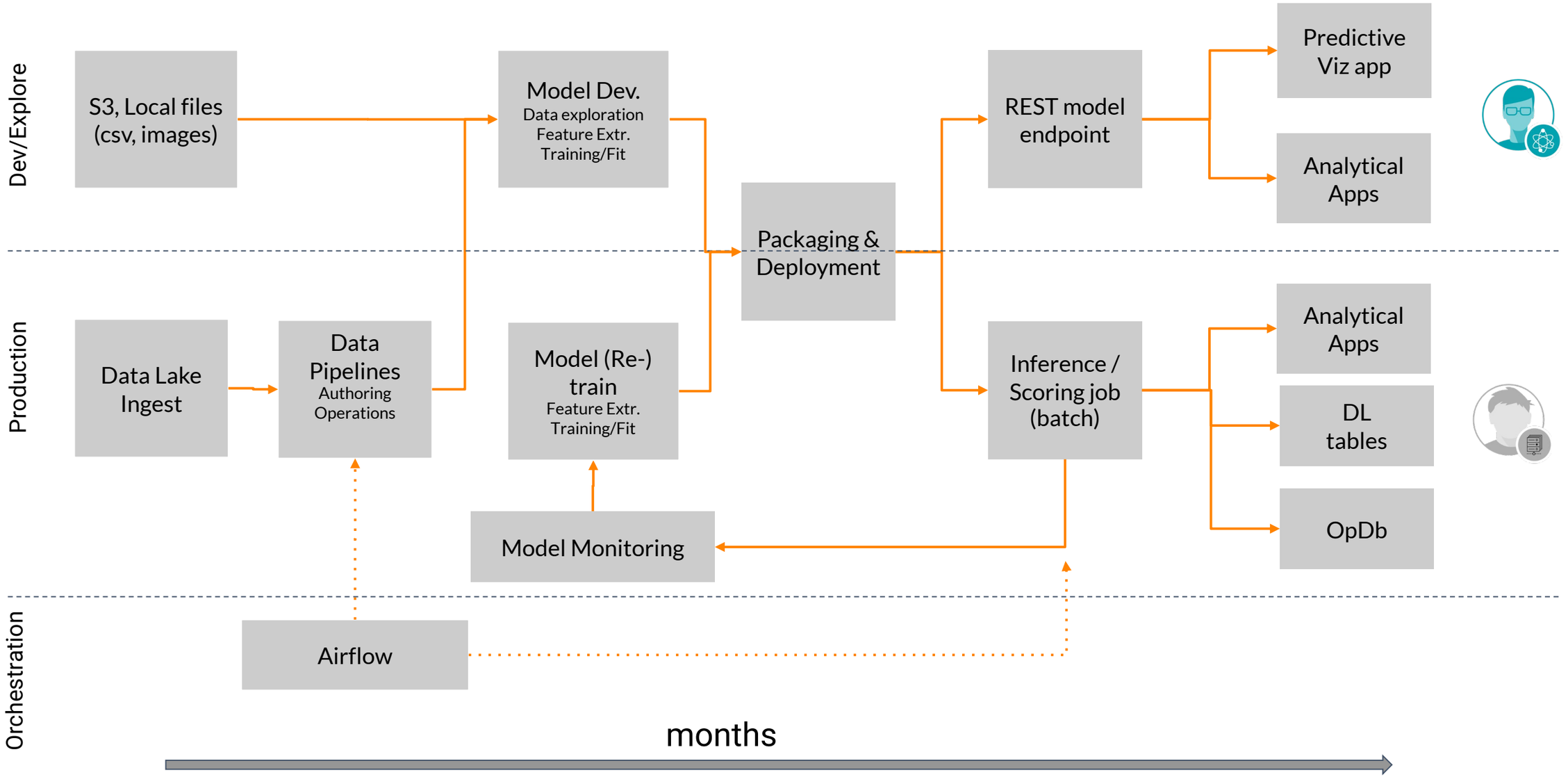
Powering ML Use Cases at Scale is Hard

Enterprises need to overcome the barriers in development and production

20%

Of ML models in the enterprise making it into production Environments —***From 3073 C-Level executives surveyed.***

Building ML Applications - Timeline

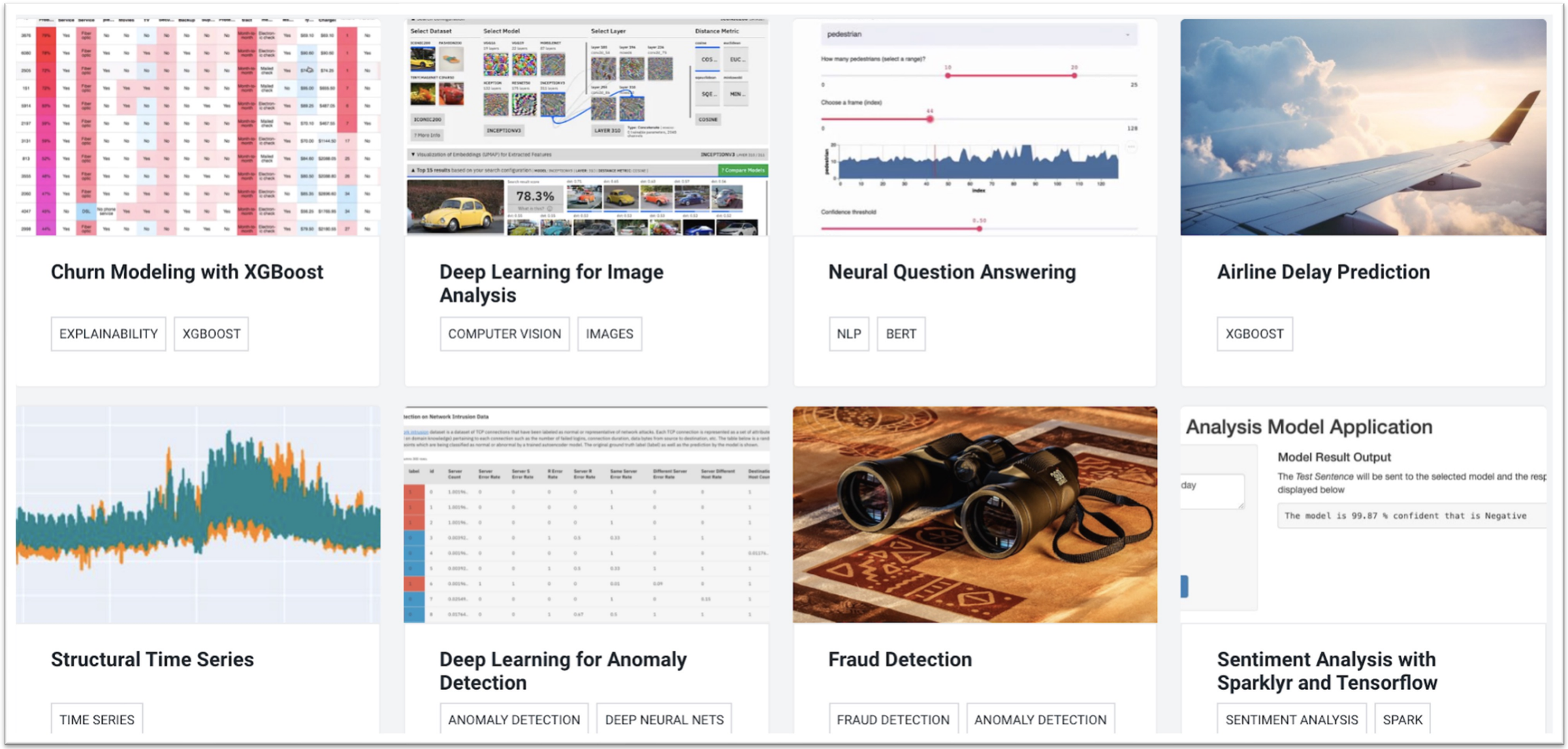


Building ML Use Cases Faster

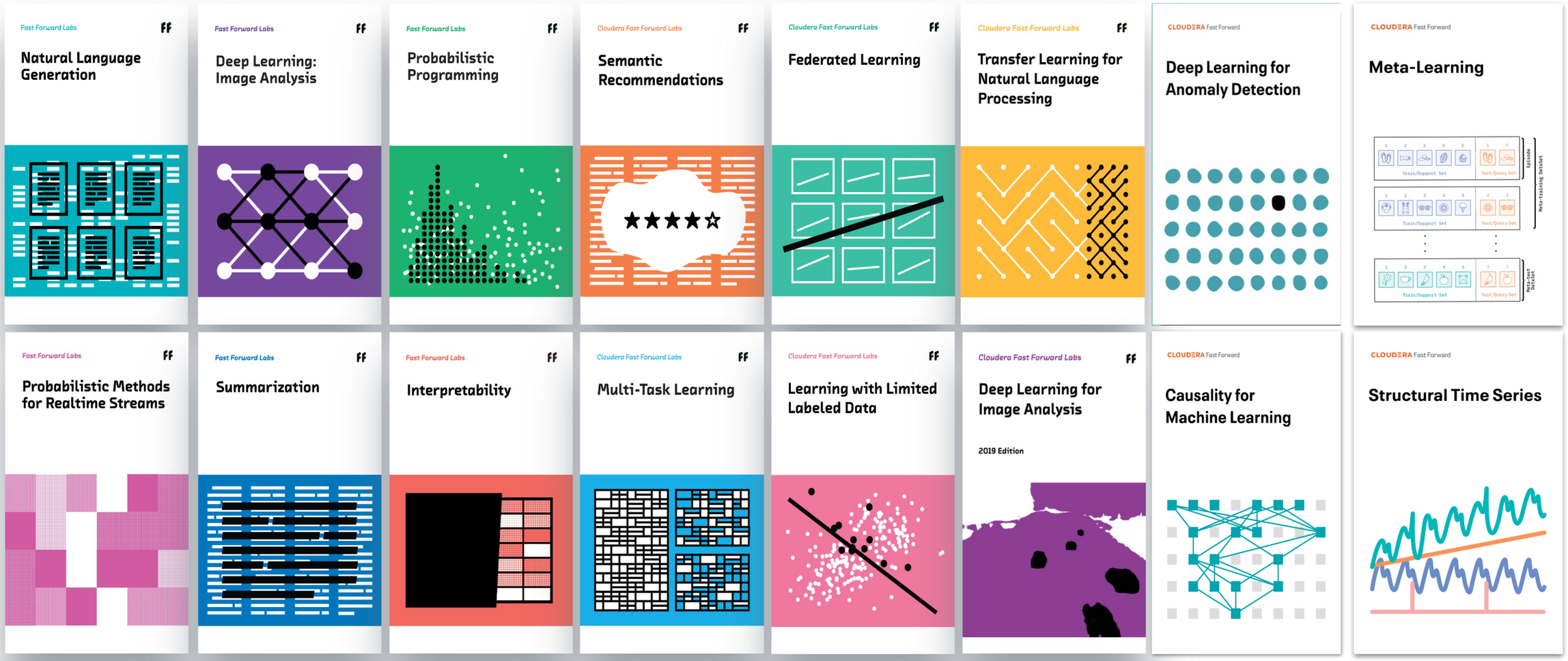
AMPs Defined

- **AMPs stands for Applied Machine learning Prototypes**
- **AMPs accelerate machine learning projects and kickstart AI use cases by providing example workflows and applications that leverage the power of CML**
  - Make it easy for novice CML users to be successful on our platform
  - Decrease time to value by providing high quality reference examples
  - Leverage Fast Forward research to showcase cutting edge ML techniques

Dozens of Use Cases Ready to Plug and Play



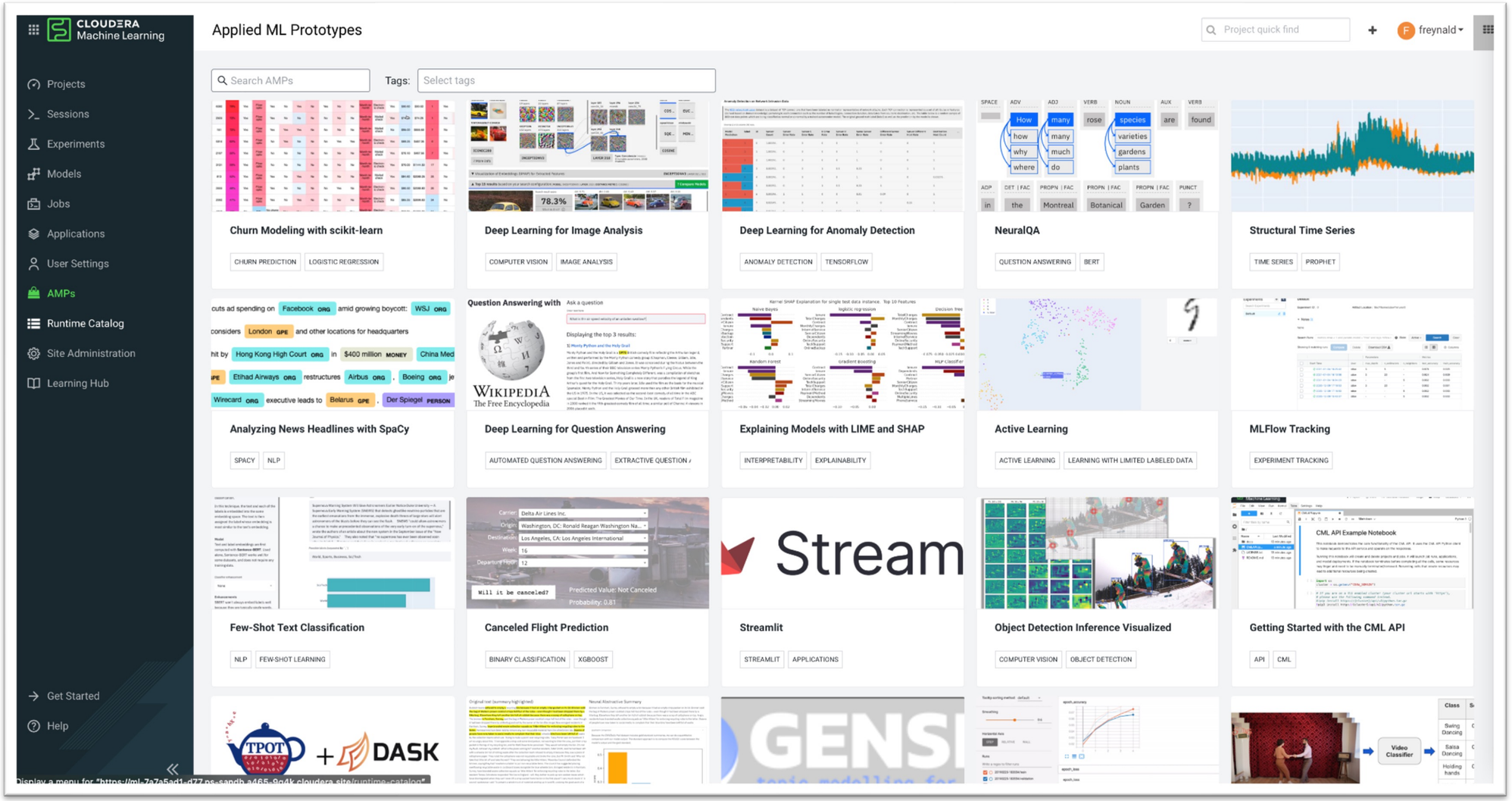
Powered by Cloudera Fast Forward Labs Research



[Preview all of our research here]

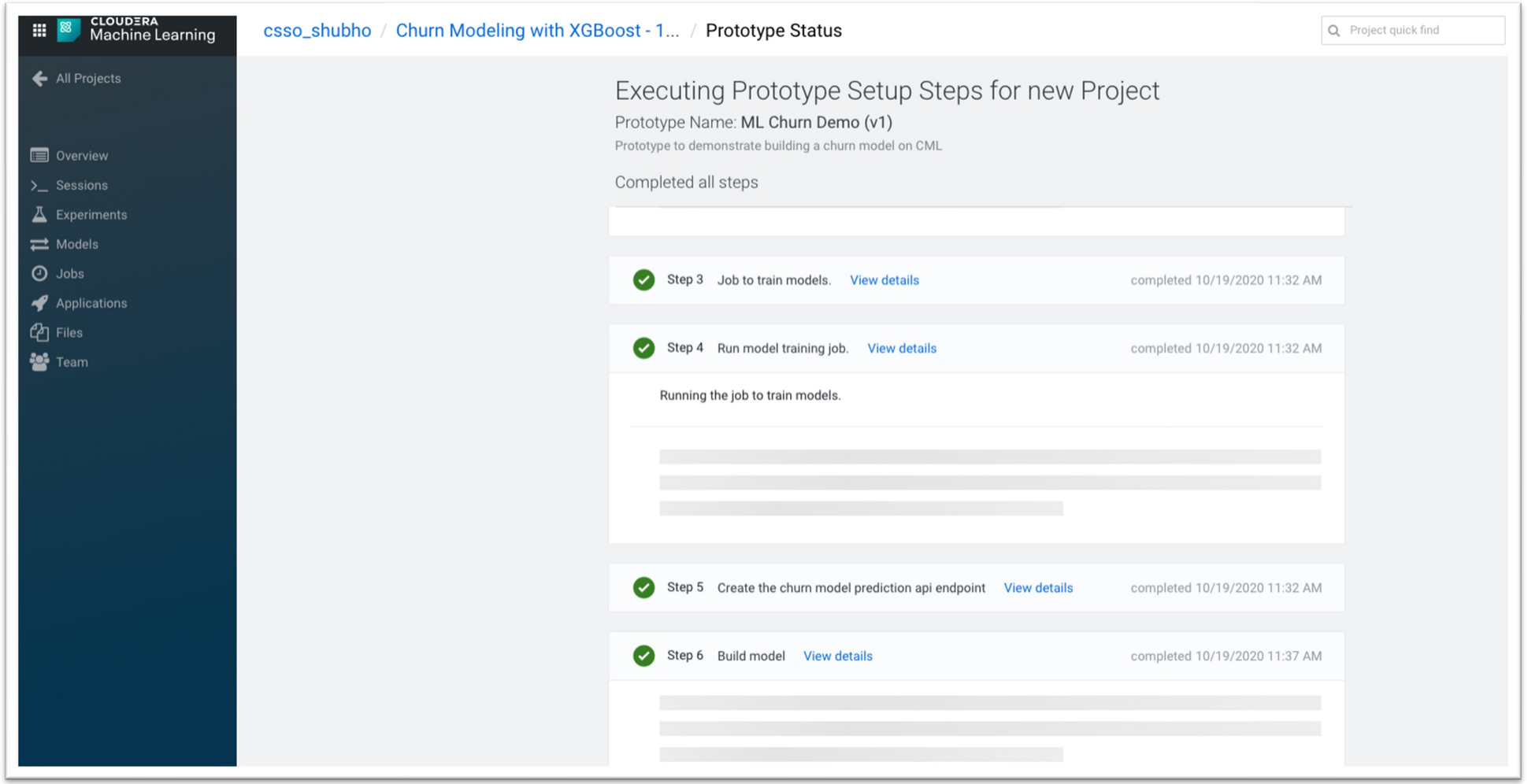
How to Use AMPs?

- **Just pick one…**



Click On It

- **Sit back and enjoy the show!**



Creating New AMPs

- **Build new AMPs**
  - Once a data science project has been built in Cloudera Machine Learning, you can package it and add it to the AMP Catalog
  - ML course exercises is an example
- **Requirements**
  - Project metadata file, which defines the environmental resources needed by the AMP
  - Setup steps to install the AMP in a Cloudera Machine Learning workspace
- **See Creating New AMPs for more information**

Essential Points

- **Run your code from within a session**
  - Can have multiple sessions running in a project
  - Use standard or third-party editor
- **CML provides full access to Git for version control**
- **Create ML web applications/dashboards and easily share them with other business stakeholders**
- **AMPs provide reference example machine learning projects in Cloudera Machine Learning**
  - Are available to install and run from the Cloudera Machine Learning user interface
  - As new AMPs are developed, they will become available
  - Can also create your own AMPs

Hands-On Exercise: Streamlit on CML

- **In this exercise, you will**
  - Use an AMP (Applied ML Prototype) to deploy a simple Streamlit application using CML
  - Deploy an AMP to understand how applications work in CML
- **Please refer to the Hands-On Exercise Manual for instructions**

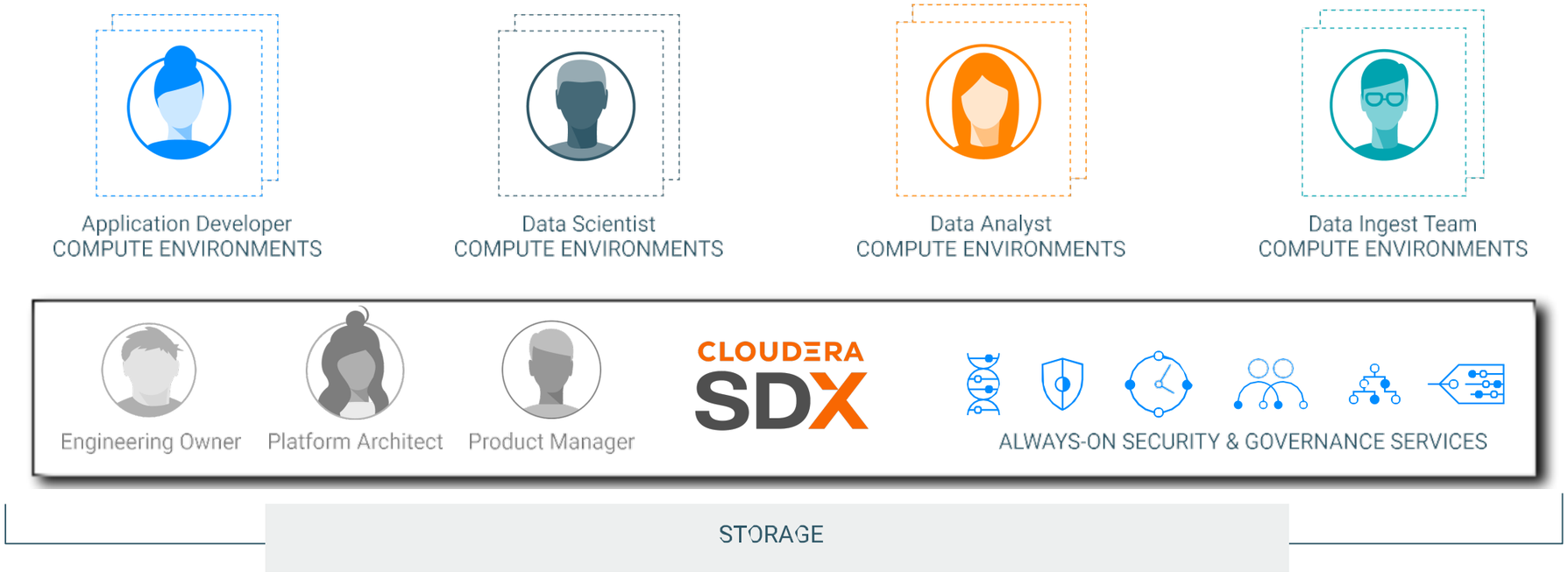
CLOUDERA

Data Access and Lineage

Data Access and Lineage

- **By the end of this chapter, you will be able to**
  - Identify which tools in Cloudera Data Platform (CDP) to use for key data governance activities
  - View and search for entities using the Data Catalog
  - Describe how Apache Ranger applies policies to allow or deny access
  - View and interpret entity's lineage using Apache Atlas

Shared Data Experience (SDX)



- **Shared catalog holds the state (structure and business context) of all data**
- **Unified security model with a consistent set of controls**
- **Consistent governance model for secure access to all data**
- **Flexible data ingest and replication to help preserve data integrity**

SDX Benefits For a Data Scientist

- **A key component within Cloudera SDX is a shared data catalog**
  - Enables self-service access to business data
  - Provides consistent security, governance, and management functions that can be leveraged for analytics applications
- **For a Data Scientist, you want know**
  - What data is available
  - How recent is the data
  - When the data was last modified
  - Is the is data clean
  - If the data was transformed
  - Who touched the data
  - How to get access to the data (for example, who to ask and what to ask)

Data Governance in Cloudera Data Platform

- **Cloudera's Shared Data Experience (SDX) applies security and governance services across all workloads**
  - Apache Ranger
  - Apache Atlas
  - Data Catalog
  - Replication Manager
  - Workload Manager

Apache Ranger

Apache Ranger is an open source application to define, administer, and manage security policies

- **Helps manage policies for files, folders, databases, tables, or columns**
  - You can set policies for individual users or groups
  - *Policies can control full access, or partial access using data masking and row level filtering*
  - Policies are enforced consistently across workloads for a data lake
- **Provides a centralized audit location**
  - Tracks all access requests in real time

Apache Atlas

Apache Atlas is an open source application for managing metadata

- **Exchanges metadata with other tools and processes**
  - Captures lineage across components
  - Import existing metadata and models from current tools
  - Export metadata to downstream systems
- **Allows modeling of assets with complex attributes and relationships**
  - Custom metadata structures in a hierarchy taxonomy
  - Classification of assets for the needs of the enterprise
    - Classifications: PII, PHI, PCI, PRIVATE, PUBLIC, CONFIDENTIAL
- **Enables search for assets using classification and attributes**
- **Includes REST API for flexible access**

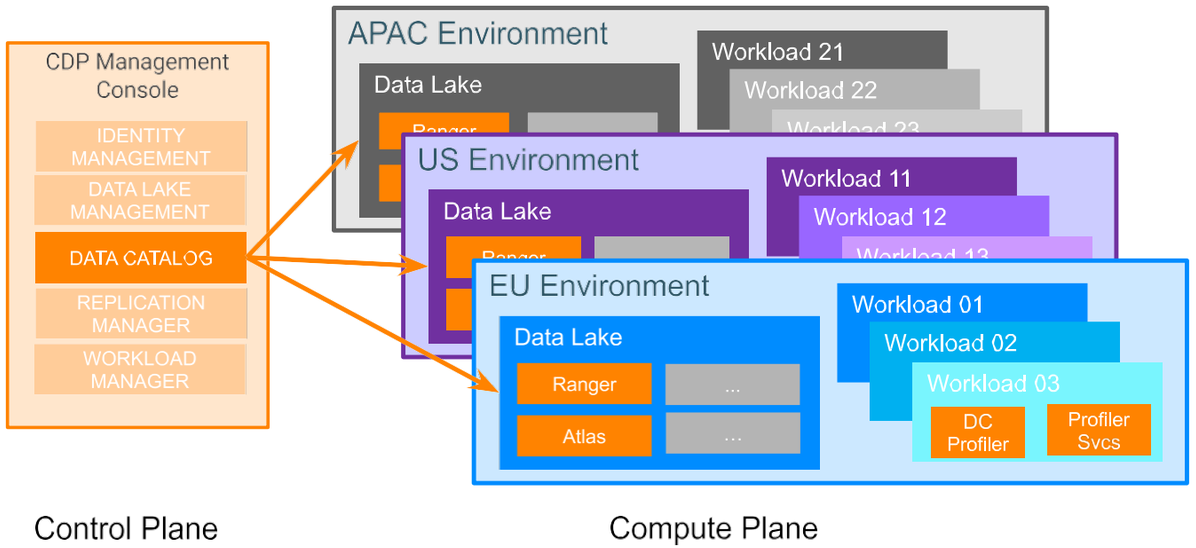
Data Catalog

Data Catalog is a service in CDP for managing, securing, and governing data assets

- **A layer over Ranger and Atlas, providing easier access to their information**
- **Features**
  - Search function for quick access to data objects
  - Quick links to Apache Atlas and Apache Ranger
  - Profilers enable automatic gathering and quick viewing of information
    - Cluster Sensitivity Profiler: Finds and tags potentially sensitive data
    - Ranger Audit Profiler: Summarizes audit logs provided by Apache Ranger
    - Hive Column Statistical Profiler: Provides summary statistics for columns in Hive tables
  - Collections of data assets called datasets facilitate collaboration
    - Bookmarking
    - Discussions and commenting

Data Catalog Overview

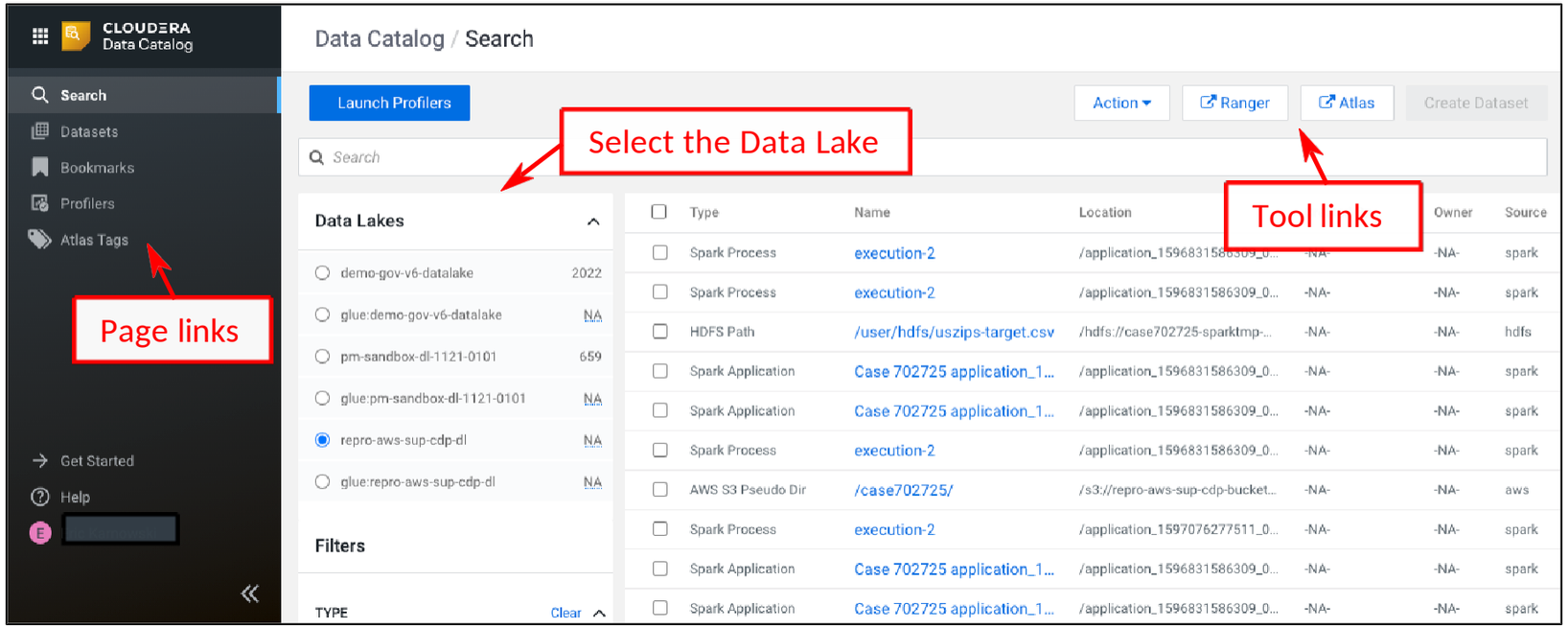
- **Data Catalog provides access to items in a particular Environment**
- **Within each Environment are**
  - A Data Lake (storage)
    - Atlas and Ranger monitor and control access to the Data Lake
  - Workloads (compute)
    - Profilers and profiler services use dedicated workload resources



Accessing Data Catalog

- **Click the Data Catalog icon in the Control Plane on the CDP home page***

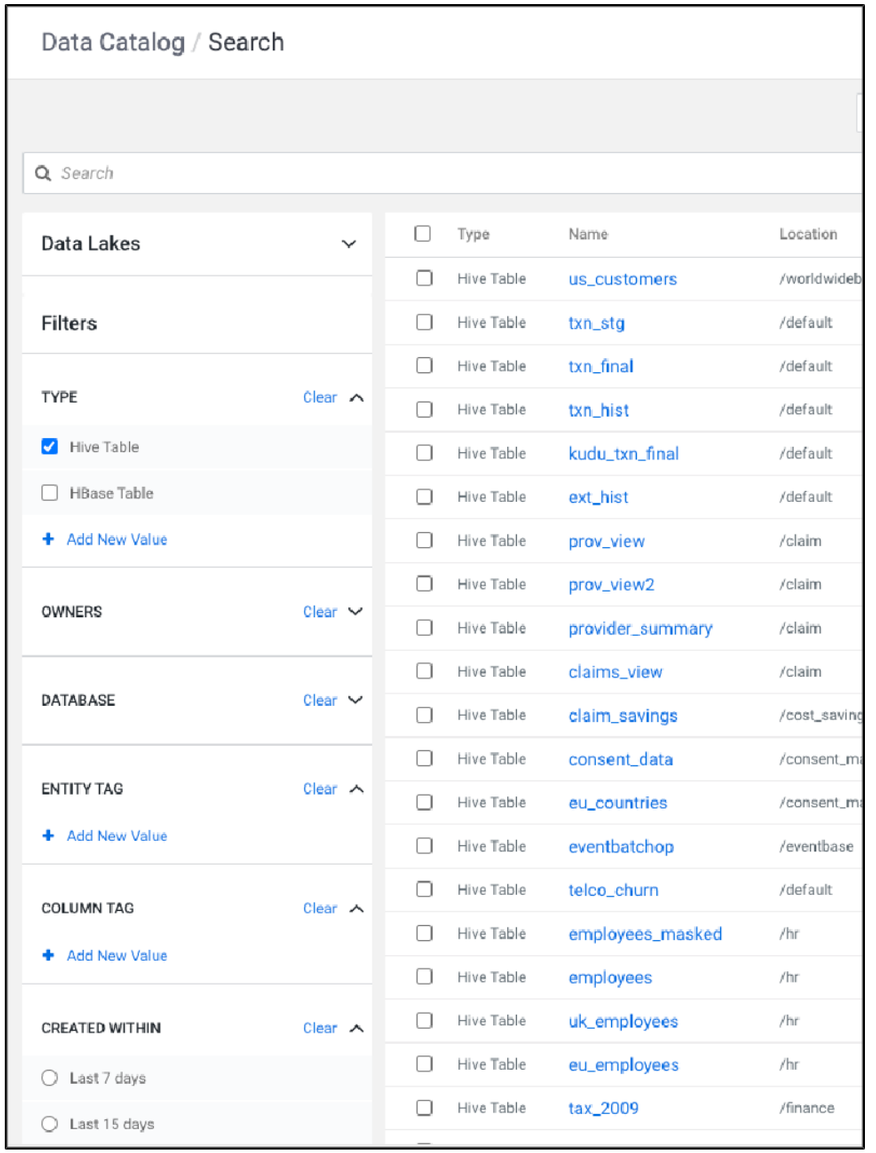
Navigating Data Catalog



- **Data Catalog presents information for a specific Data Lake**
- **Page links provide different features within Data Catalog**
- **Tool links take you to Atlas and Ranger UI pages for the same Data Lake**

Searching in Data Catalog

- **Opens to the Search page**
  - Fast access to Entities
  - Sufficient for most search needs
- **Uses faceted filtering**
- **Includes filters for specific attributes, for example**
  - Owner
  - Database
  - When created



Asset Details: Schema

- **Hive column statistical profiler provides statistics for each column of a table**

  Unique values*
  Null values
  Minimum value
  Mean of values
  Maximum value

- **Use the Schema tab for the Hive table page**

Chart Type	Column Name	Type	Unique Values *	Null Values	Max	Min	Mean
	age	int	33	NA	84	19	46.34
	dateofbirth	date	0	NA			
	email	string	52	NA			
	id	int	53	NA	979,607	134,841	476,579.68
	name	string	49	NA			
	phone	string	52	NA			
	region	string	4	NA			
	salary	int	49	NA	197,537	42,005	114,979.2

* Approximate values as being computed using HLL algorithm.

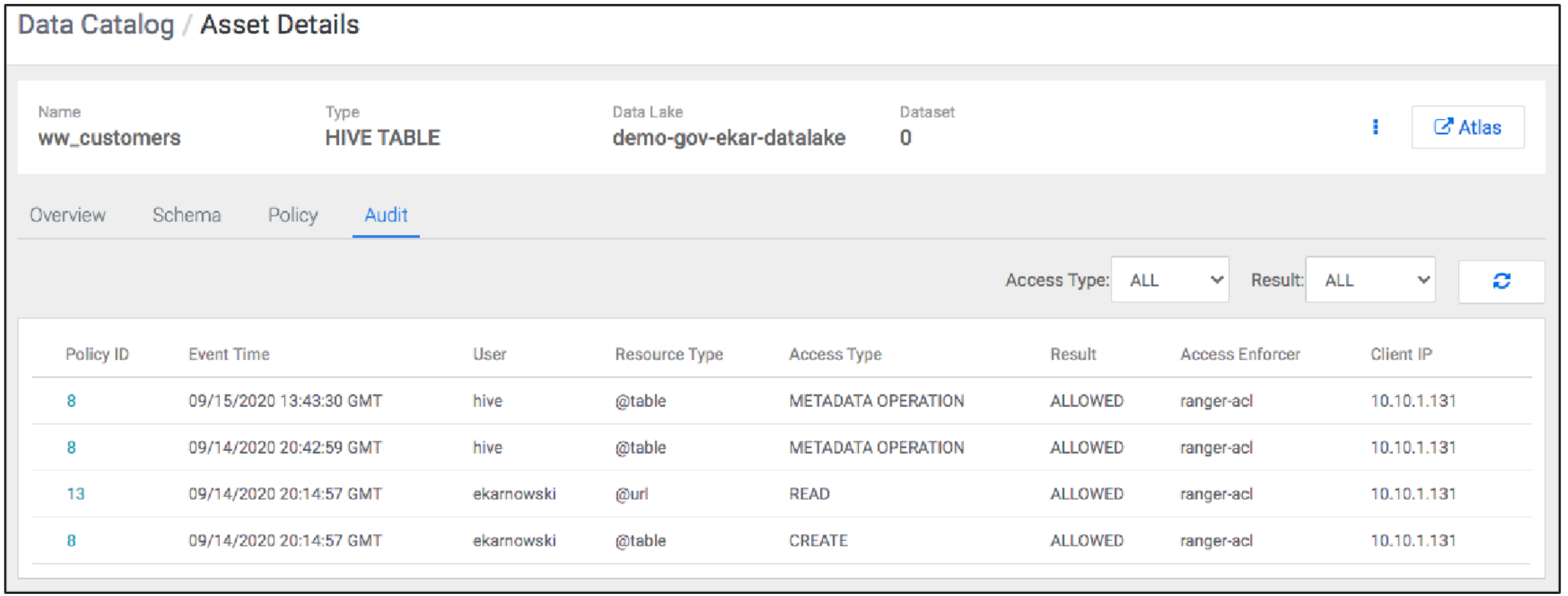
* Approximation, not actual count

What About the Rest of Your Data?

- **Data includes how the entities have changed**
  - **Atlas**: Data lineage
- **In addition, data also includes**
  - **Ranger**: Metadata
    - Who created the table
    - How it was created
    - …
  - **Atlas**: Audit Logs
    - Details of access to data and metadata

Asset Details: Audit

- **Data Catalog provides quick assess to audit information**
  - Operational metadata information from Atlas
  - Access information from Ranger

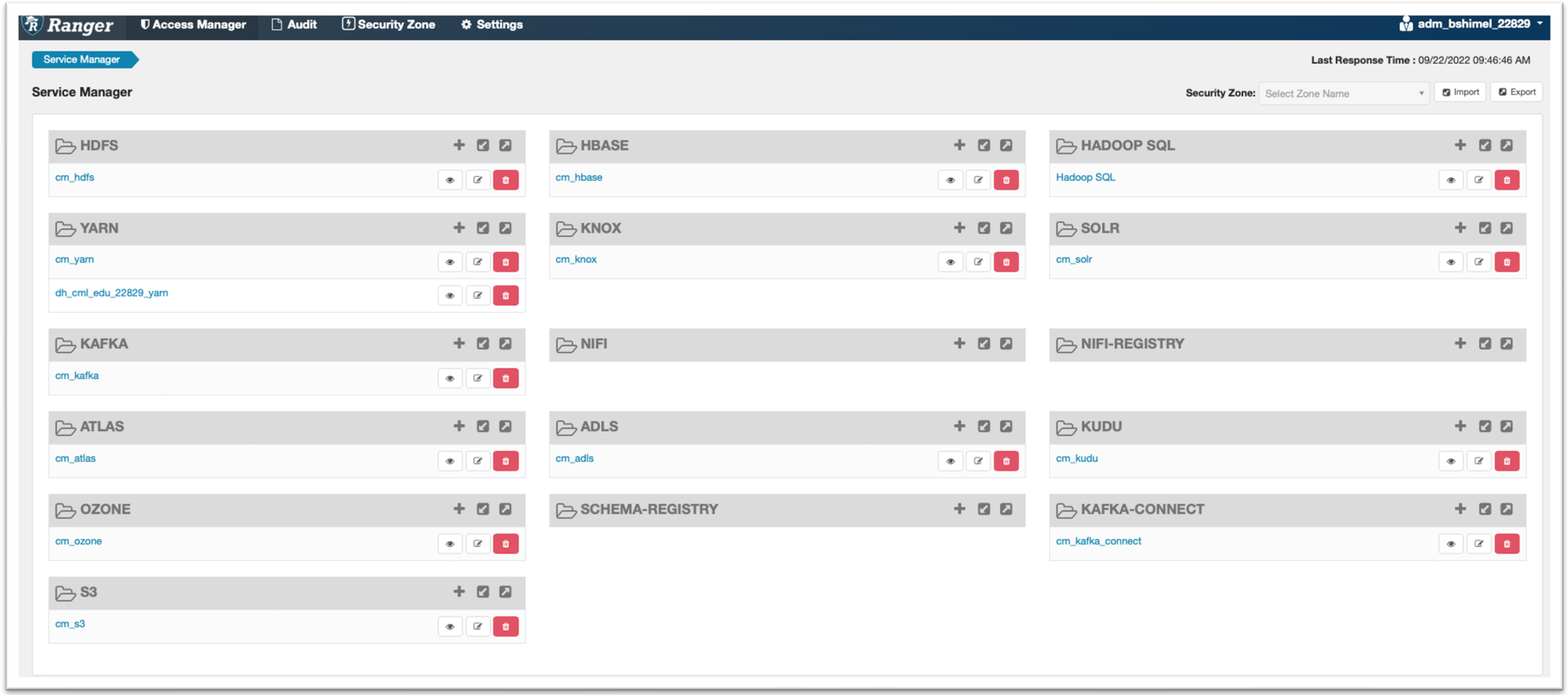


Authorization

- **Controlling access to a resource**
  - Limit the scope of tools or resources available to a user
  - Limit the scope to data available to a user
- **Depends on authentication**
- **Multiple strategies including**
  - Role-based access control (RBAC)
  - Attribute-based access control (ABAC)

Data Access Using Apache Ranger

- **The comprehensive policy management system across CDP and CDF**
- **Managed using**
  - Browser-based UI
  - REST API

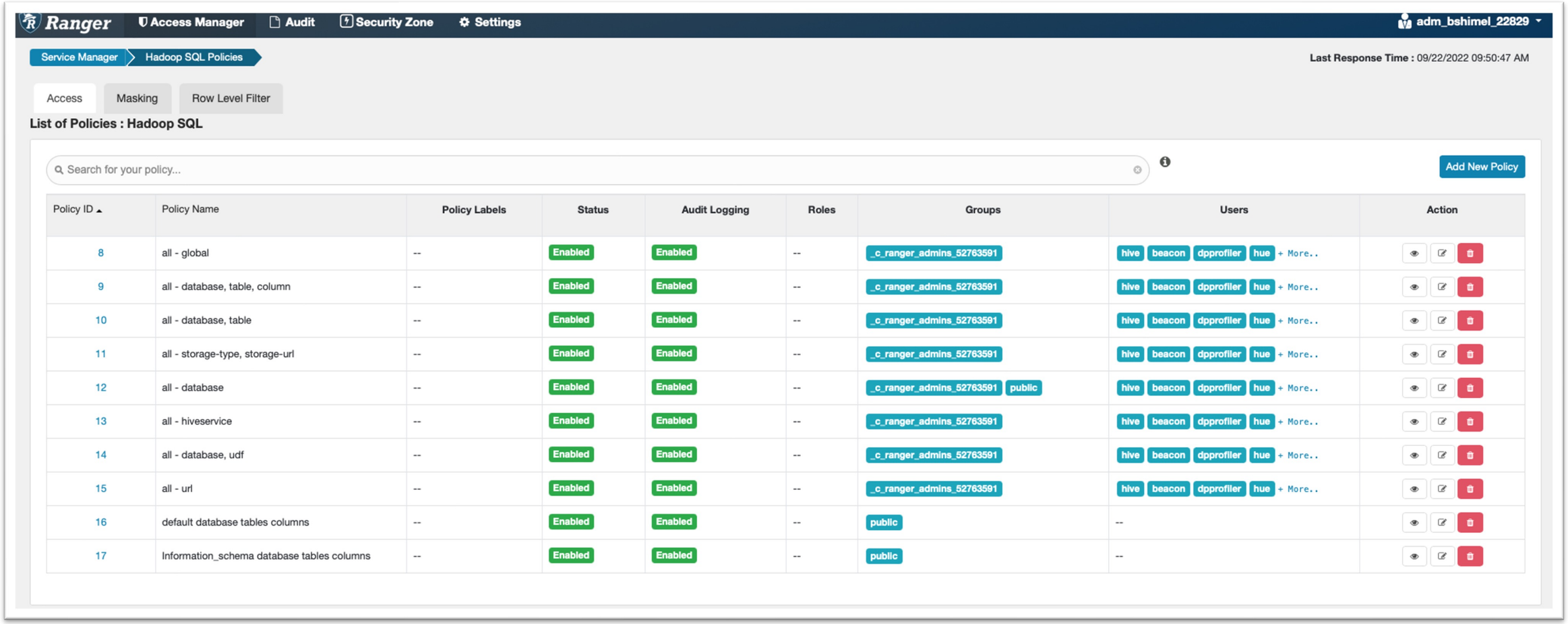


Policies in Apache Ranger

- **Policies provide the rules to allow or deny access to an entity based on a**
  - Role
  - Group
  - User
- **Resource-based policies are associated with a particular service**
  - Identify who can use the resource
    - Perform specific actions
    - Access specific assets
  - Create or edit through the plugin for the service
- **Attribute or tag based policies**
  - Restrict access using classifications and other attributes

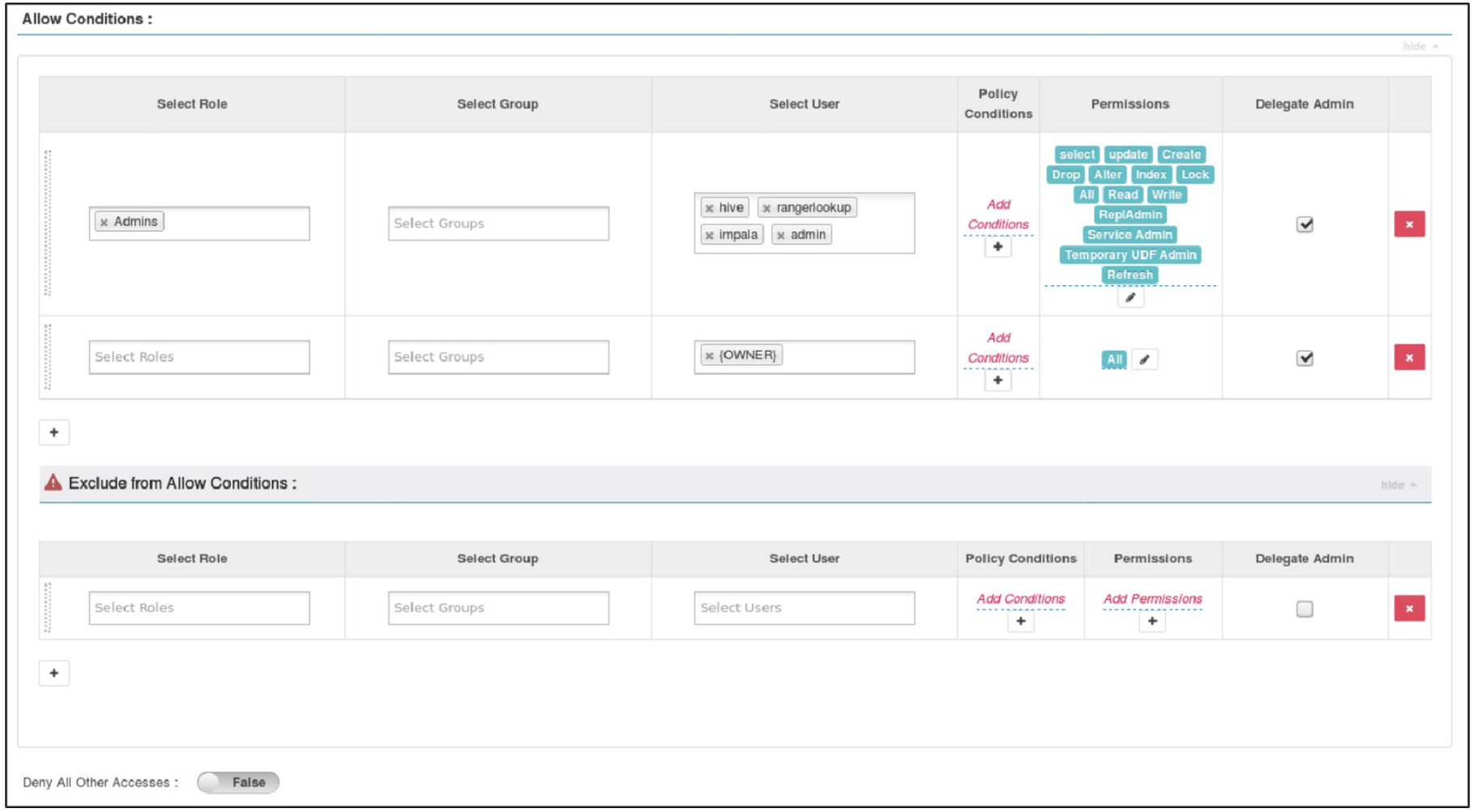
Resource-Based Policies

- **Policies specific to the tool being used (Hive, HDFS, Kafka, and so on)**



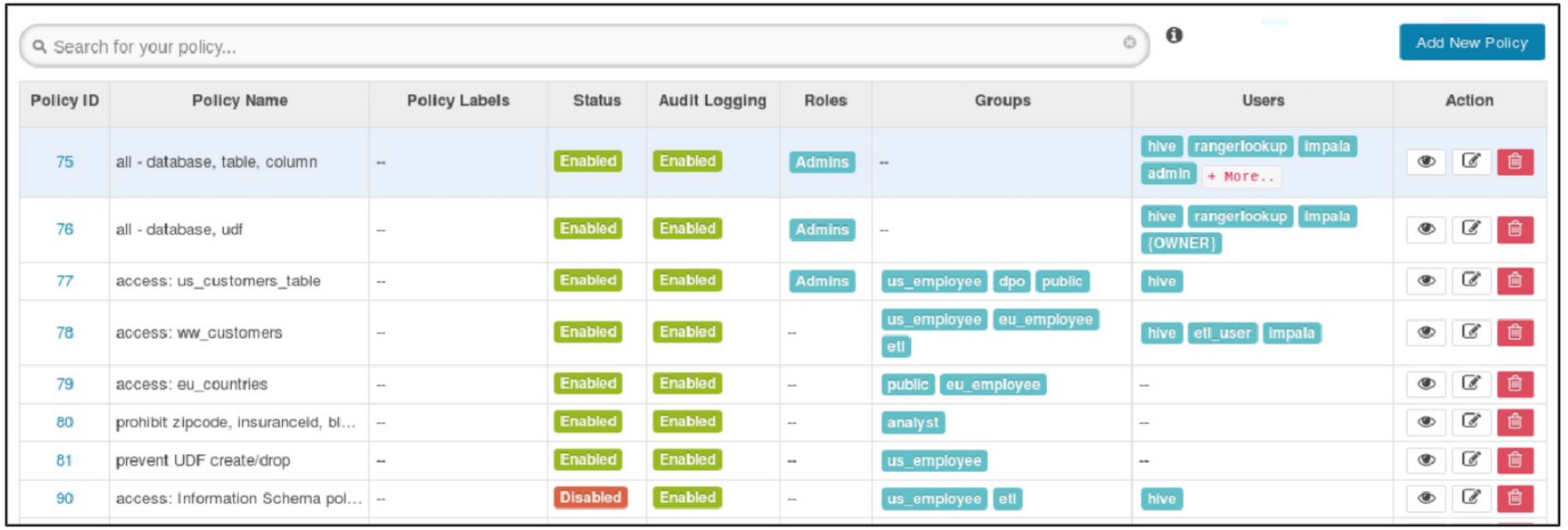
Policy Conditions

- **Policies are defined using allow and deny conditions**



View Policies

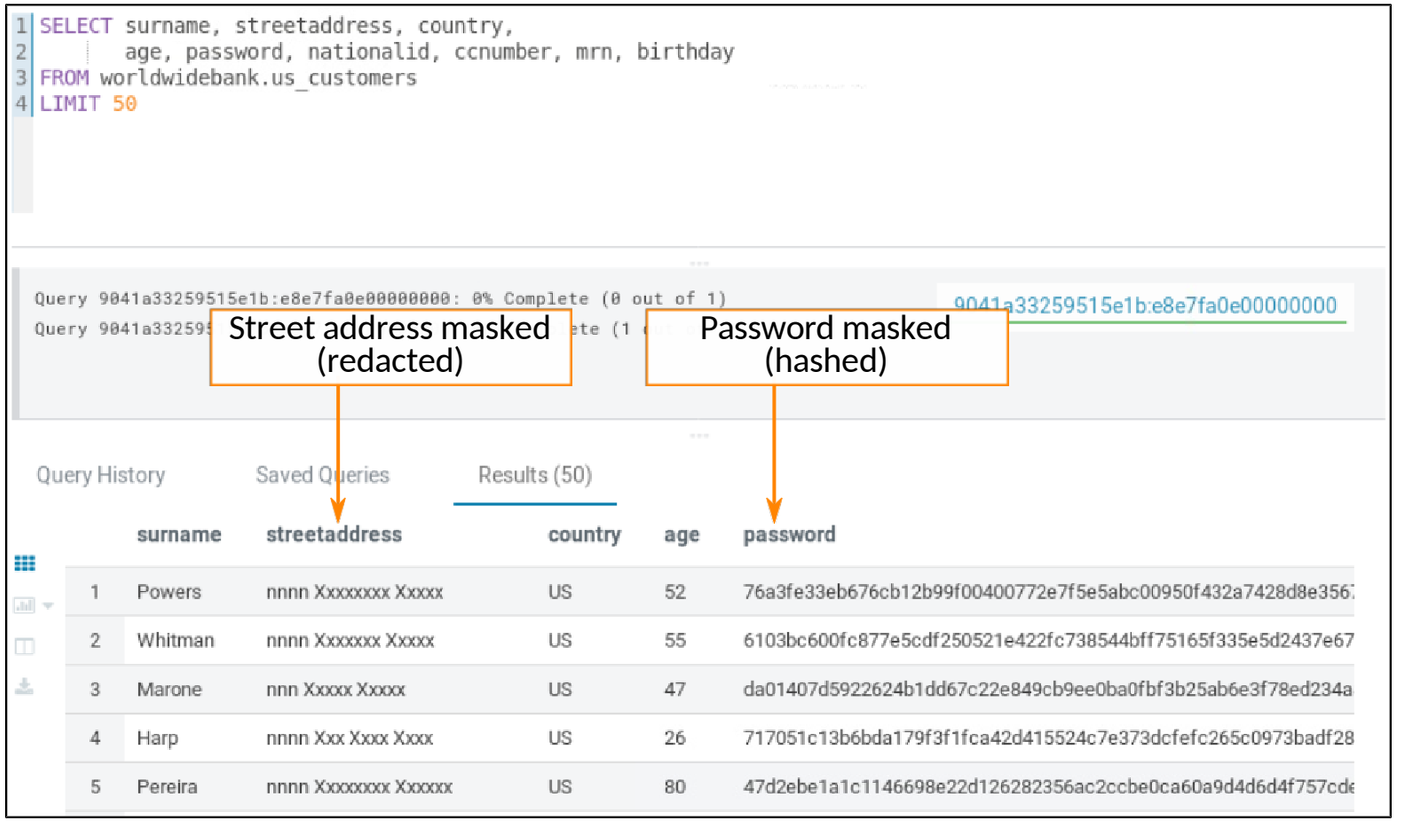
- **List of policies provides overview**



- Policy ID links to policy details
- Labels for organizing and easy search
- Status (enabled or disabled)
- Roles, groups, and users mentioned in the policy

Policies that Mask Data

- ***Masking data* provides results without values**
  - Possibly showing the data exists without exposing it
  - Supported for Hive tables (accessed with Hive or Impala)

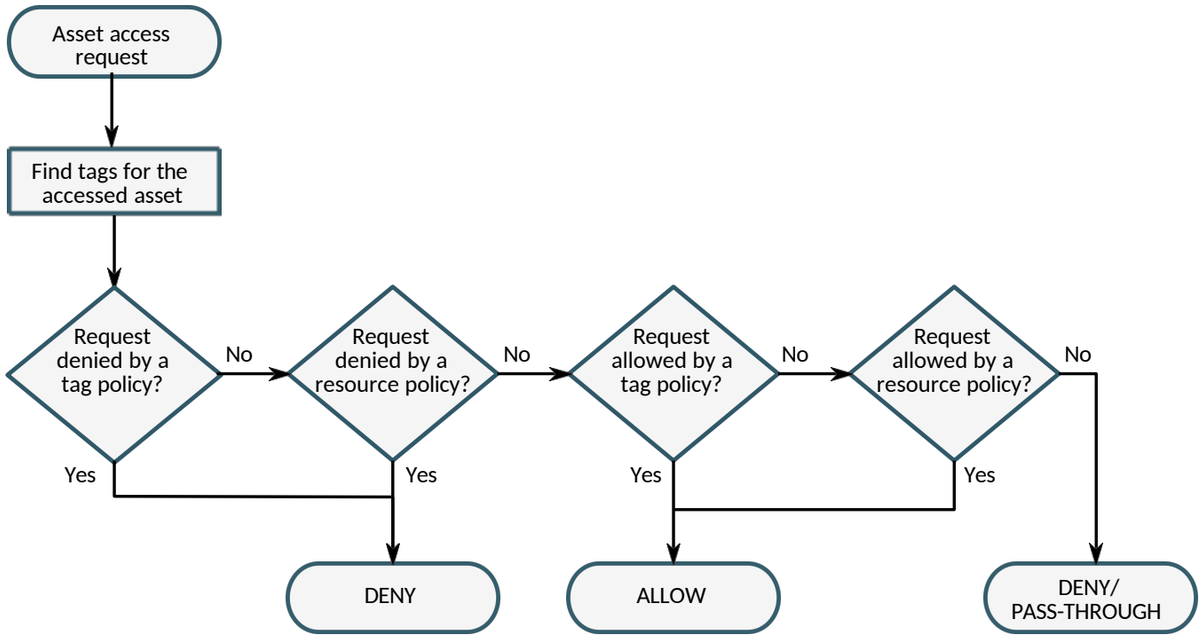


Masking Options

Masking Option	Description	Example
Redact	Replace alphabetic characters with “x” and numeric characters with “n”	`nnn Xxx Xxxxx`
Show last 4	Show only the last four characters	`xxx-xx-8366`
Show first 4	Show only the first four characters	`5.20xxxx+xx`
Hash	Replace all characters with a hash of entire cell value	`6103bc600fc877`
Nullify	Replace all characters with `NULL`	`NULL`
Show only year	Show only the year portion of a date string and default the month and day to 01/01	`01/01/2022`
Custom	Use a valid Hive expression to specify custom value (must return the same data type)	*varies*

Policy Evaluation Flow

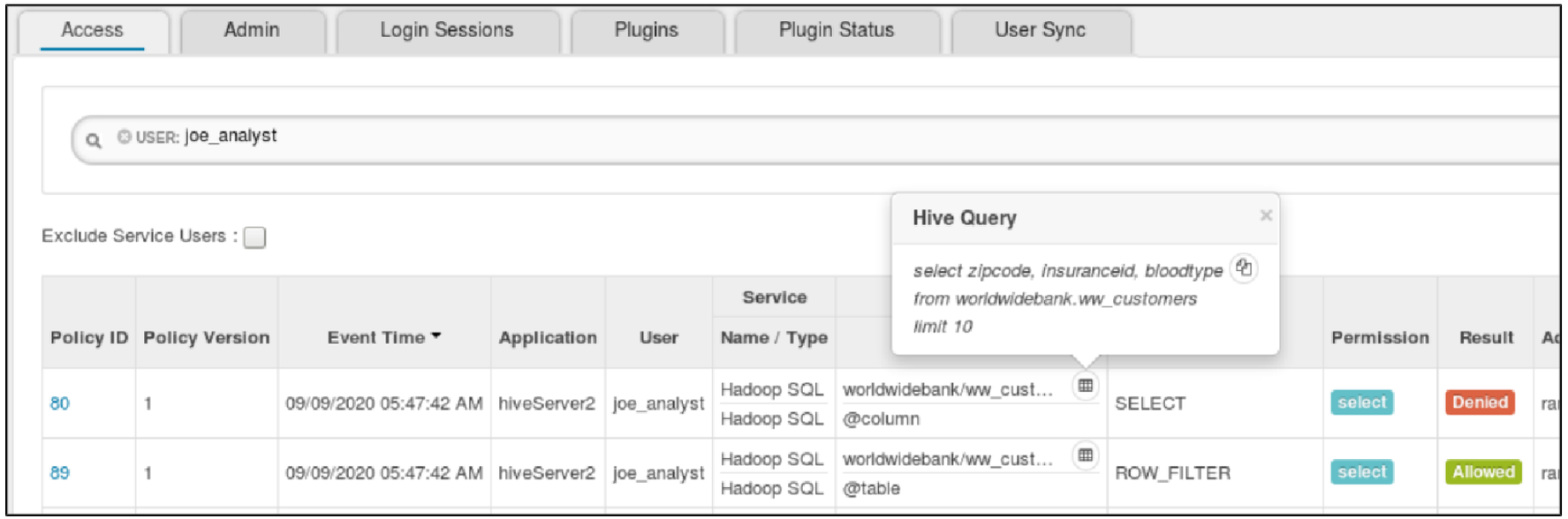
- **Deny conditions are checked first, then allow conditions**
  - This is opposite of the order they are presented on the policy page in Ranger



Audits in Apache Ranger

- **Use the Audit link in the Ranger menu bar**
- **From this page you can get information about several aspects of the system, for example:**
  - Access: Which data was accessed, when, and by whom
  - Admin: Any administrative tasks, such as creation of policies and assignment of user roles
  - Login Sessions: Login attempts to Ranger Administration; when, by whom, and whether successful
  - User Sync: Service activity data for all usersync processes

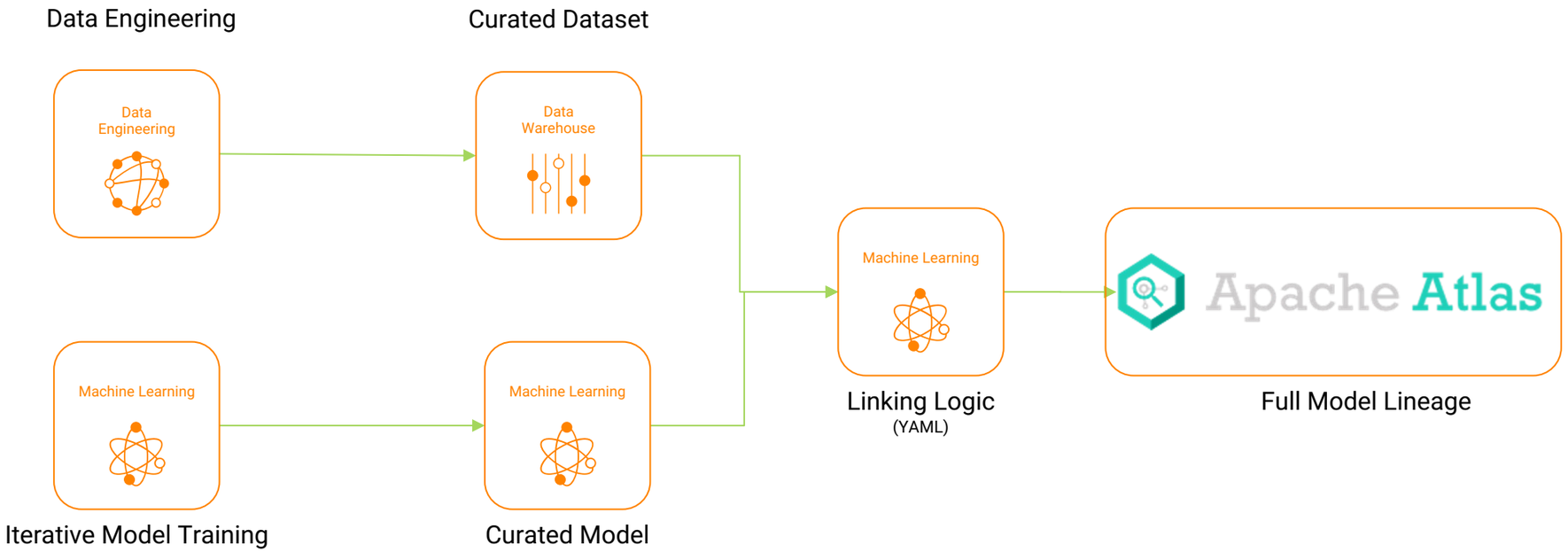
Access Audits in Ranger



- **Use to get details of access including**
  - Applicable policy ID, with link to policy details
  - Time of access event
  - User
  - Resource accessed, with a pop-up showing the query
  - Result (access allowed or denied)

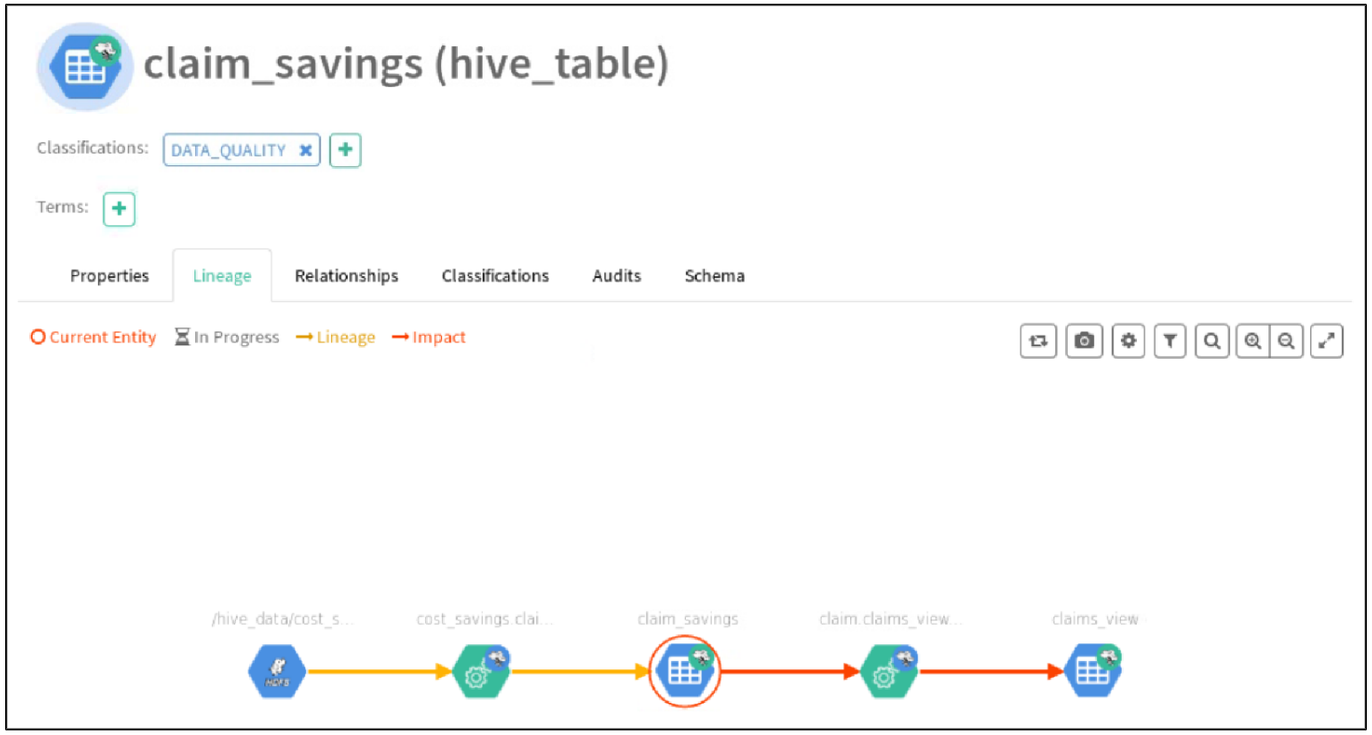
Inspecting Data using Apache Atlas

- **View and interpret the lineage for a table or other data object**
  - Find when errors were introduced
  - Consider impact a change in a table might have
  - Assess the suitability of a table's data for a given purpose
  - Reduce duplication of assets
  - Informs future project decisions



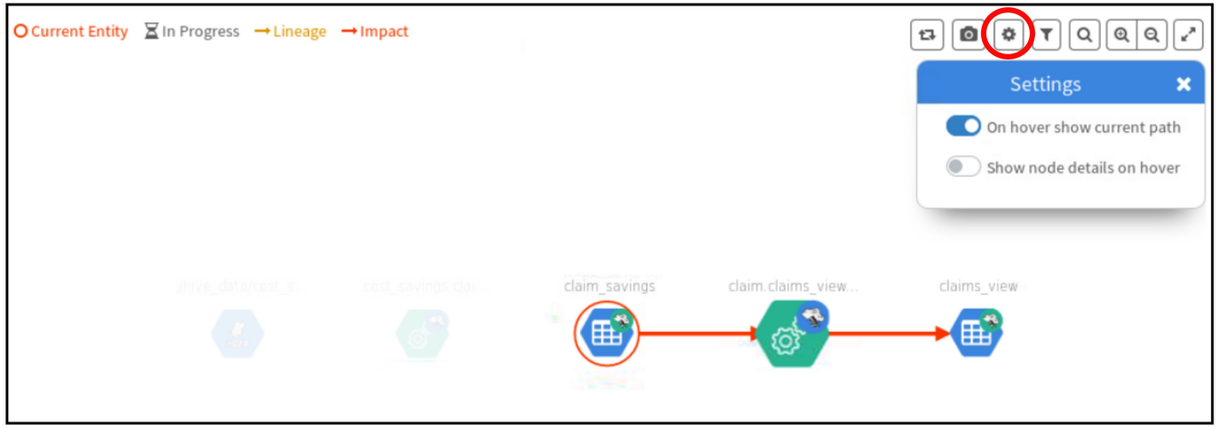
Viewing an Object's Lineage in Apache Atlas

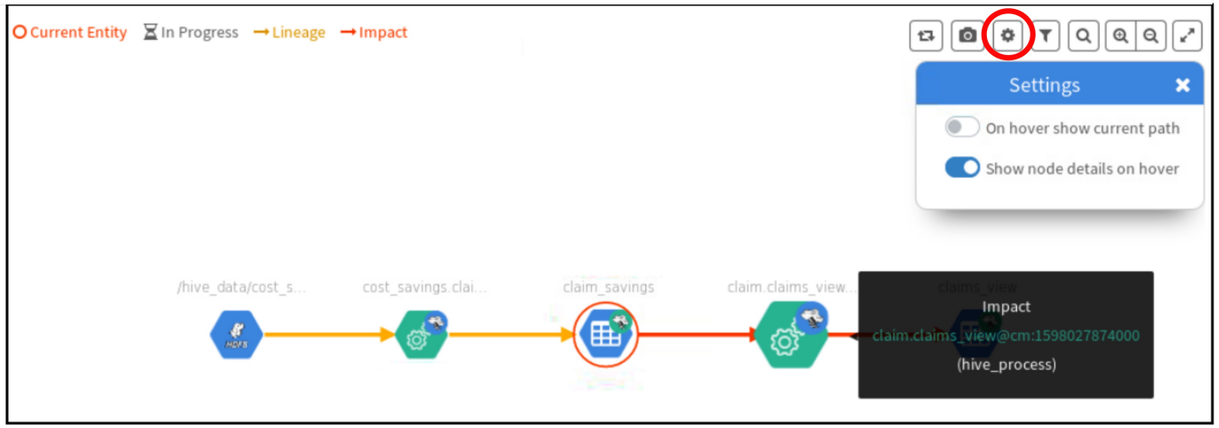
- **Use the Lineage tab on an entity's details page**
  - Appears as a directed graph
  - Differentiates between lineage (upstream) and impact (downstream)
- **Provides information such as who created the entity, and when**
  - Does not provide information about access queries



Lineage Nodes

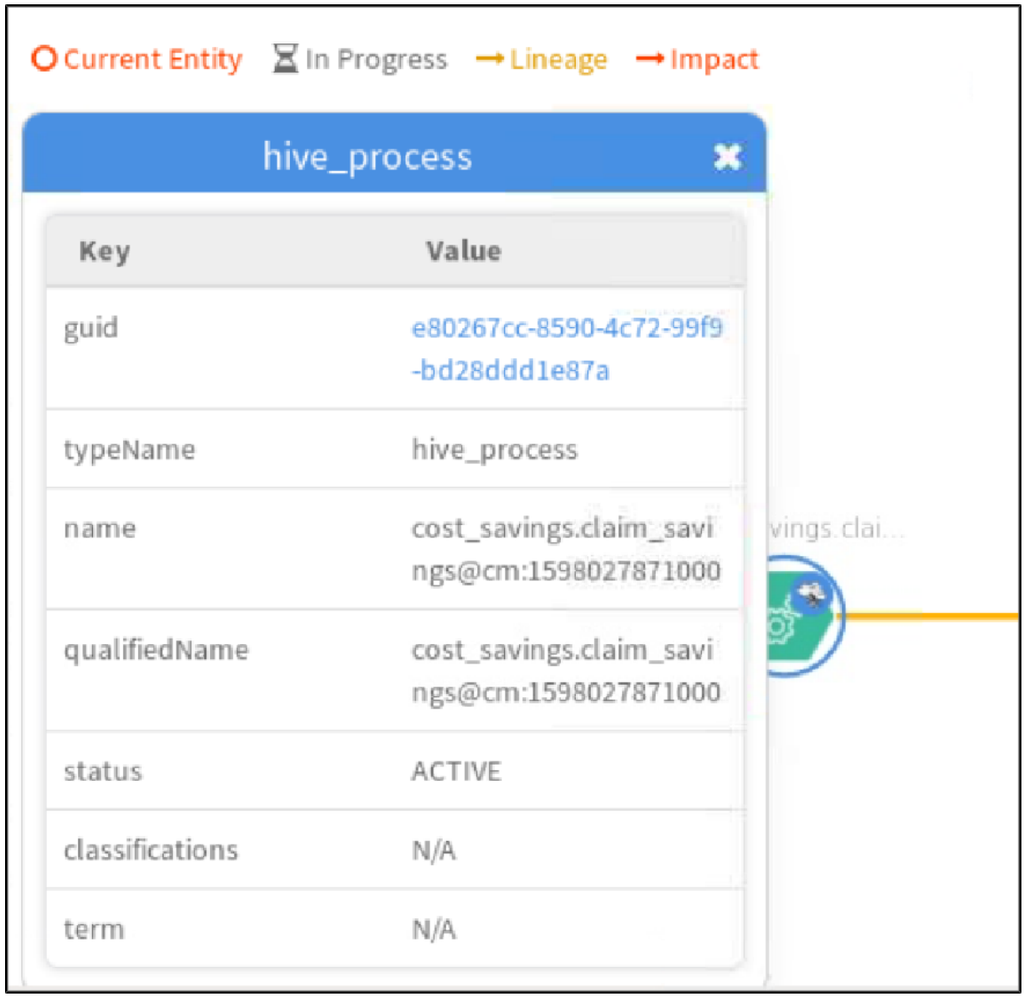
- **Nodes represent data objects**
  - Tables, views, processes, …
- **Settings control what you see when hovering over a node**
  - Current path (immediate previous and next nodes)
  - Details (entity name and type)





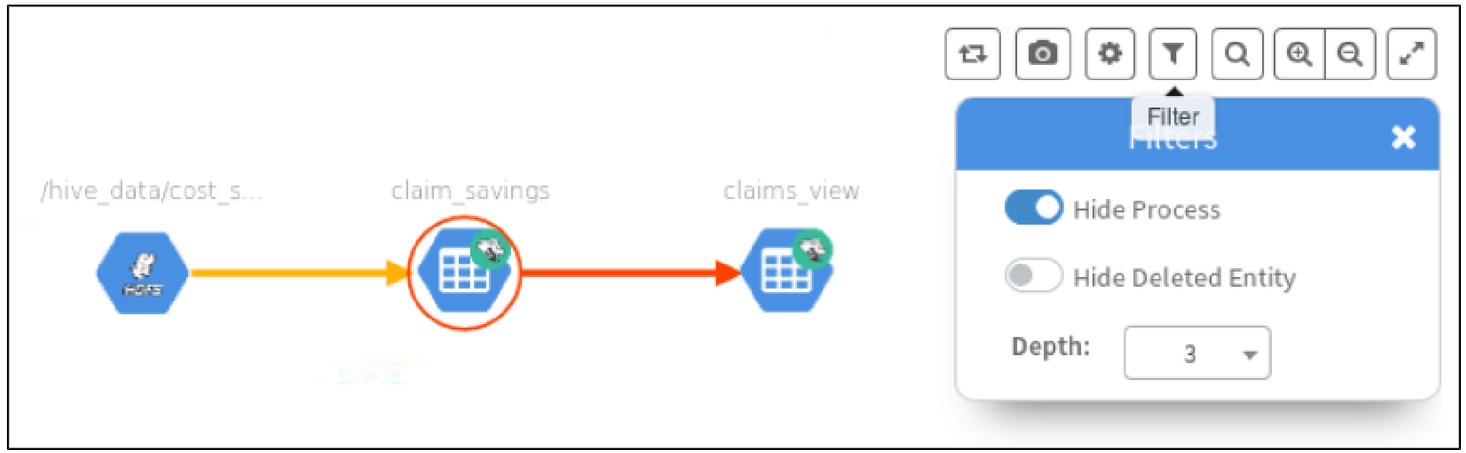
Viewing Entity Details from Lineage

- **Click on a node for more detail**
- **Includes the guid as a link to that node's entity page**
- **Also shows classifications and terms attached to the entity, if any**



Filtering

- **Show or hide**
  - Process nodes
  - Deleted entities
    - Depth determines how many assets (excluding processes) before and after will be shown



Ranger Policies for Apache Atlas

- **Ranger supplies access control for Apache Atlas**
- **Some example policies are provided by default to the following users and groups:**
  - **admin**: the initial Atlas administrator user has full access to all Atlas actions
  - **rangertagsync**: the TagSync service user has read access to entity metadata
  - **rangerlookup**: the Ranger lookup service user has read access to entity metadata
  - **public**: all users are granted access to read Atlas entity metadata
  - {**USER**}: users can save searches for later Atlas sessions

Essential Points

- **SDX holds the state (structure and business context) of all data**
- **Data Catalog provides access to a Data Lake and Workloads**
- **Policies are created in Ranger to provide the rules to allow or deny access to an entity**
- **Use Altas to display lineage**
  - Provides the history of entities and processes

Hands-On Exercise: Data Access

- **In this exercise, you will**
  - View policies and identify users that have access to specific data or services
  - Inspect the lineage details of data assets
- **Please refer to the Hands-On Exercise Manual for instructions**

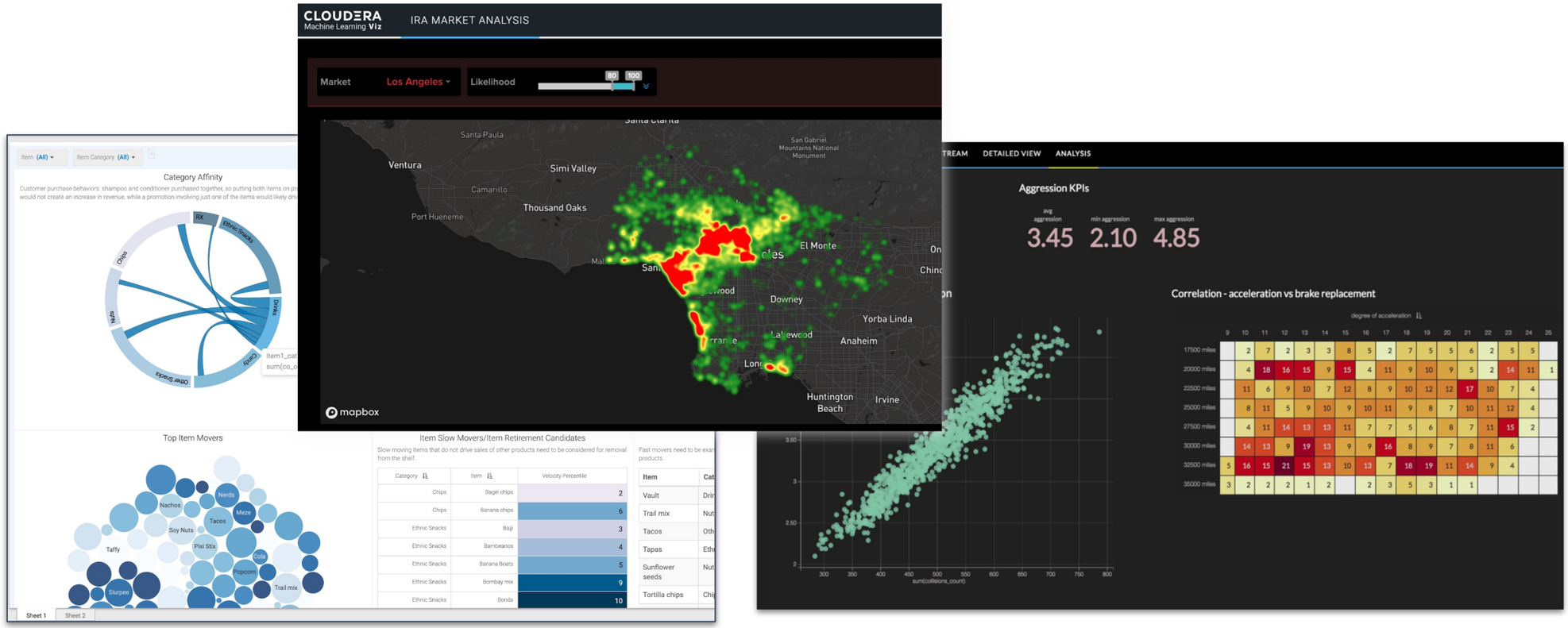
CLOUDERA

Data Visualization

Data Visualization in CML

- **By the end of this chapter, you will be able to**
  - Understand the importance of data visualization in the context of data science
  - List the concepts used in the CML Data Visualization application
  - Create your own dashboard in CML Data Visualization

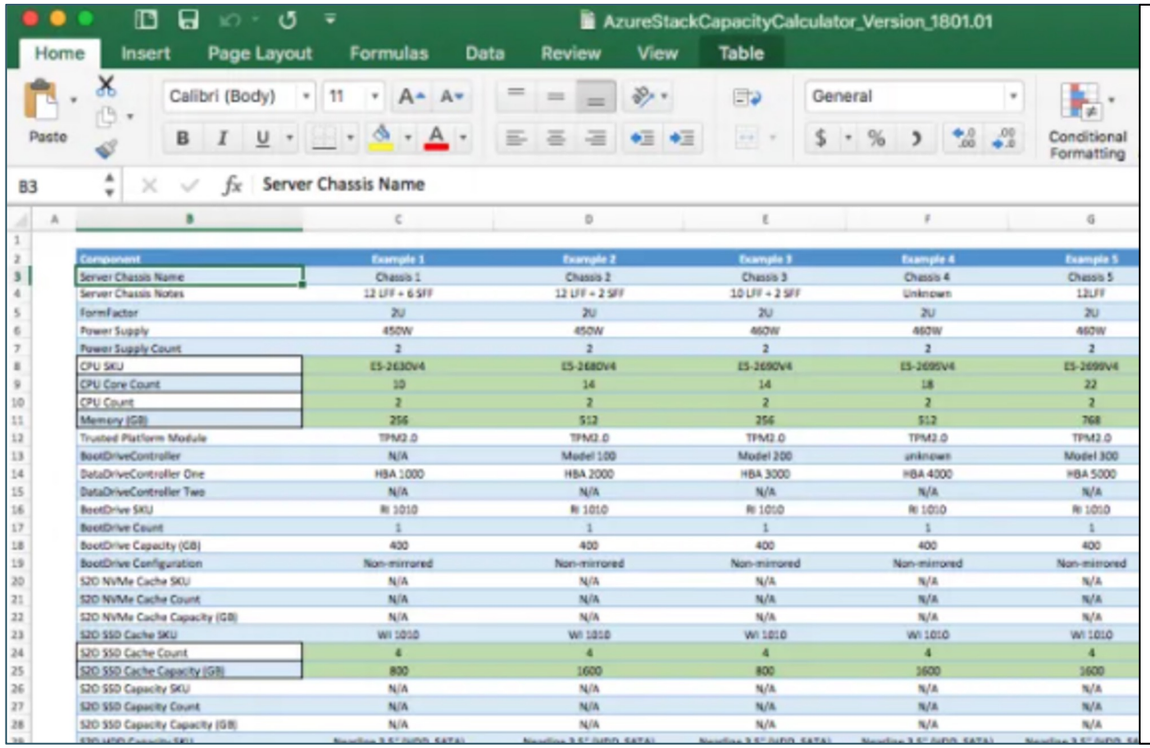
What is CDP Data Visualization?



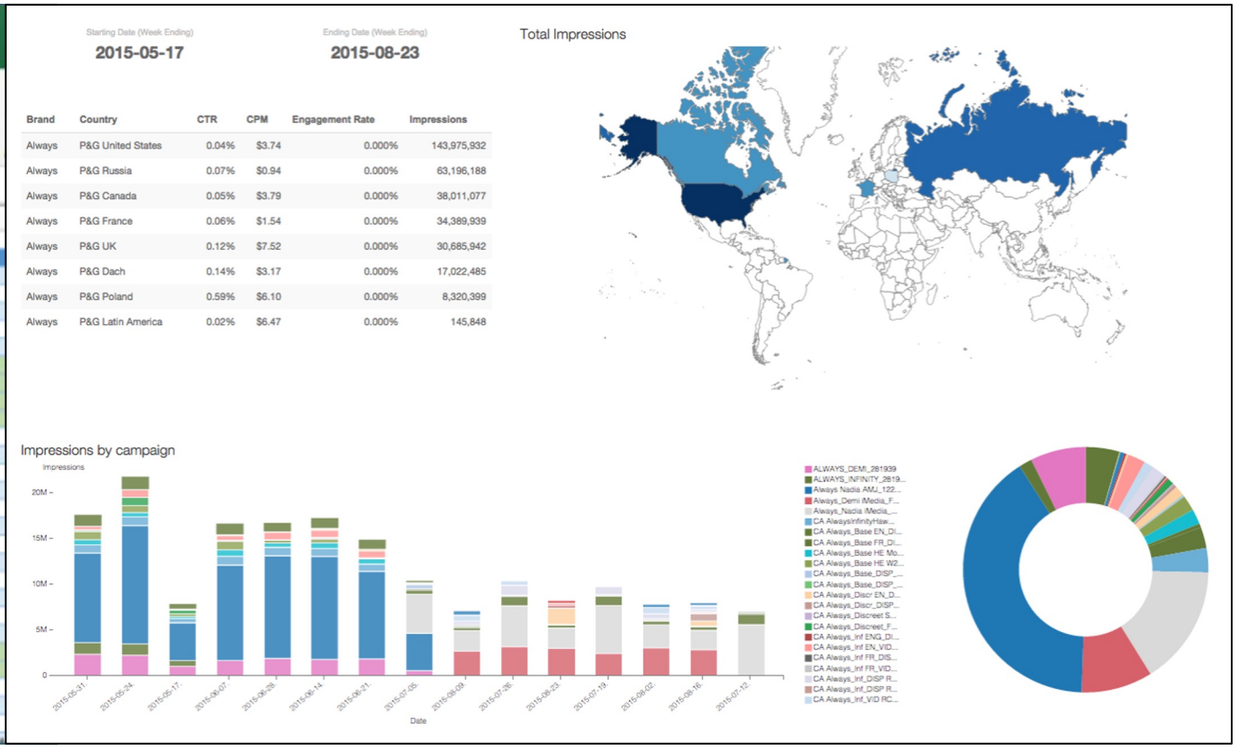
CDP Data Visualization (CDV) enables data engineers, business analysts, and data scientists to **quickly and easily** explore data, collaborate, and communicate **explainable insights across the data lifecycle.**

Visualizations Are Essential

- **Table-based data is great for calculation and organization, but hard to use for decision making when working with large sets of data**
- **Data visualizations enable humans to make inferences and draw conclusions about large sets of data based on visual input alone**

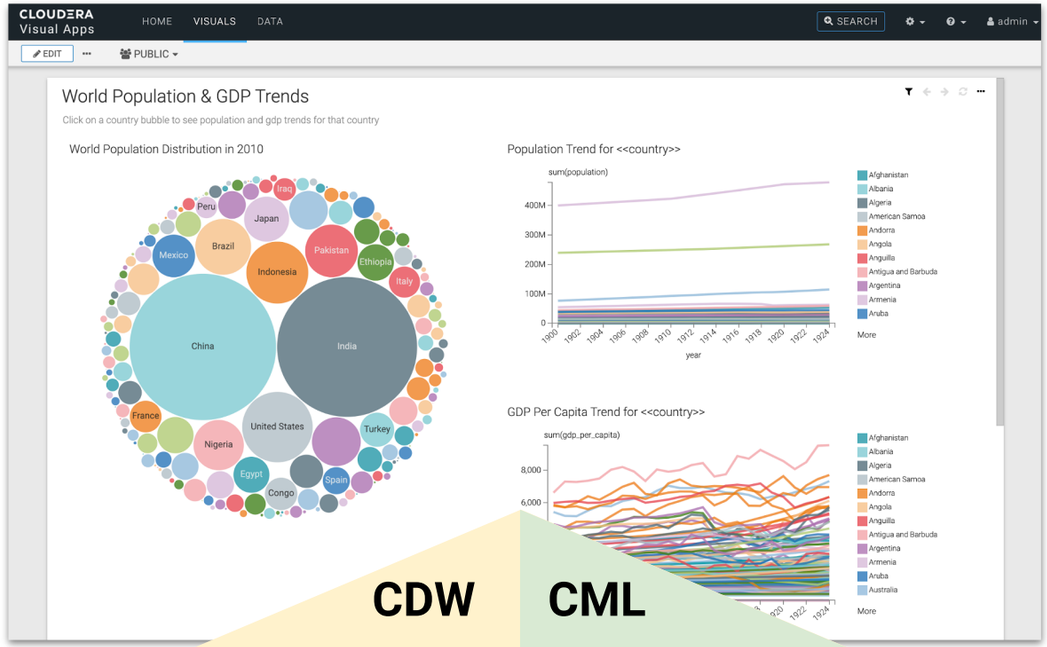


Component	Example 1	Example 2	Example 3	Example 4	Example 5
Server Chassis Name	Chassis 1	Chassis 2	Chassis 3	Chassis 4	Chassis 5
Server Chassis Notes	12 LFF + 6 SFF	12 LFF + 2 SFF	10 LFF + 2 SFF	Unknown	12LFF
FormFactor	2U	2U	2U	2U	2U
Power Supply	450W	450W	460W	460W	460W
Power Supply Count	2	2	2	2	2
CPU SKU	E5-2630V4	E5-2680V4	E5-2690V4	E5-2695V4	E5-2699V4
CPU Core Count	10	14	14	18	22
CPU Count	2	2	2	2	2
Memory (GB)	256	512	256	512	768
Trusted Platform Module	TPM2.0	TPM2.0	TPM2.0	TPM2.0	TPM2.0
BootDriveController	N/A	Model 100	Model 200	unknown	Model 300
DataDriveController One	HBA 1000	HBA 2000	HBA 3000	HBA 4000	HBA 5000
DataDriveController Two	N/A	N/A	N/A	N/A	N/A
BootDrive SKU	RI 1010	RI 1010	RI 1010	RI 1010	RI 1010
BootDrive Count	1	1	1	1	1
BootDrive Capacity (GB)	400	400	400	400	400
BootDrive Configuration	Non-mirrored	Non-mirrored	Non-mirrored	Non-mirrored	Non-mirrored
S2D NVMe Cache SKU	N/A	N/A	N/A	N/A	N/A
S2D NVMe Cache Count	N/A	N/A	N/A	N/A	N/A
S2D NVMe Cache Capacity (GB)	N/A	N/A	N/A	N/A	N/A
S2D SSD Cache SKU	WI 1010	WI 1010	WI 1010	WI 1010	WI 1010
S2D SSD Cache Count	4	4	4	4	4
S2D SSD Cache Capacity (GB)	800	1600	800	1600	1600
S2D SSD Capacity SKU	N/A	N/A	N/A	N/A	N/A
S2D SSD Capacity Count	N/A	N/A	N/A	N/A	N/A
S2D SSD Capacity Capacity (GB)	N/A	N/A	N/A	N/A	N/A



Brand	Country	CTR	CPM	Engagement Rate	Impressions
Always	P&G United States	0.04%	$3.74	0.000%	143,975,932
Always	P&G Russia	0.07%	$0.94	0.000%	63,196,188
Always	P&G Canada	0.05%	$3.79	0.000%	38,011,077
Always	P&G France	0.06%	$1.54	0.000%	34,389,939
Always	P&G UK	0.12%	$7.52	0.000%	30,685,942
Always	P&G Dach	0.14%	$3.17	0.000%	17,022,485
Always	P&G Poland	0.59%	$6.10	0.000%	8,320,399
Always	P&G Latin America	0.02%	$6.47	0.000%	145,848

Bring All Your Data Together



CDW

Real Time Event Store

Real Time Data Mart

Data Discovery & Exploration

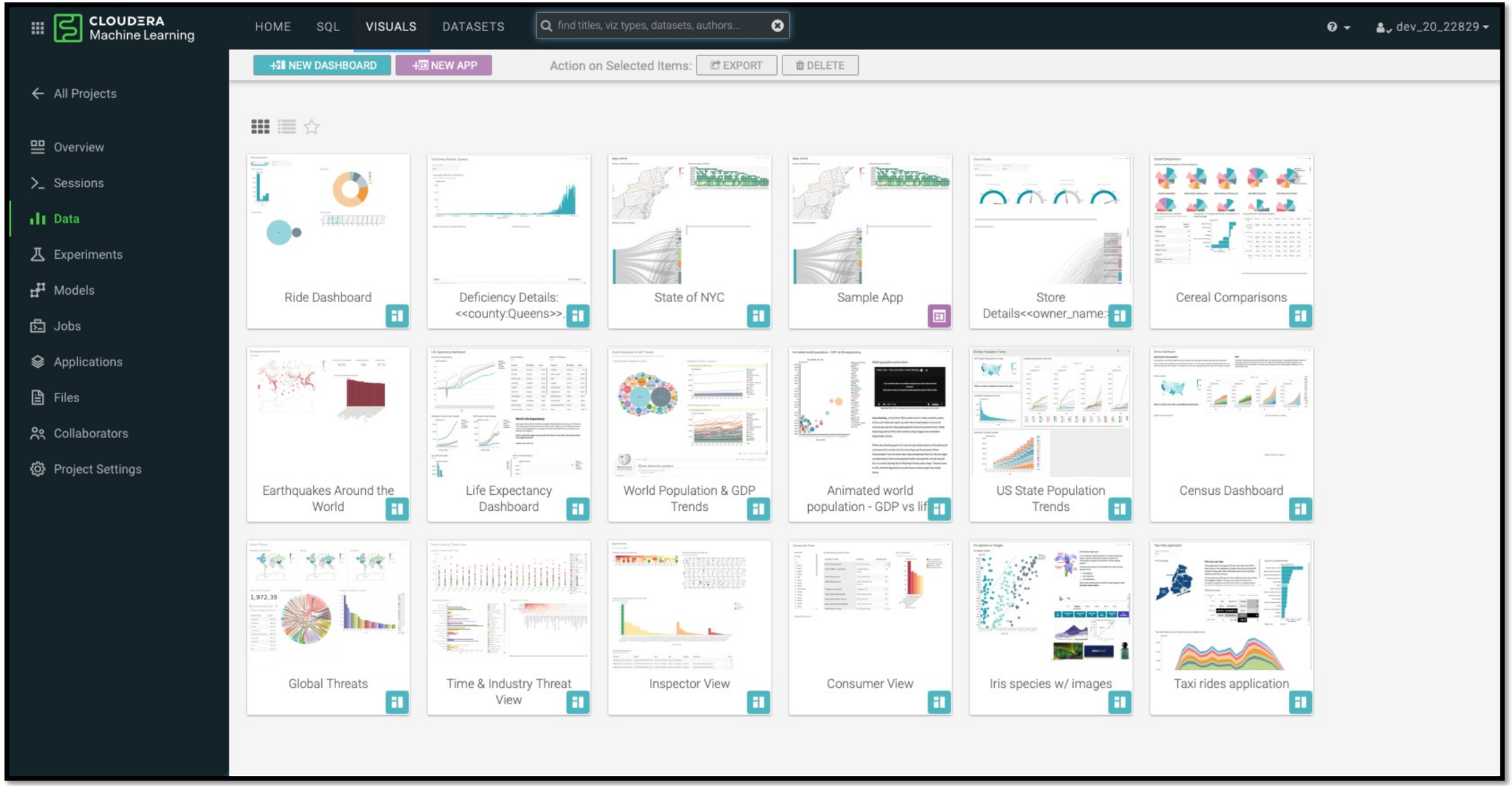
CML

Data Stream Analytics

Predictive Models

Data Science

Native Data Visualizations in CML

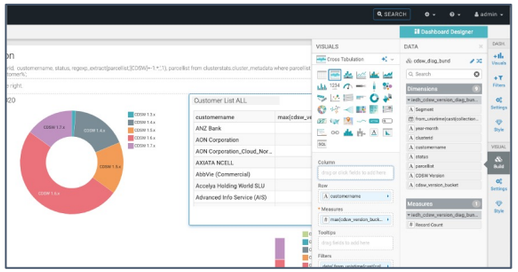


Integrated Data Visualizations for ML Workflows

Create fast, explainable insights from ML models

1

Exploratory data analysis

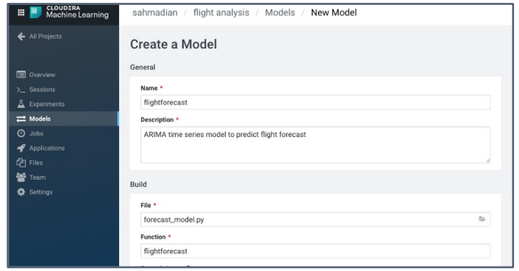


2

Develop predictive models

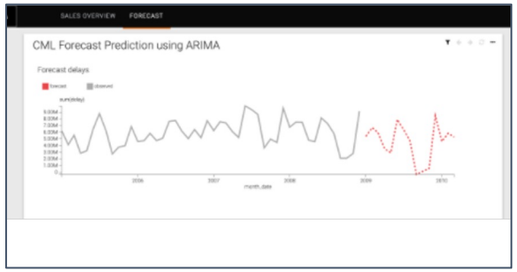
3

Deploy & serve models



4

Call & visualize model predictions within interactive dashboards

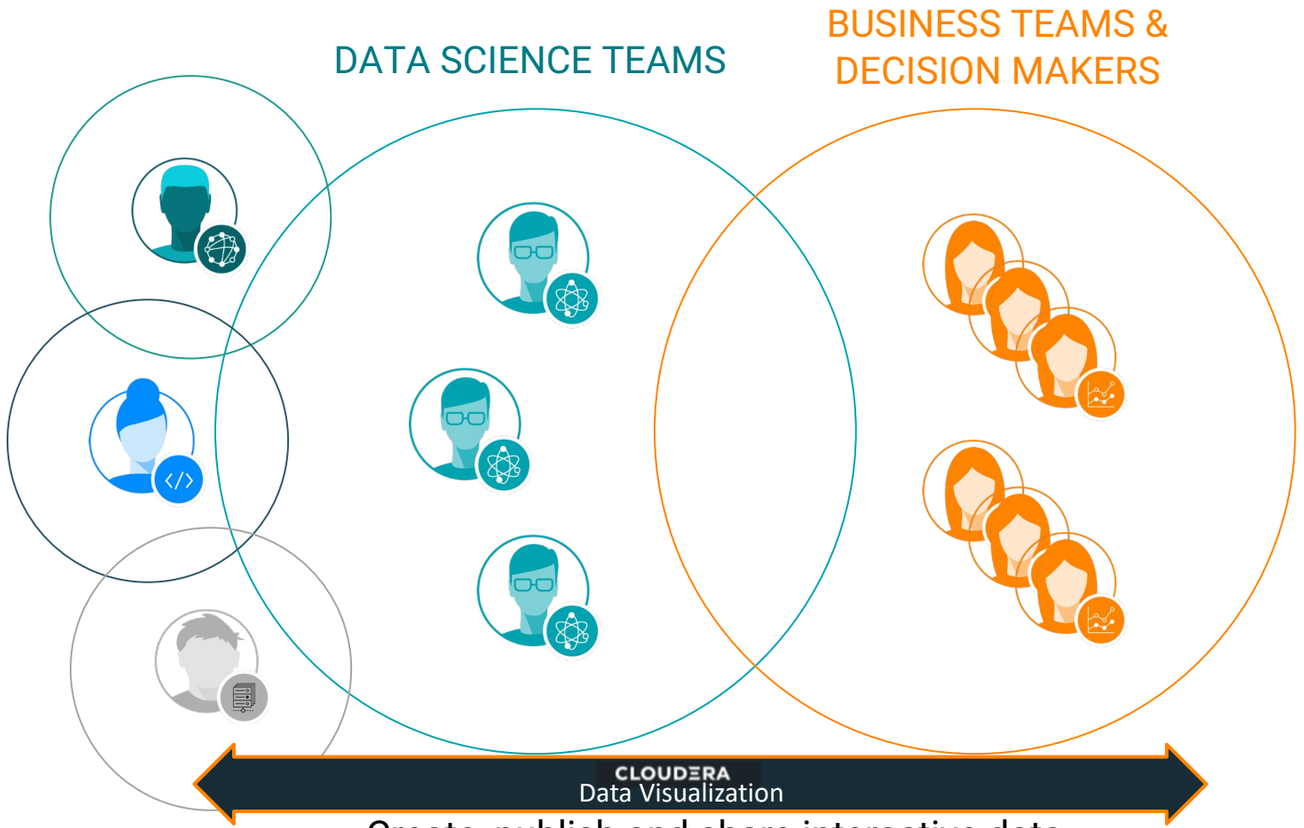


Visual Collaboration for Getting to Production

Accelerating Production ML Workflows from Raw Data to Business Impact

Communicate & collaborate effectively across the data lifecycle

Experiment, train, deploy models across the business



Leverage Explainable AI

Make informed business decisions

Create, publish and share interactive data visualizations

Visualize, Share, and Collaborate

FAST, SELF-SERVICE DATA EXPLORATION

- Intuitive drag & drop UI
- Integrated with CDP, no moving data or data silos
- Inherently secure with SDX - no data extraction needed

EXPEDITE CROSS-TEAM COLLABORATION

- Self-service for everyone
- Fast insight sharing across the lifecycle
- Same sharable experience everywhere

POWER ANALYTICAL AUTOMATION

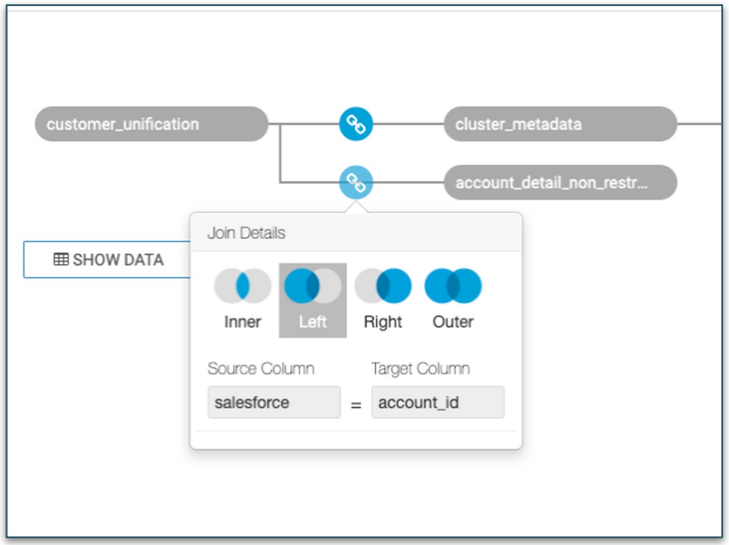
- Real time custom dashboards natively in CDP
- Visualize across the data lifecycle to discover optimization opportunities

Visualization Benefits

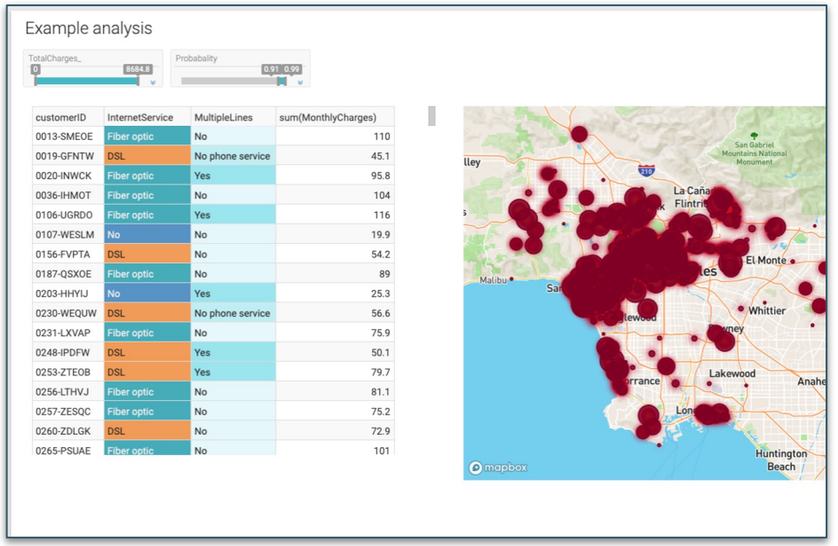
- **Work within the workflow of ML**
- **Simple drag & drop interface**
- **Interactive, dynamic dashboard**
- **Completely web based for integration through links, embedding, HTML/JS**
- **Many ways to share and collaborate**
  - Bookmark link
  - Emailed reports
  - Customized applications
- **Access to all of CDP sources (Impala, Hive, ML models, etc.)**

Build, Publish, and Share

Visual Types & Custom Extensions



Data Modeling for sharable logical semantic layer



UI-powered custom interactive dashboards & applications

Concepts

- Connections
- Datasets
- Visuals
- Dashboards
- Applications

Connections

- **Create and manage connections to many types of external data sources such as**
  - SQL (Impala, Hive, MySQL)
  - Events and time series data (Impala over Kudu)
  - Unstructured data (Solr)
  - ML workloads (Spark)
- **Using Cloudera Machine Learning (CML), you can**
  - Connect to an Impala or a Hive data warehouse
  - Tie in data from predictive CML models
- **Using CDP Public Cloud with Cloudera Data Warehouse (CDW)**
  - The data connection is automatically set up, but you can connect to other data sources as well

Supported Types

- Impala
- Hive
- Druid
- MariaDB
- MySQL
- PostgreSQL
- Solr
- Spark SQL
- SQL Stream Builder
- SQLite

Datasets

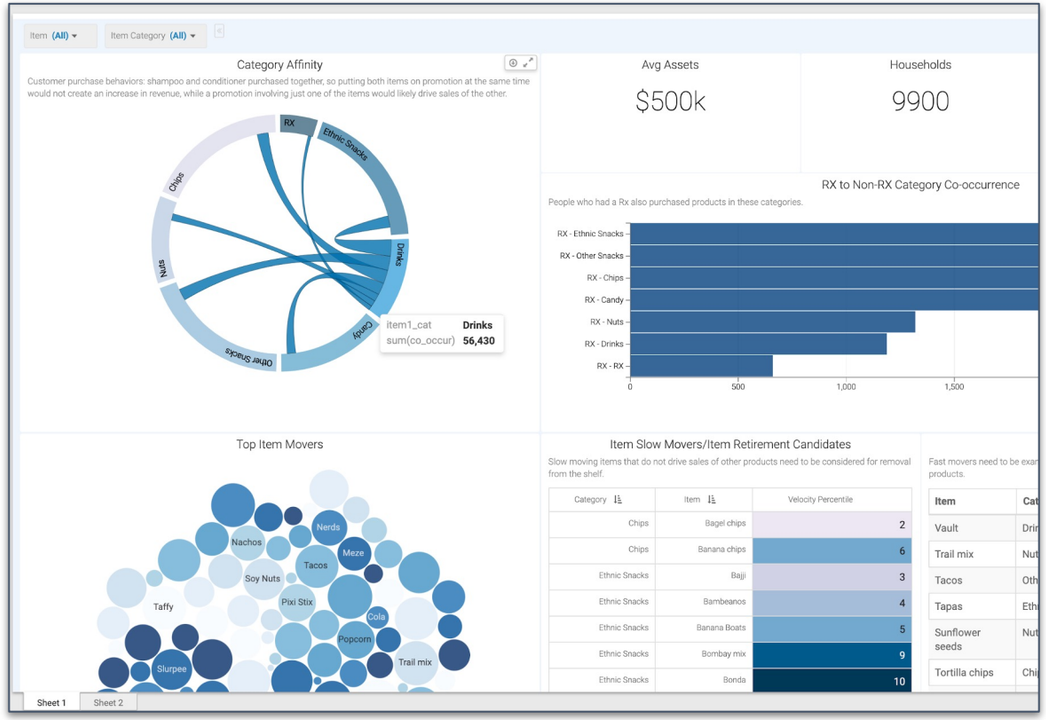
- **The foundation and starting point for visualizing your data**
  - The ***semantic layer on top of your data tables*** and views in the data store
  - Allows you to model it into what you need for your visual application without changing underlying data or tables
- **Represents a single data table or data matrix from several tables on the same connection**
- **Can be modeled using**
  - Table joins
  - Calculated fields
  - Modification of data types, dataset fields, and default aggregation of fields

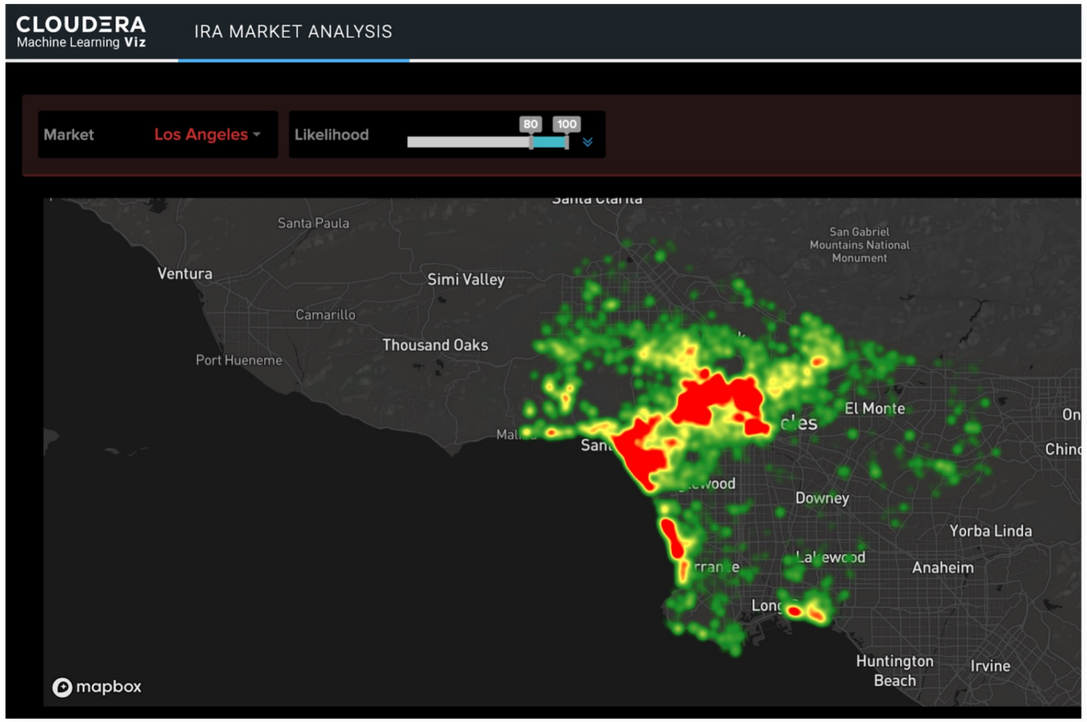
Visuals

- **A visual is a *single piece of visualized data*, for example**
  - Pie charts
  - Histograms
  - Heatmaps
- **It translates large data sets and metrics into a visual representations**
  - Easier to identify insights about the information represented in the data
- **CDP Data Visualization has a rich offering of different types of visualization to assist you in analyzing your data**

Dashboards

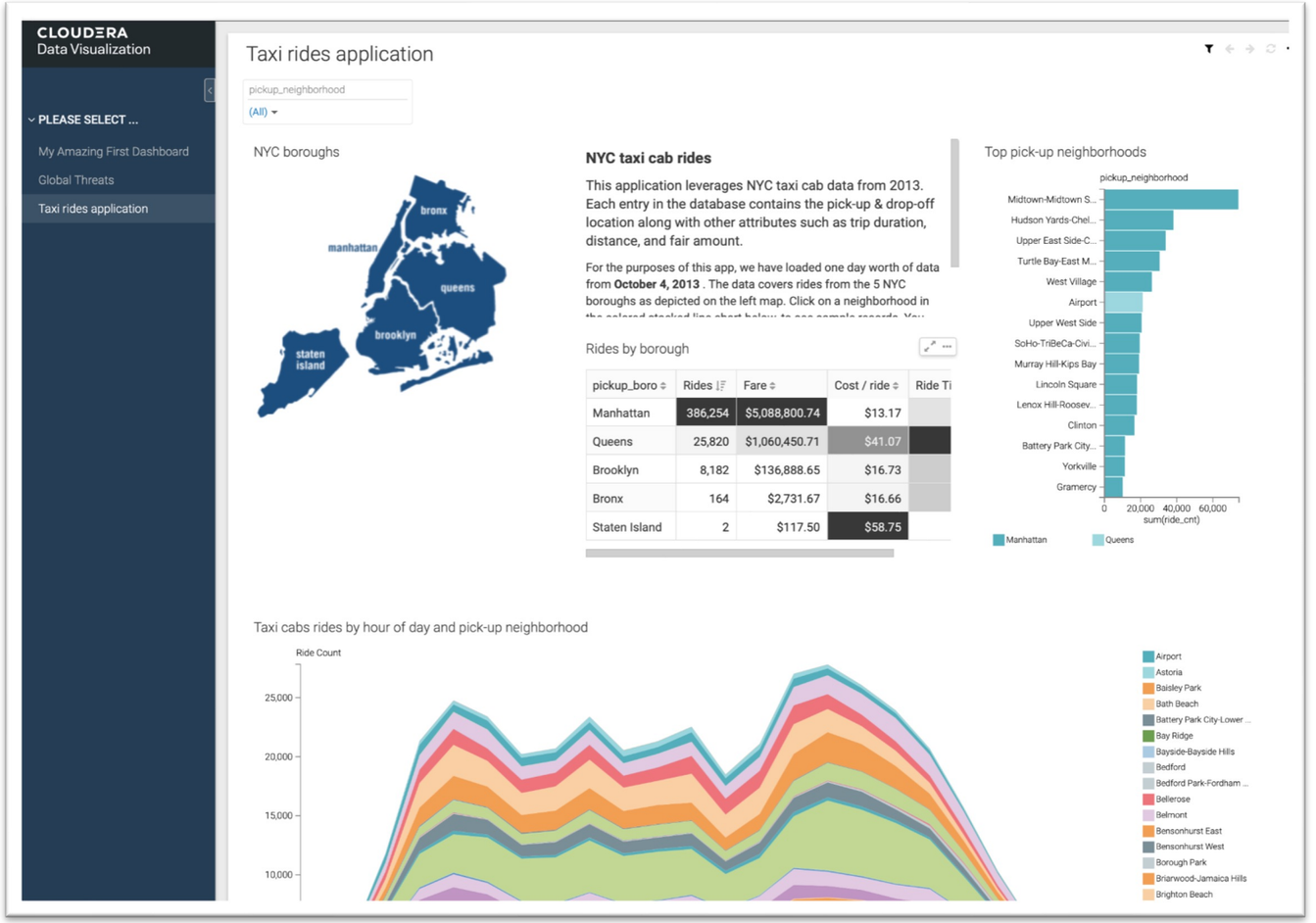
- **CDP Data Visualization dashboards consolidate related visualizations**
  - Display and link visuals that are based on different datasets across different connections
  - Provide optional run-time filtering on all referenced information





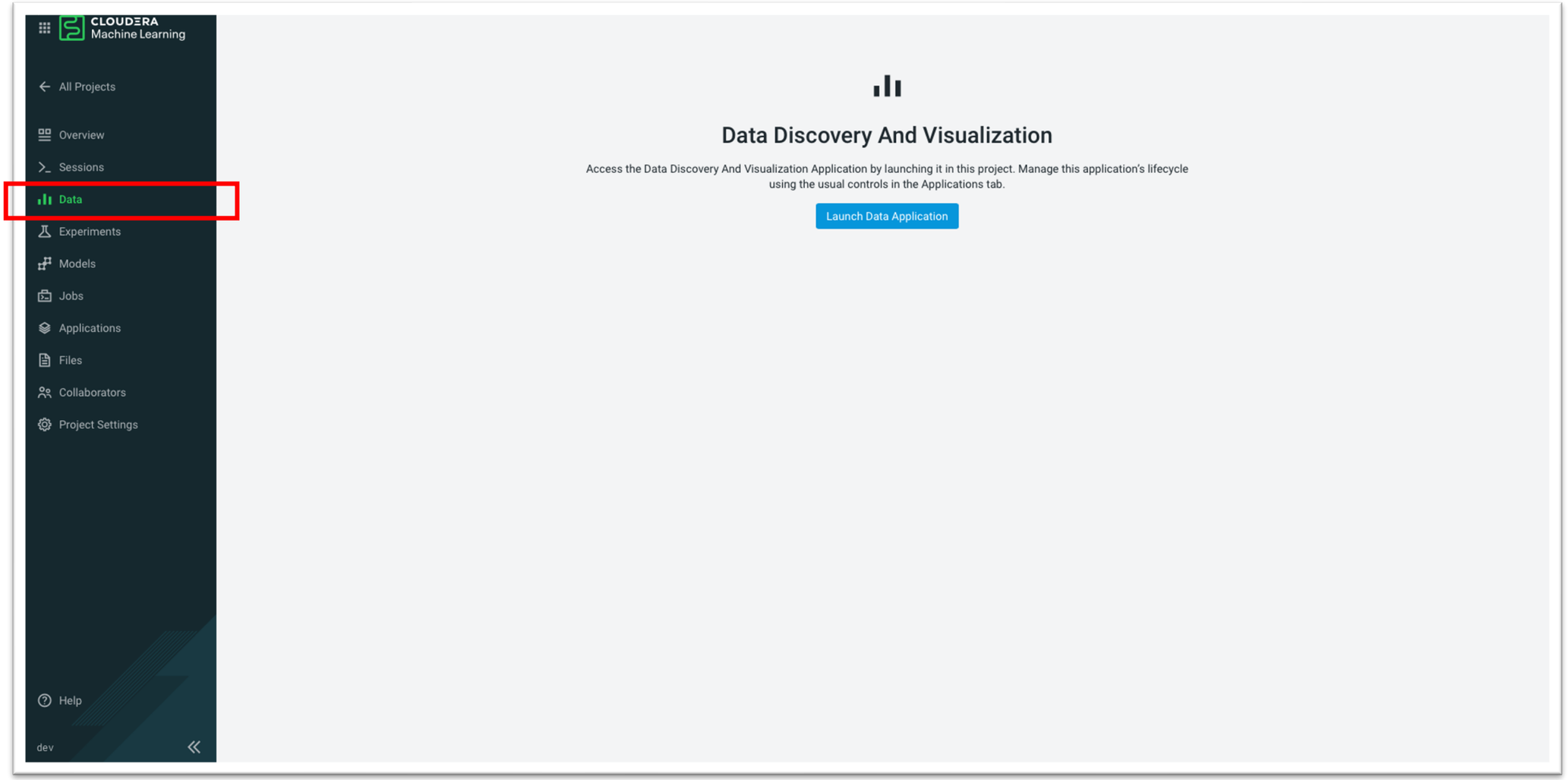
Applications (Apps)

- **A collection of dashboards tied together that can be launched as a standalone, branded data visualization tool**
- **The App Designer interface enables you to**
  - Build applications
  - Customize the tabs of the apps
  - Populate the tabs with relevant dashboards

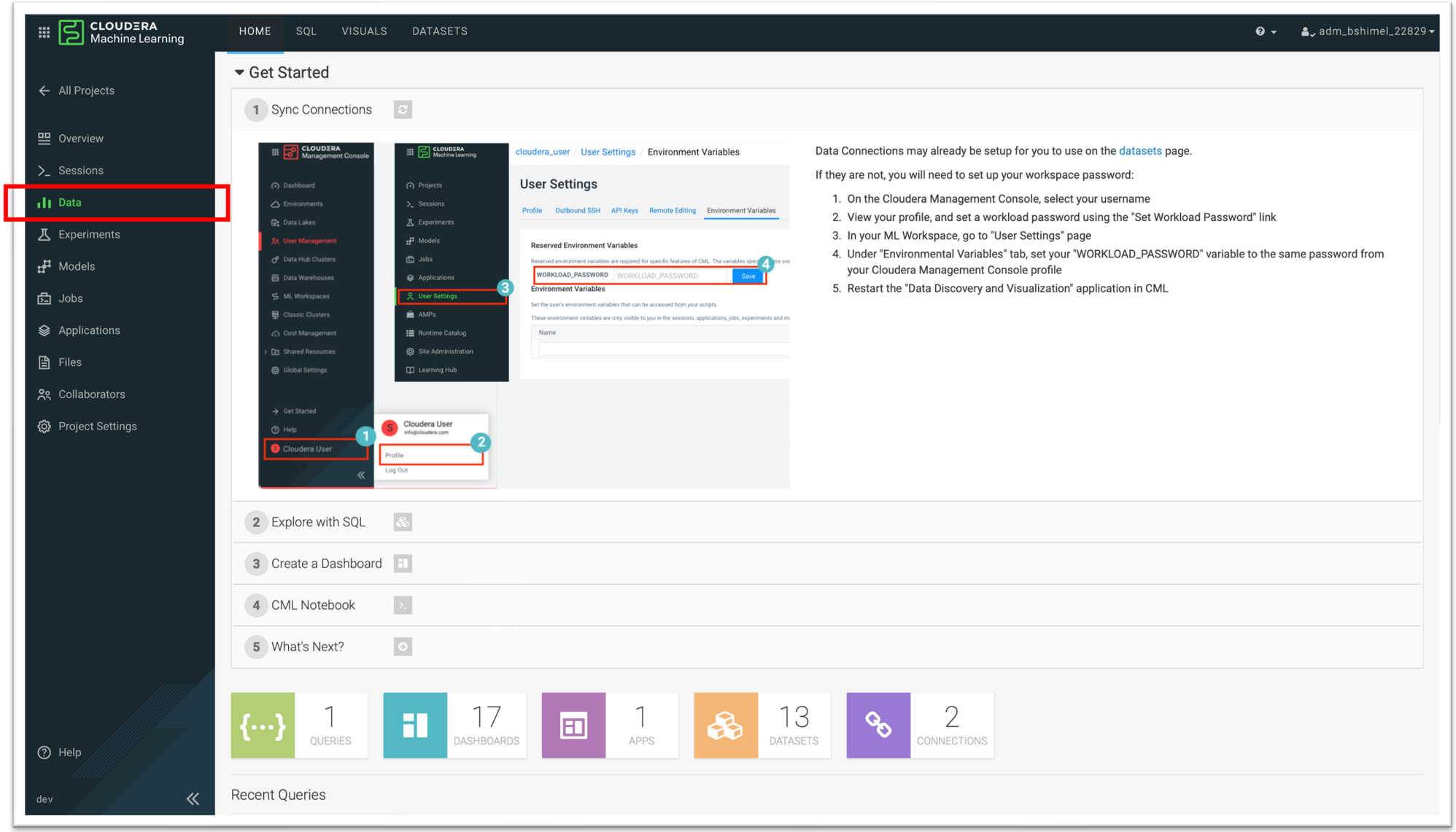


Creating Data Visualization Application

- **Create a new application that hosts visualization by selecting Data on the left sidebar**



CDV Home Page

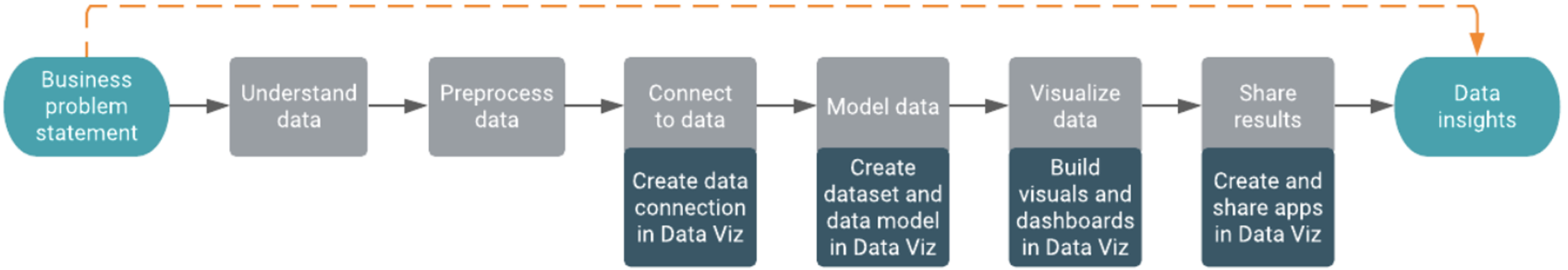


Home Page Views

- The interface of Data Visualization has three views:
  - HOME
  - VISUALS
  - DATA

General Workflow

1. Create a data connection
2. Create a dataset using your data connection
3. Create a dashboard based on your new dataset
4. Add visuals to your dashboard
5. Create an application to share your dashboard with business users



Essential Points

- **CDP Data Visualization provides out of the box functionality without additional integration efforts, moving data, or creating security issues**
  - Easily and quickly build interactive dashboards and instantly share insights across your business
  - Fully integrated data visualization from within ML Workflows
  - More than 34 visual types are available which help in representing the data in the most suitable format rather than using rows and columns to present the information
  - In addition to out-of-the-box visuals, you can also build applications using custom extensions

Hands-On Exercise: Build a Visualization Application

- **In this exercise, you will use the DuoCar data to**
  - Connect to the data
  - Create a dataset
  - Display data using different visuals
  - Create an application
- **Please refer to the Hands-On Exercise Manual for instructions**

CLOUDERA

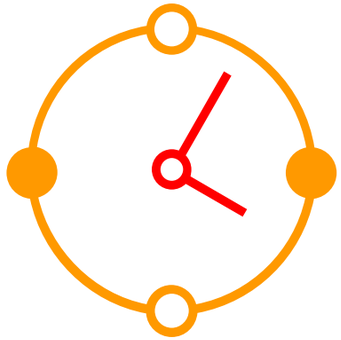
Data Warehouse Service

WHY CLOUDERA DATA WAREHOUSE?

Self-service analytics

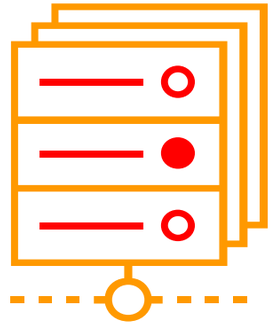
Businesses demand data democracy

Faster Insights for faster decisions



LOBs cannot wait

Access to modern data use cases



competitive

Benefits of Cloudera Data Warehouse

Zero to Query in minutes

Self-Service Data Analytics

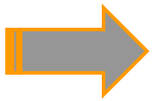
Unlimited Concurrency

Security & governance

Hybrid Flexibility

Autoscaling

Shared Data Catalog

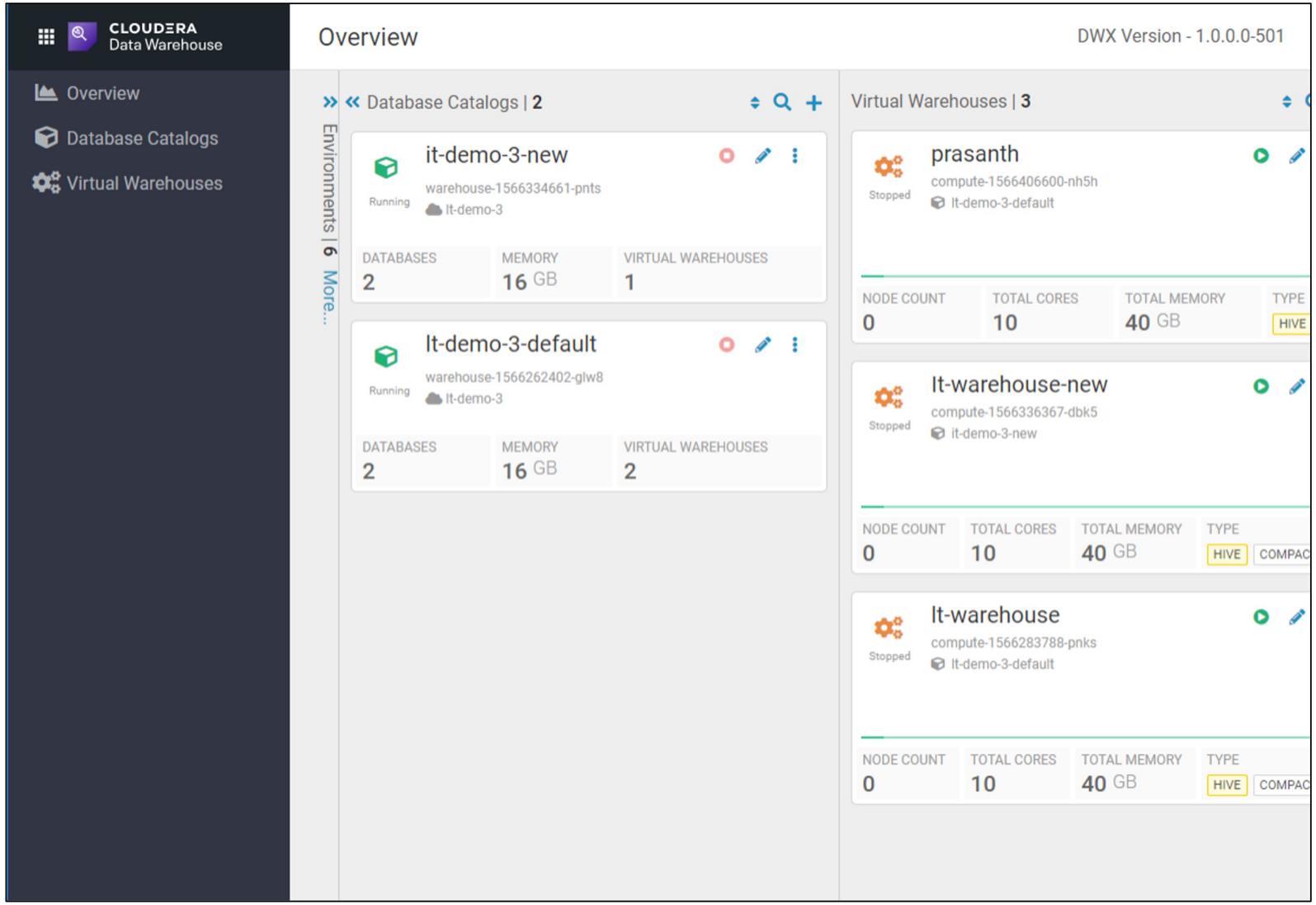


Agile, auto provisioning

Cloudera Data Warehouse

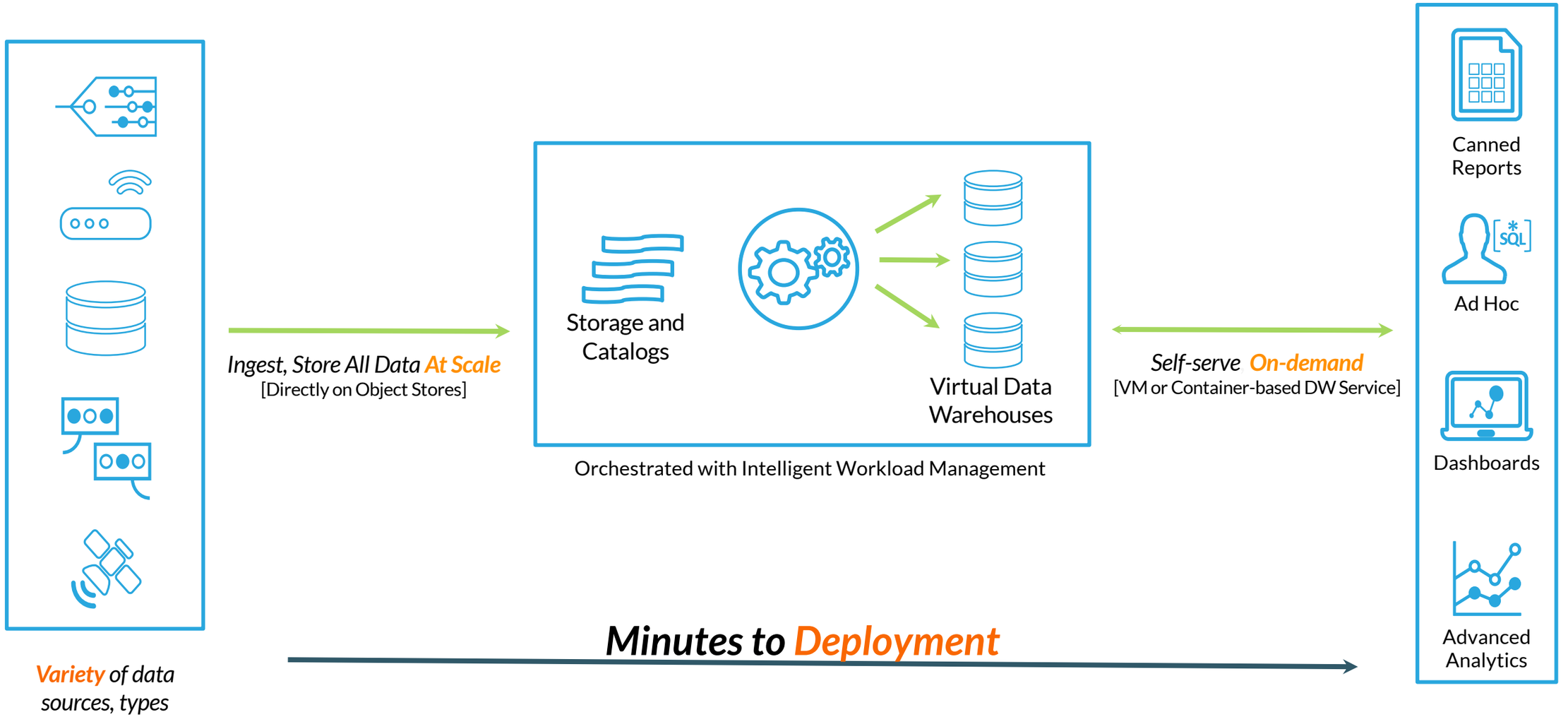
A data warehousing service optimized for concurrency, caching, and isolation

1. Quick Time to Value
   - Ease of Use
   - Integrated Visual Applications
   - Shared Data Catalog
   - Traditional & New Data Types
2. Lower Cost of Ownership
   - Multi-Engine & lowest Price/Performance
   - Elasticity, Automation & Intelligence
   - Workload Management



Elegant: Modern Data Warehouse

Guarantees faster time to value



Ease of Use

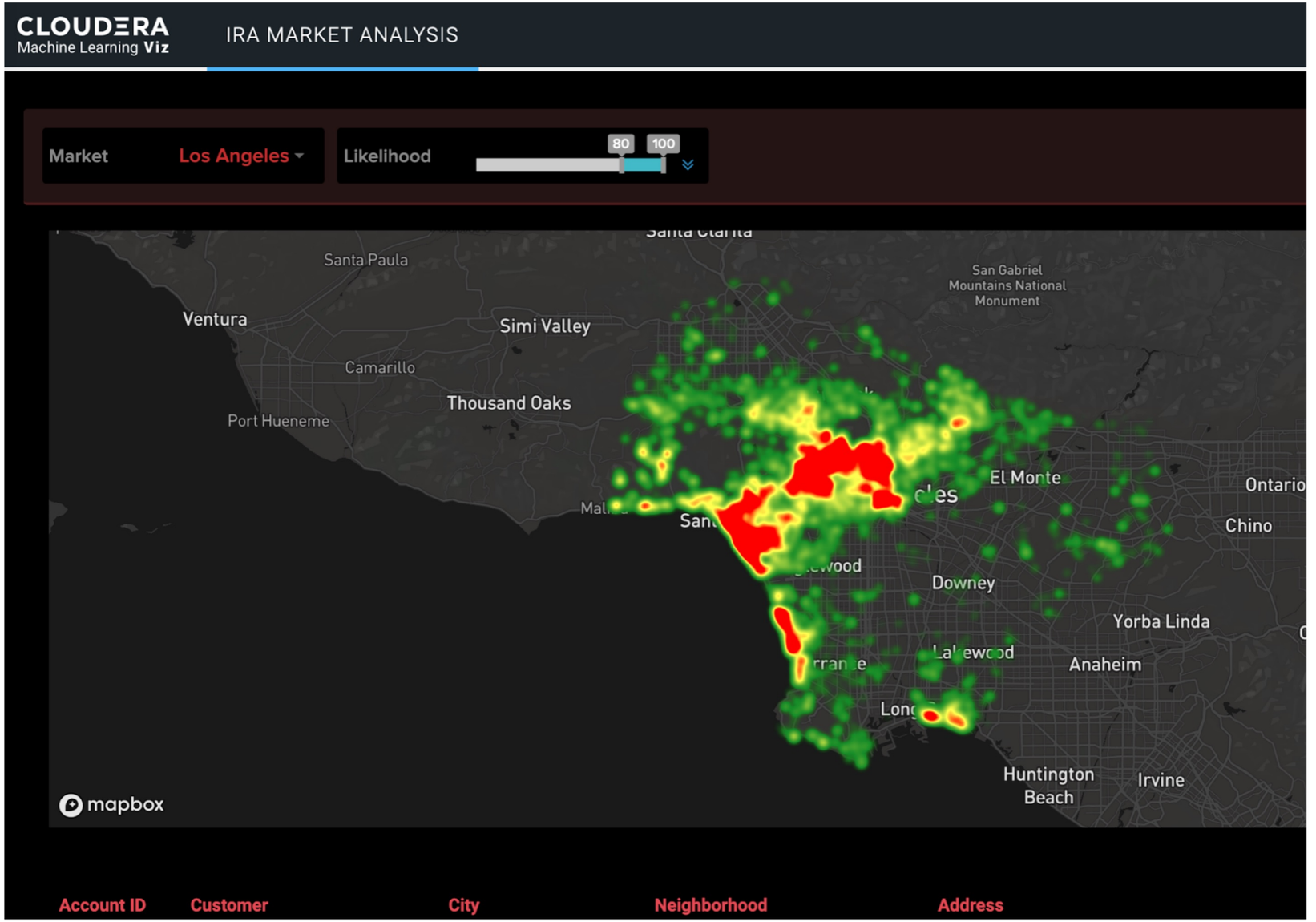
Zero to Query in 10 minutes or less

- **Simplified provisioning**
  - Relieve Central IT from LoB projects with the simplicity to enable self-service
- **Simplified data ingest**
  - Integrated Data Flow and Engineering enables any data at any speed
- **Elastic architecture**
  - Simplified capacity planning with automated scale up, down, & auto-suspend to ZERO nodes
  - Guardrails prevent runaway costs
- **Contention elimination**
  - Simplify resource allocation and workload optimization with isolated compute resources
  - Simplify SLA adherence, as users access the same data with independent compute clusters
  - Simplify data availability with sharing between experimental and production workloads

Visualization

Integrated visualization and dashboarding with CDW

- **Collection of dashboards styled for the organization that built it**
- **Consolidate related visualizations on same screen**
- **Easy collaboration by linking and sharing visuals**
- **Customized using styles and out of the box settings**



Shared Data Catalog

Organize & curate data

- **Find**
- **Interpret**
- **Trace**
- **Audit**
- **Profile**