

Cloudera Data Services





Introduction to Data Services

CLOUDERA Data Platform

Data Services



DataFlow Data Engineering Data Warehouse Operational Database Machine Learning

Data Management



Data Hub Clusters Data Catalog Replication Manager Workload Manager Management Console

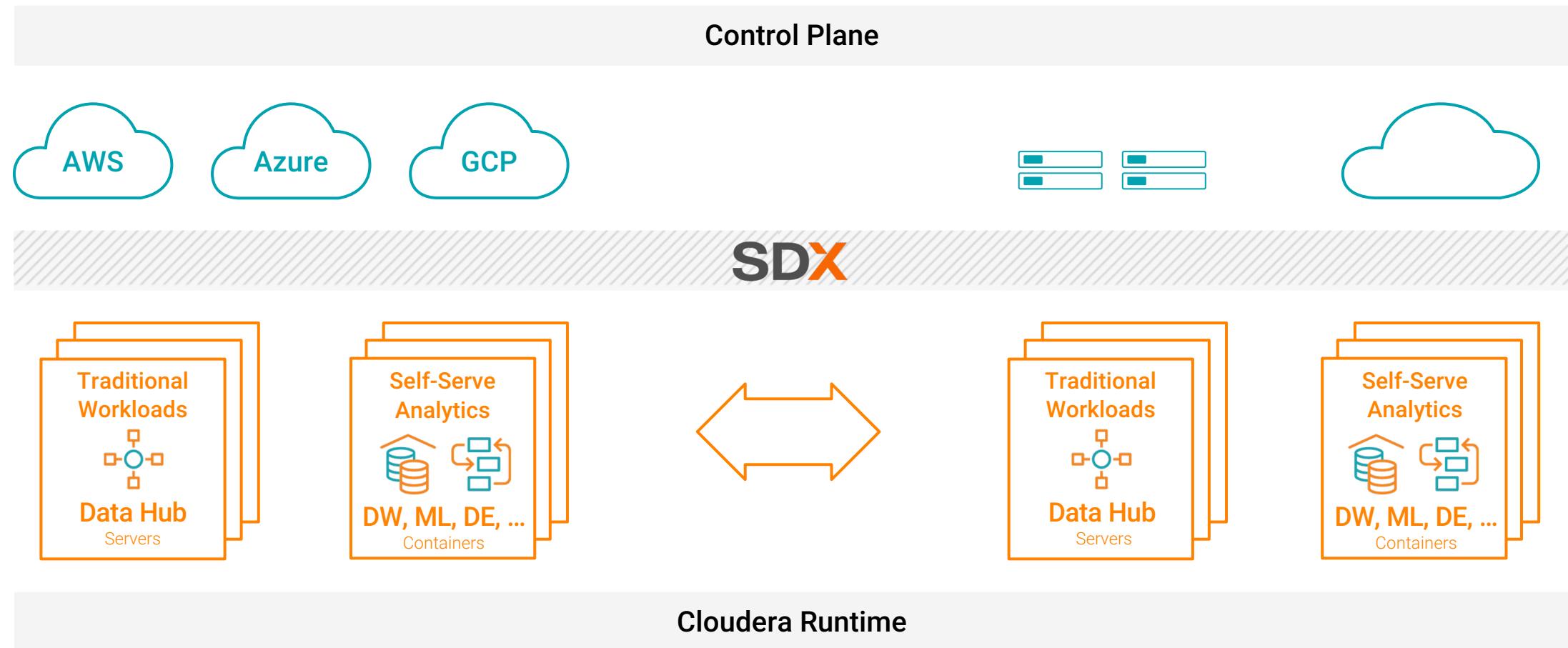
Feedback

Powered by Cloudera

One Platform – Two Form Factors

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

CDP Private Cloud (platform as installable software)



Key Concepts & Components

The screenshot shows the Cloudera Management Console interface. On the left is a dark sidebar with navigation links: Dashboard, Environments (which is selected and highlighted in blue), Data Lakes, User Management, Data Hub Clusters, Data Warehouses, ML Workspaces, and Classic Clusters. The main content area is titled "Environments / List" and shows a table of environments. The table has columns for Status, Name, Cloud Provider, and Region. There are 8 environments listed:

Status	Name	Cloud Provider	Region
Available	cdp-cldr-pri-env	aws	US East (Ohio)
Available	cldr-cdp-prim-az	aws	Central US
Available	wwbank-02	aws	US East (Ohio)
Available	cdp-cldr-prim-env	aws	US East(N. Virginia)
Available	cdp-cldr-bkup-env	aws	US East (Ohio)
Available	cdp-cldr-virginia-env	aws	US East(N. Virginia)
Available	all-se-env-1026	aws	US East (Ohio)
Available	prod-dw	aws	EU (Ireland)

ENVIRONMENTS



1:1



1:N



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

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

- 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.



Data Hub Clusters

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

ML, DW, DE, OD, DF



Data Warehouse



Machine Learning

What is CDP Private Cloud?



New set of data analytics applications

Featuring use-case optimized interfaces



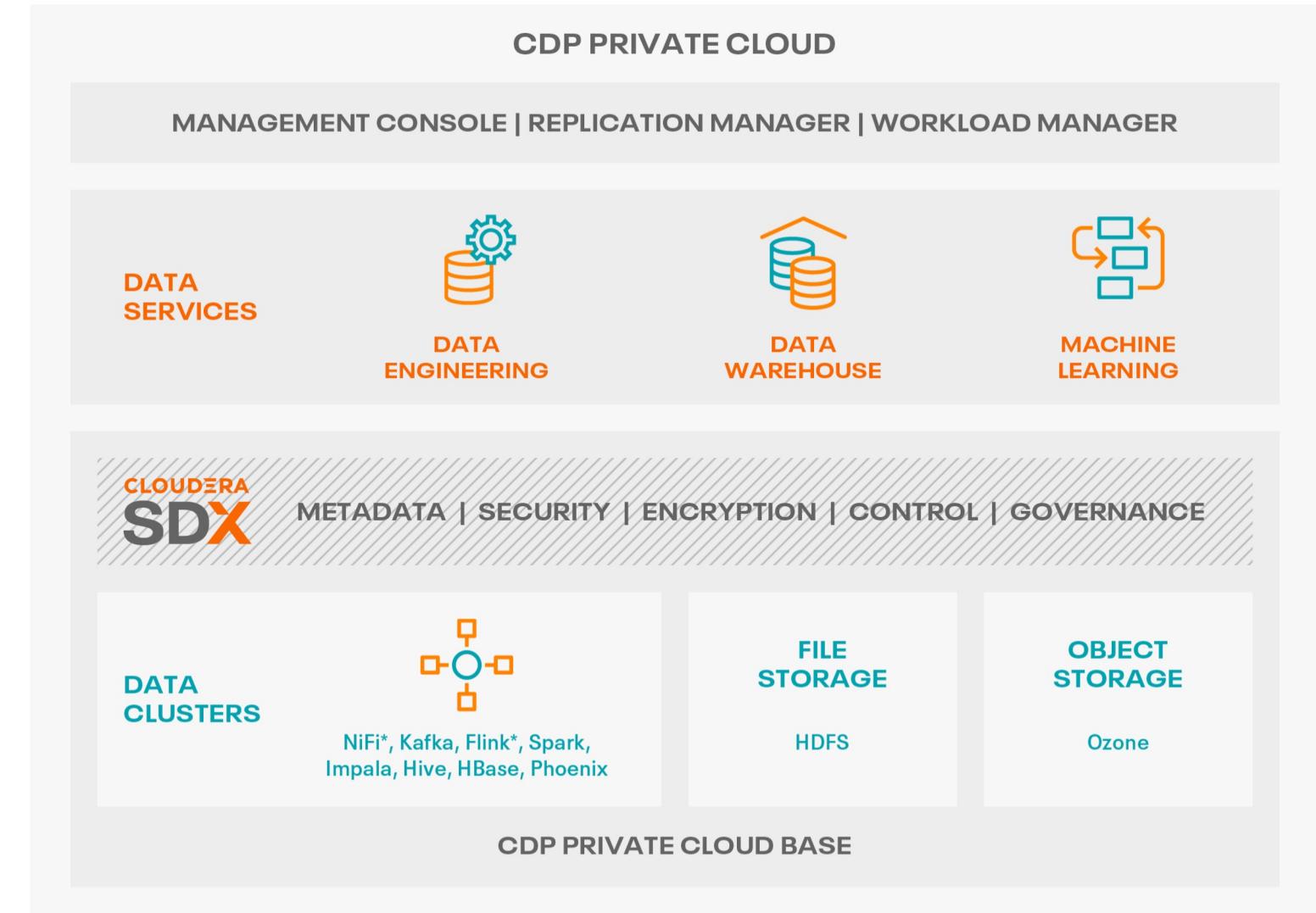
Running on its own container cloud

Simplified deployment, with fast provisioning and efficient scaling



With access to a shared data lake

That is secured and governed



*Standalone SKU

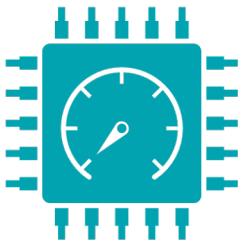
Impact of Challenges



Noisy
neighbors



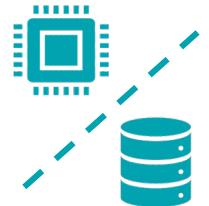
Complex
upgrades



Lack of
elasticity



Extended
onboarding
cycles



Co-location of
Storage &
Compute

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

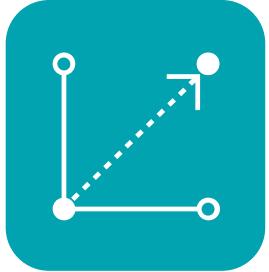
"We are still on 2.x, you'll have to wait to be able to use 3.x"

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

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

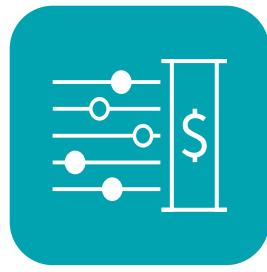
"We can't grow compute independently of storage"

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

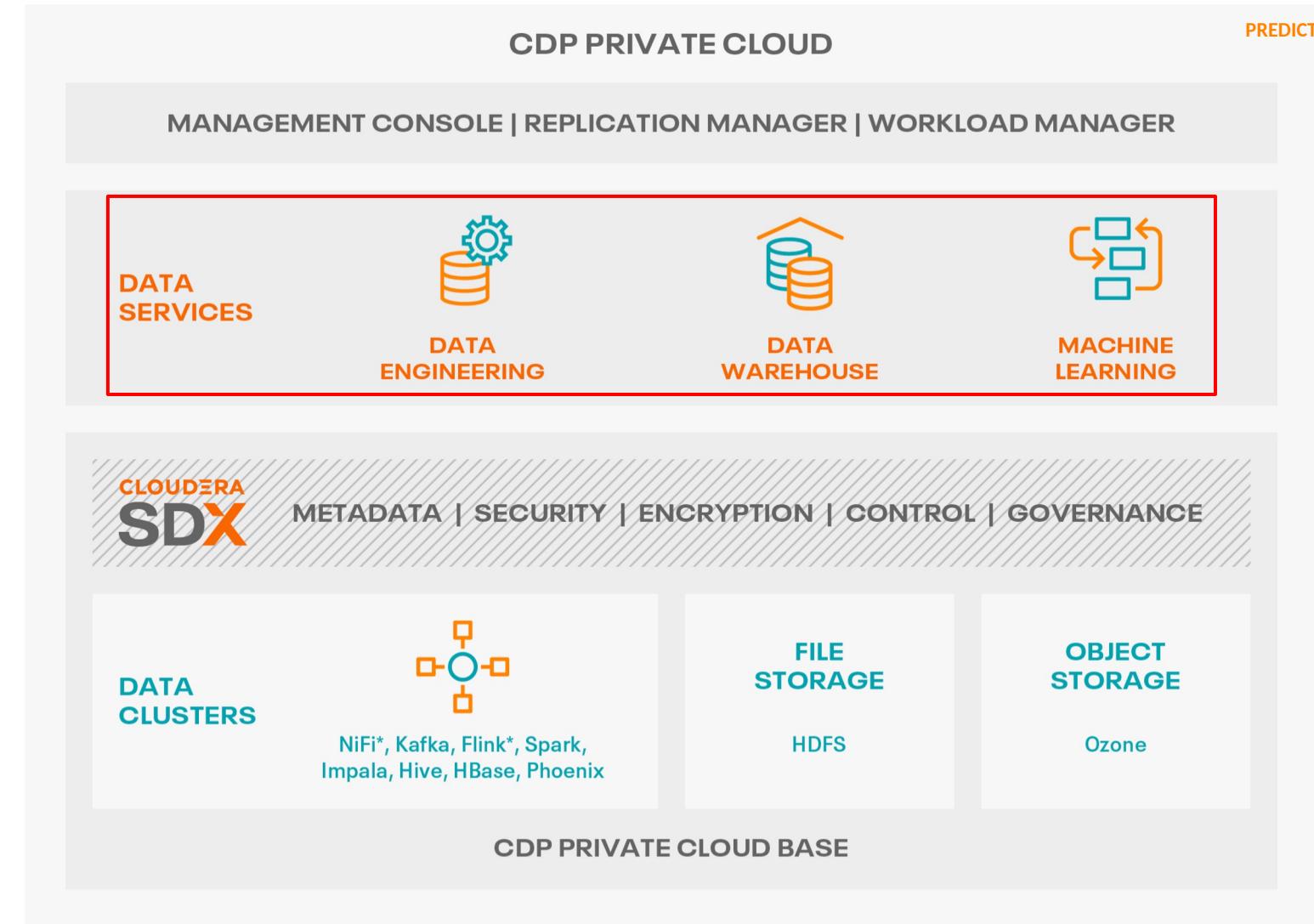
Simplified provisioning and onboarding of new use-cases

A Deeper Look at CDP Private Cloud

Predictable workload performance with tenant isolation



PREDICTABLE



*Standalone SKU

A Deeper Look at CDP Private Cloud

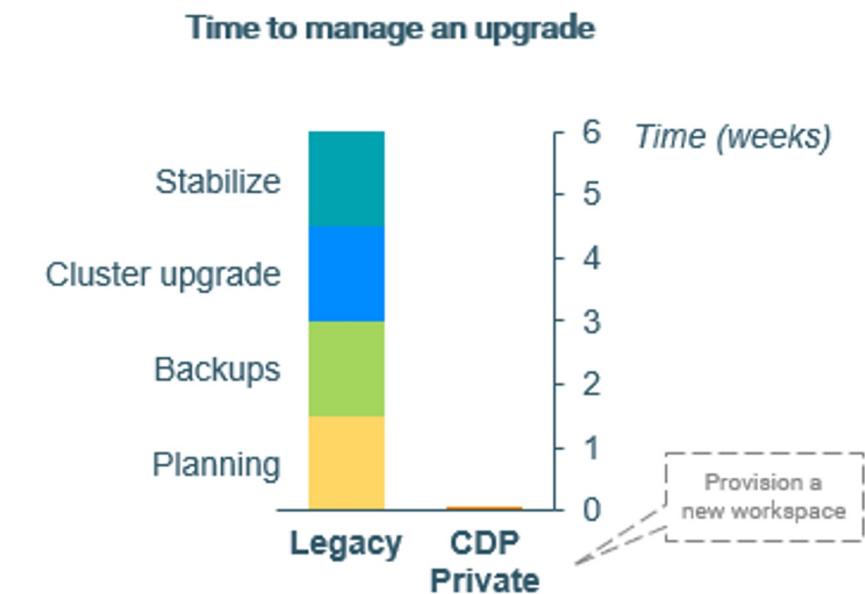
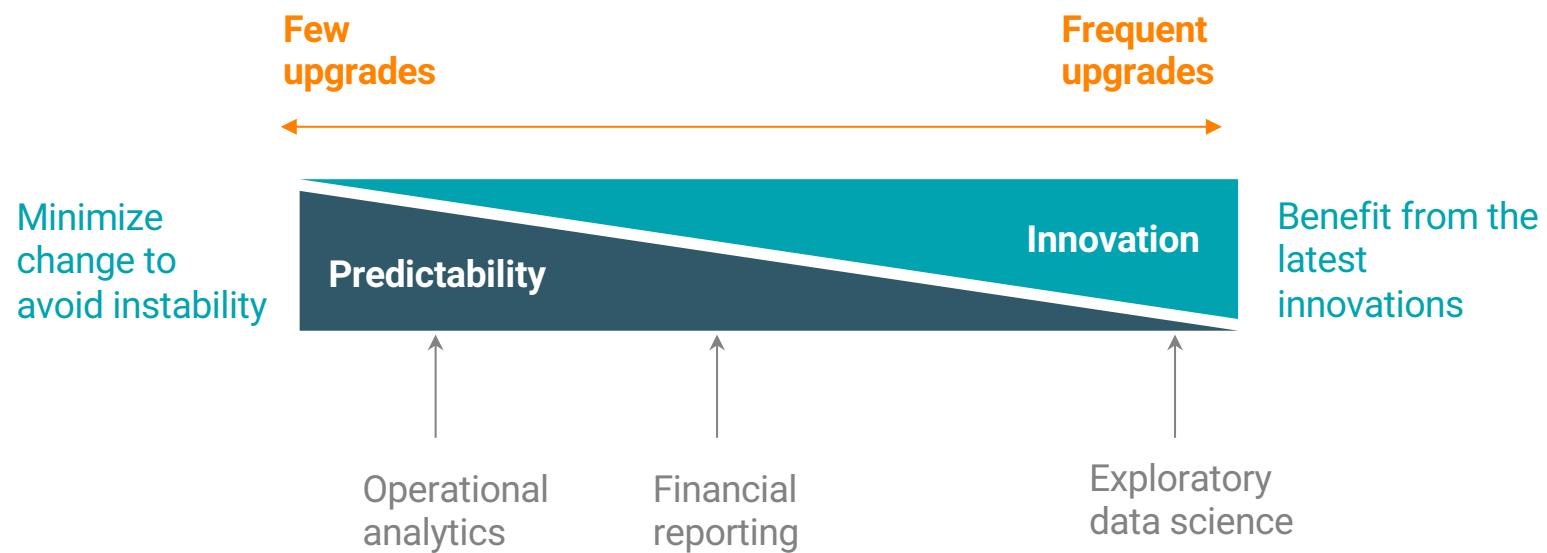
Better multi-tenancy for ‘upgrade agility’



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

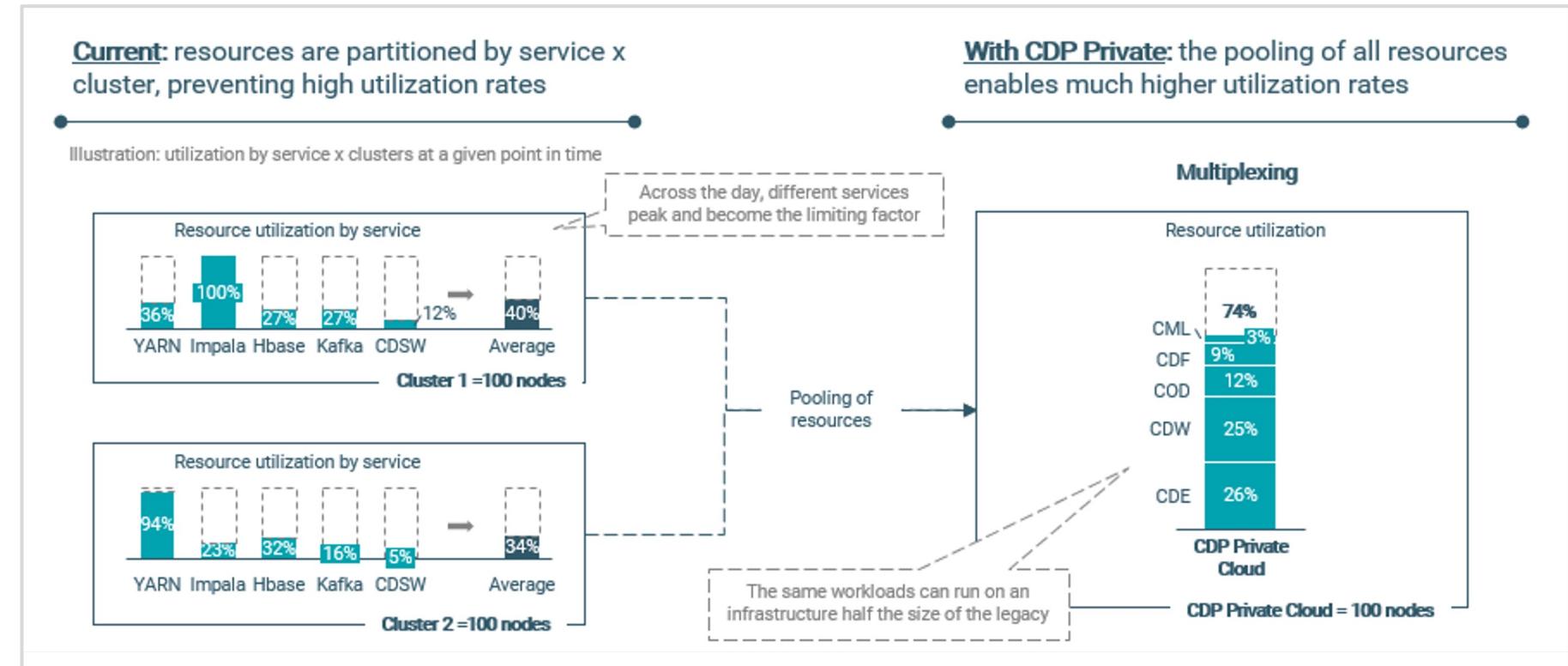
Performance without replicating data or creating silos



COST EFFICIENT

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



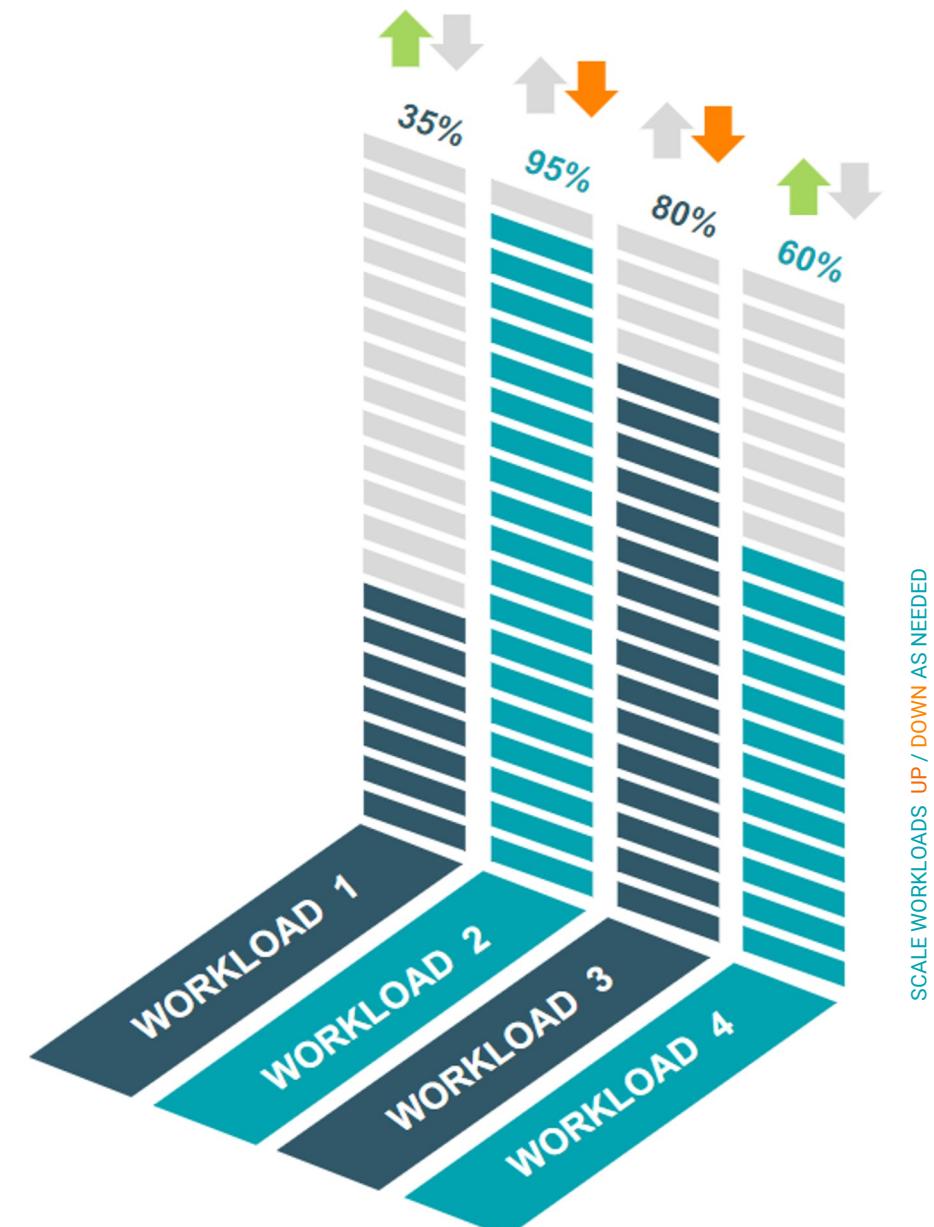
COST EFFICIENT

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



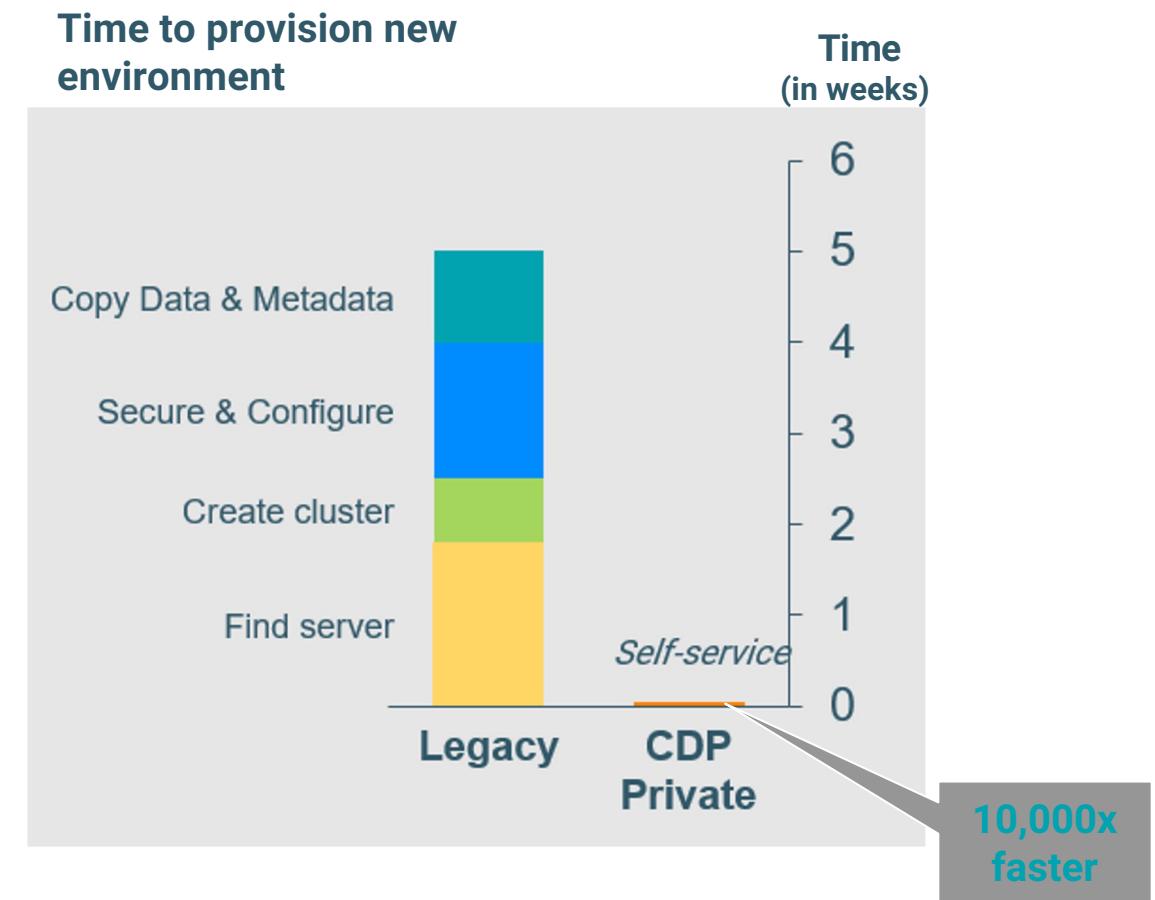
RAPID

Push-button Provisioning

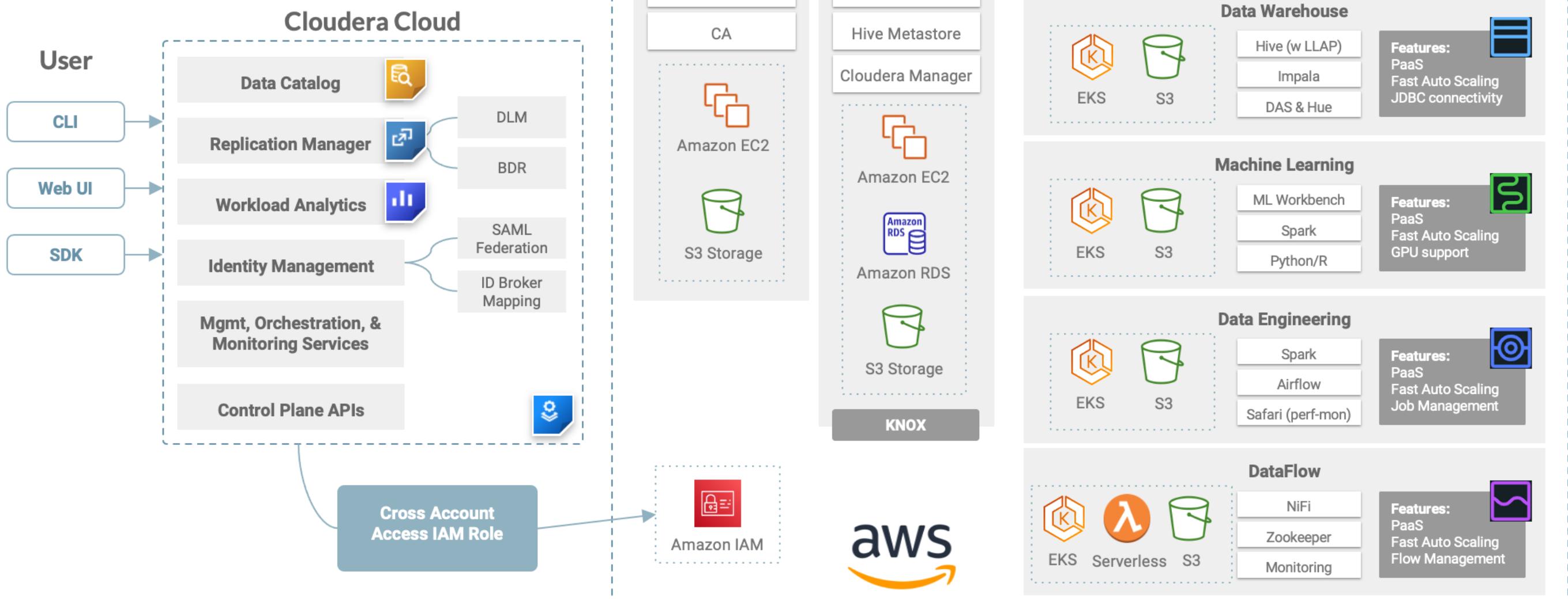
- Near instantaneous provisioning - reduce weeks of work down to minutes

Redesigned User Interfaces

- Workflows optimized for self-service analytical experiences



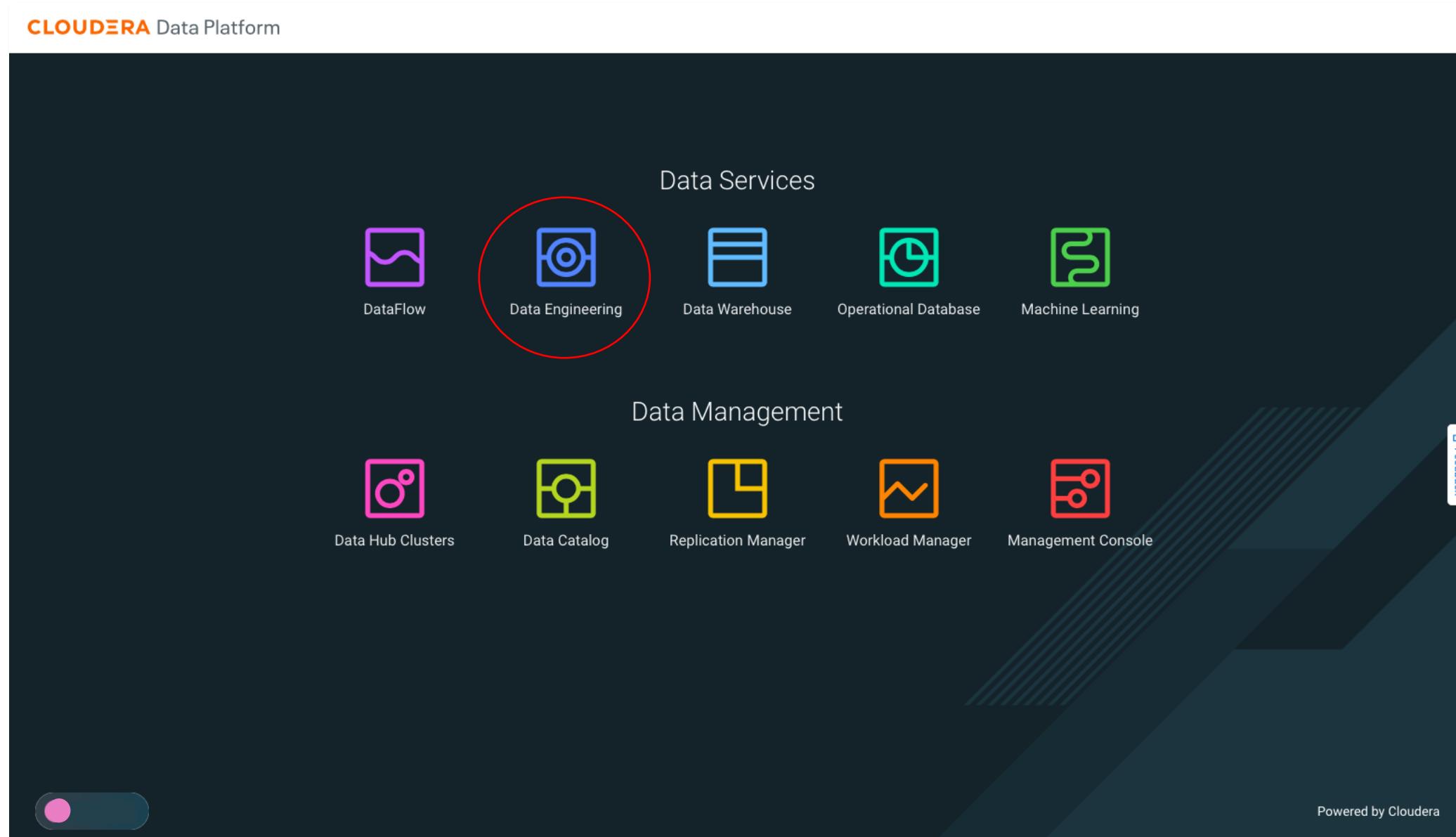
CDP on AWS High Level Architecture



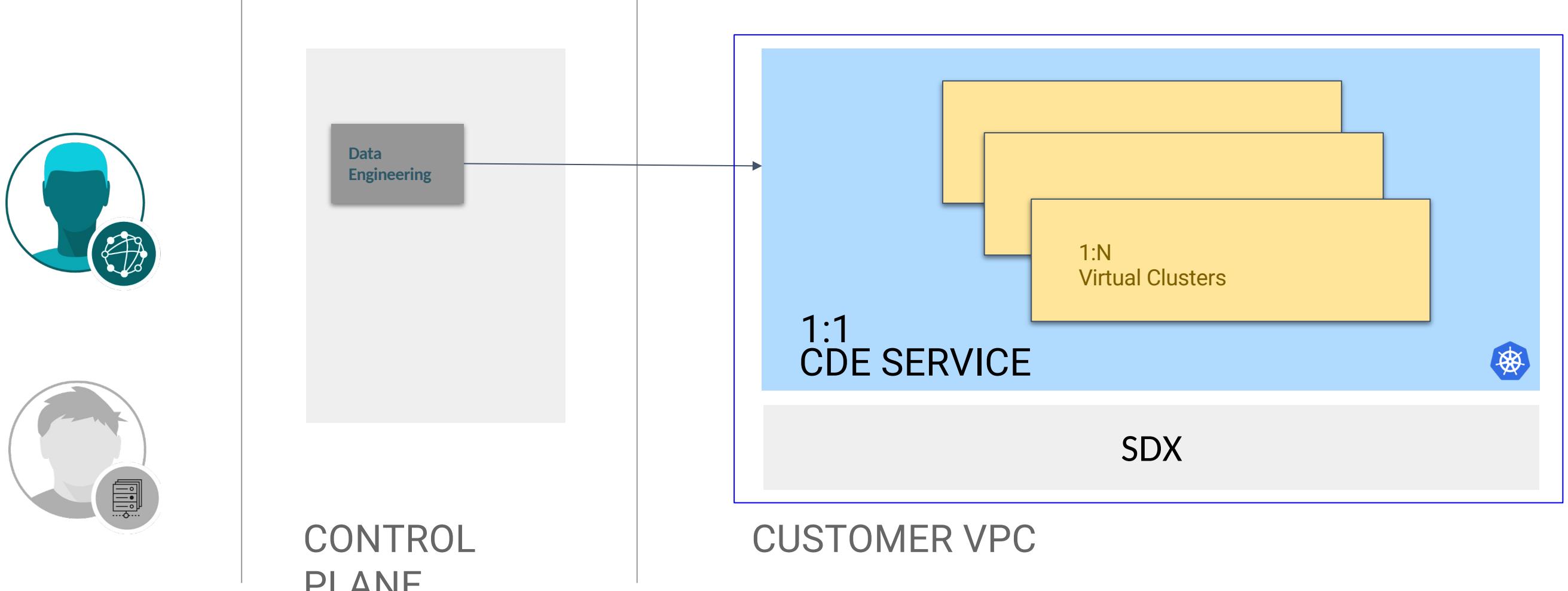


Data Engineering Service

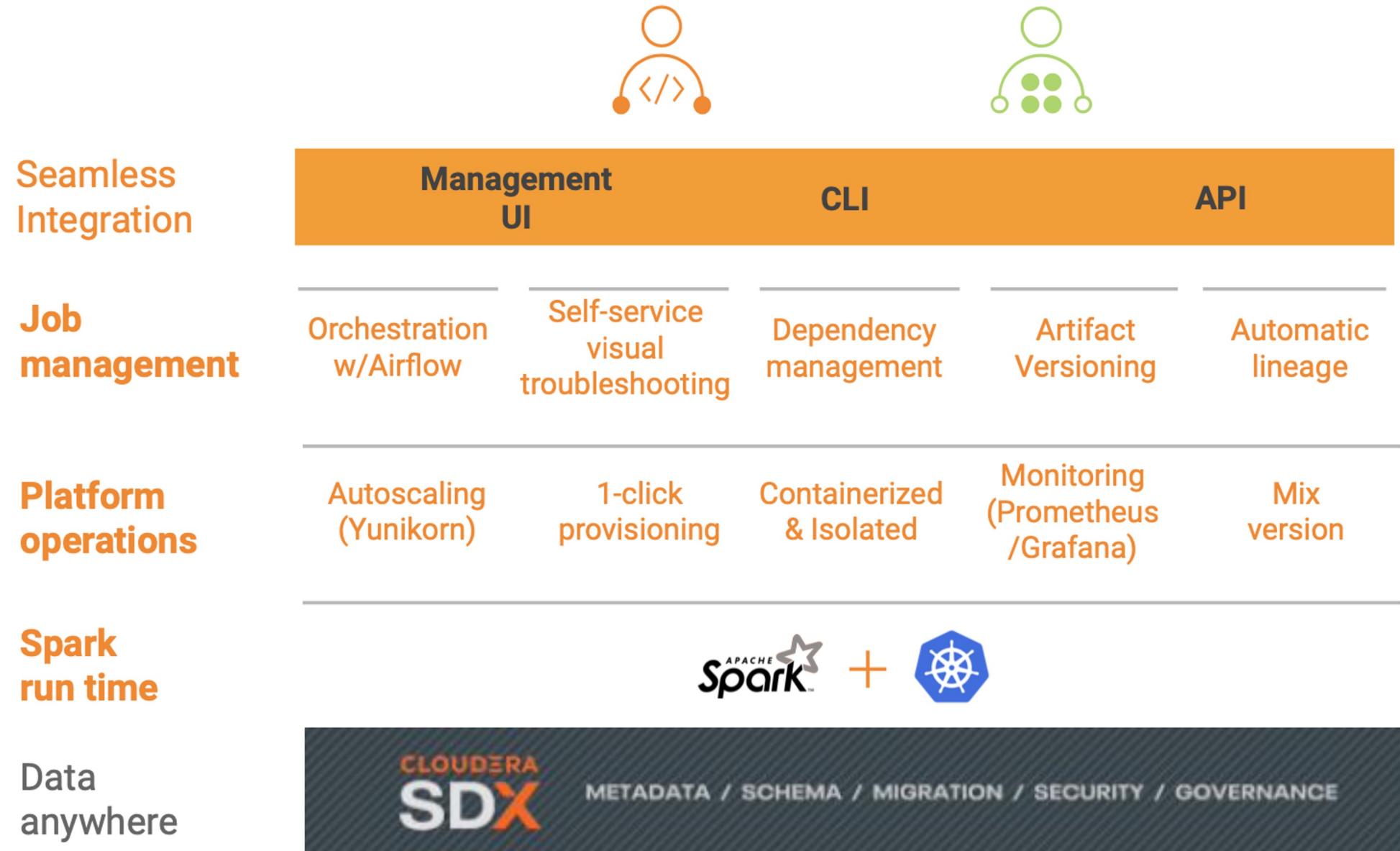
Data Engineering Service



CDE Service Components and Relationships



Cloudera Data Engineering



Autoscaling

The screenshot shows the Cloudera Data Engineering interface. On the left, there's a sidebar with a 'CLOUDERA Data Engineering' logo and an 'Overview' button. The main area has two tabs: 'Environments [1]' and 'Virtual Clusters / dex-mow-int-env [2]'. The 'Environments' tab shows one environment named 'dex-mow-int-env' running on AWS, with 3 nodes, 24 CPU units, and 92 GB of memory. A red notification bell icon with '1' is visible. The 'Virtual Clusters' tab shows two workloads: 'HeavyETL-workload' and 'SalesOps-workload', both running on the 'dex-mow-int-env' environment. Each workload has 9 pods, 4.5 CPU units, 9 GB of memory, and 0 jobs. A blue arrow points from the text 'Isolated virtual clusters' to the 'Virtual Clusters' tab. Another blue arrow points from the text 'Autoscaling' to a step-up line graph at the bottom right of the cluster details.

Isolated virtual clusters

Autoscaling

Cloud Engineering Data Engineering

Overview

Cloud Environments [1]

dex-mow-int-env

aws

Enabled

NODES 3

CPU 24

MEMORY 92 GB

Enable new CDE

Virtual Clusters / dex-mow-int-env [2]

HeavyETL-workload

dex-mow-int-env

Running

PODS 9

CPU 4.5

MEMORY 9 GB

JOB 0

SalesOps-workload

dex-mow-int-env

Running

PODS 9

CPU 4.5

MEMORY 9 GB

JOB 0

Active

1

Autoscaling

Provision Isolated Workloads with Quotas

CLOUDERA Data Engineering

Overview / Create a Cluster

3

Cluster Name: SalesOps-VirtualCluster

CDE Service: dex-dev-us-west-2

Auto-Scale Range

CPU: 35 (Max 50)

Memory (GB): 100 (Max 200)

Summary

- Cluster Name: SalesOps-VirtualCluster
- CDE Service: cluster-rk6pbs89
- Auto-Scale Range:
 - CPU: 35 Max
 - Memory: 100 GB Max
- Airflow enabled

SalesOps-VirtualCluster

dex-dev-us-west-2

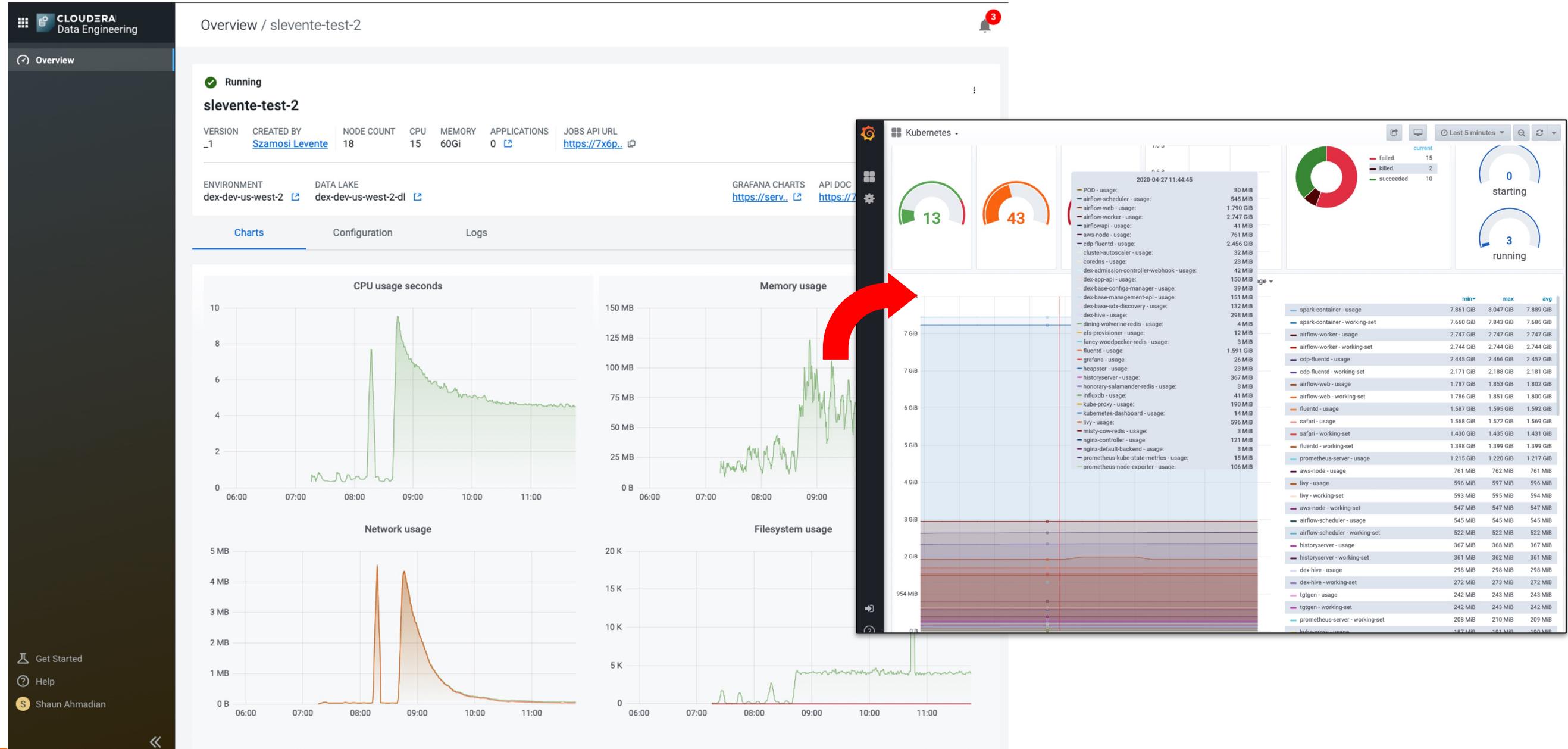
35

100

Summary

- Cluster Name: SalesOps-VirtualCluster
- CDE Service: cluster-rk6pbs89
- Auto-Scale Range:
 - CPU: 35 Max
 - Memory: 100 GB Max
- Airflow enabled

Monitor Capacity & Resource Usage Thru Grafana



Easy Job Deployment

CLOUDERA Data Engineering

Job Runs

Jobs

Schedules

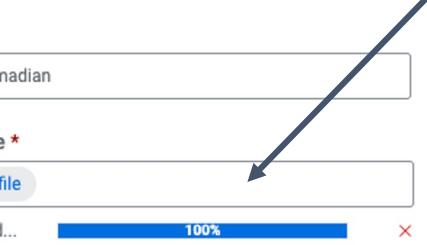
Jobs / Create Job

Quick upload of JAR / Python code

Job Details

Name * ETL-job-2

User * csso_sahmadian

Upload File * Choose file insurance-cd... 100% 

Main Class * org.cloudera.cde.app.Application

Arguments (Optional) Argument 

Configurations (Optional) spark.yarn.access.hadoopFileSystems 

Advanced Configurations >> Upload additional files, customize no. of executors, driver and executor cores and memory

Advanced Config

Advanced Configurations << Upload additional files, customize no. of executors, driver and executor cores and memory

Upload Jars (dependencies)  Select jar file(s)

Upload dependencies  Select file(s) to upload

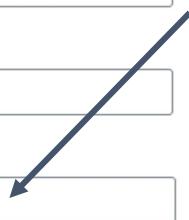
Executors  1 20 50  20

Driver Cores  1 4 16  4

Executor Cores  1 8 16  8

Driver Memory (GB)  1 5 32  5

Executor Memory (GB)  1 32  32



Schedule *

Run Now Schedule

Every year on every day of every month at 6,8,18 : 15,30
At 15 and 30 minutes past the hour, at 06:00 AM, 08:00 AM, and 06:00 PM

Use a Cron Expression
* * * * *

Start Date Tuesday, May 5, 2020 at 4:32:21 PM

UTC Time: Tuesday, May 5, 2020 at 11:32 PM UTC

End Date Wednesday, May 6, 2020 at 4:32:21 PM

Scheduling Configurations

Enable Catchup
Kick off Job Runs for intervals that has not been run along with the current interval

Depends on Previous
If the job runs are dependant, then the catchup will occur serially

Flexible Scheduler

Flexible Orchestration Backed By Airflow

Embedded Apache Airflow

The screenshot shows the Cloudera Data Engineering interface with the 'Schedules' tab selected. The main area displays a list of DAGs:

	i	DAG	Schedule	Owner	Recent
1	On	analytics-pyspark-job1	0,15,30,45 * * * *	Airflow	1
2	On	demo-etl-job	0,15,30,45 * * * *	Airflow	1
3	Off	demo-job-1	0,10,20,30,40,50 * * * *	Airflow	1
4	On	ingestion-ETL-job-1	0,15,30,45 * * * *	Airflow	1

An annotation labeled "Auto-generated Pipelines" points to the "ingestion-ETL-job-1" row. A modal window titled "DAG: ingestion-ETL-job-1" is open, showing the Python code for the DAG:

```
from dateutil import parser
from datetime import timezone
from airflow import DAG
from cde_job_run_operator.operator import CDEJobRunOperator

default_args = {
    'owner': 'Airflow',
    'depends_on_past': False,
    'wait_for_downstream': False,
    'start_date': parser.isoparse('2020-08-04T15:12:47.262Z').replace(tzinfo=timezone.utc),
    'end_date': parser.isoparse('2020-08-05T15:12:47.262Z').replace(tzinfo=timezone.utc),
    'connection_id': 'cde_runtime_api',
    'job_name': 'ingestion-ETL-job-1',
    'user': 'sahmadian',
}

dag = DAG(
    'ingestion-ETL-job-1',
```

An annotation labeled "Pipelines as Python Code" points to the code editor area.

Cloudera Data Engineering UI Elements:

- Job Runs
- Jobs
- Resources
- Schedules
- Schedule (button)
- DAGs (selected)
- Data Profiling
- Browse
- Admin
- HeavyETL-Workload (button)
- Search bar
- Graph View, Tree View, Task Duration, Task Tries, Landing Times, Gantt, Details, Code buttons
- Refresh, Delete buttons
- Toggle wrap button
- Page navigation buttons («, <, 1, >, »)
- Hide Paused DAGs button
- User profile: sahmadian
- Cloudera Educational Services logo

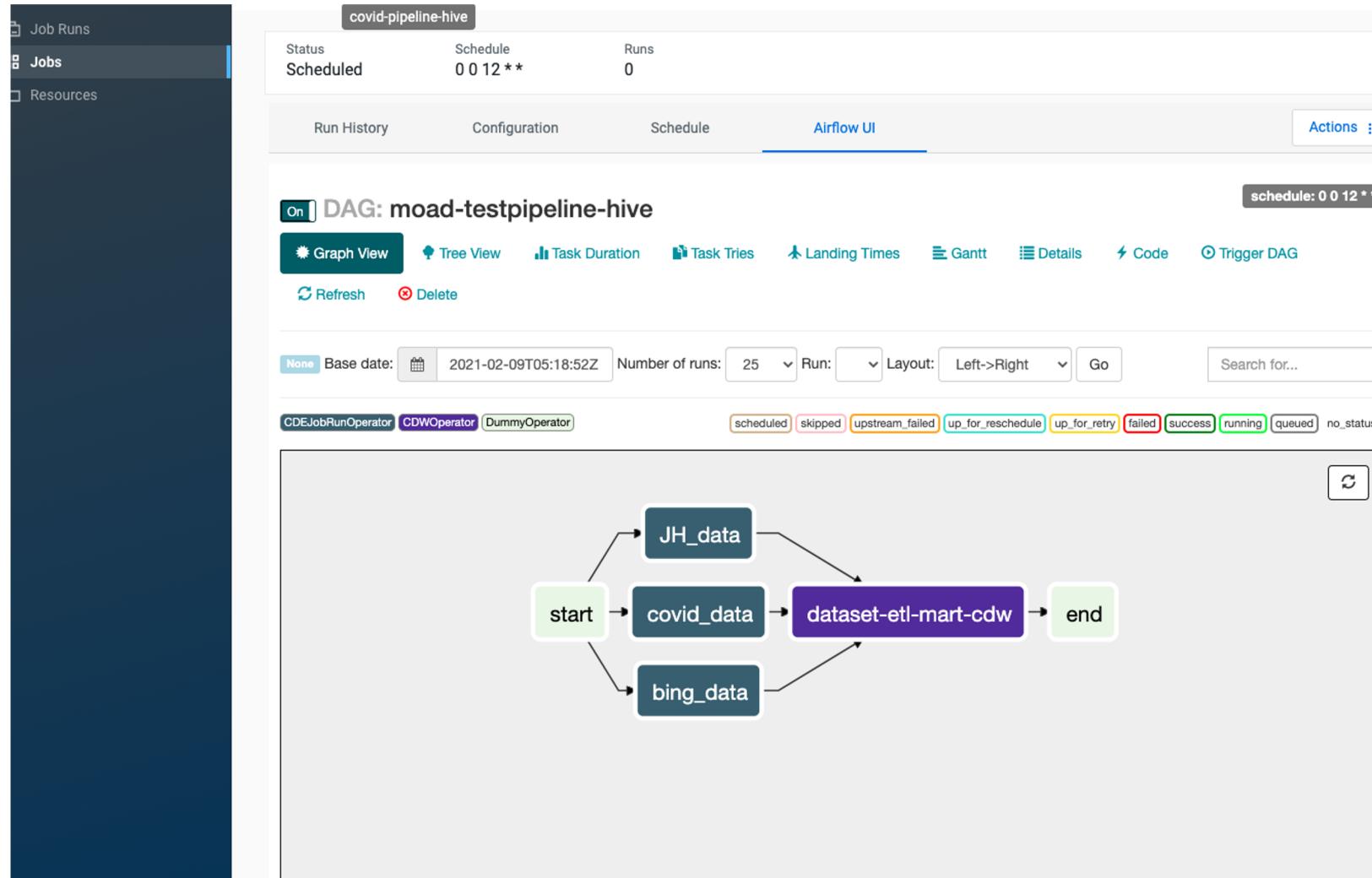
Deploy Airflow Jobs

The screenshot shows the Cloudera Data Engineering interface. On the left, a sidebar menu includes 'Job Runs', 'Jobs' (which is selected), and 'Resources'. The main area is titled 'Jobs / Create Job' and contains a form for 'Job Details'. The 'Job Type' field has two options: 'Spark' (unchecked) and 'Airflow' (checked). A red arrow points to the 'Airflow' option. Below it, there's a 'Name *' field with a placeholder 'Job Name', a 'DAG File *' section with 'File' (checked) and 'Resource' (unchecked) options, and an 'Upload File *' section with a 'Choose file' button set to 'Python file'. At the bottom are 'Cancel' and 'Create and Run' buttons. To the right, a list of existing jobs is displayed in a table:

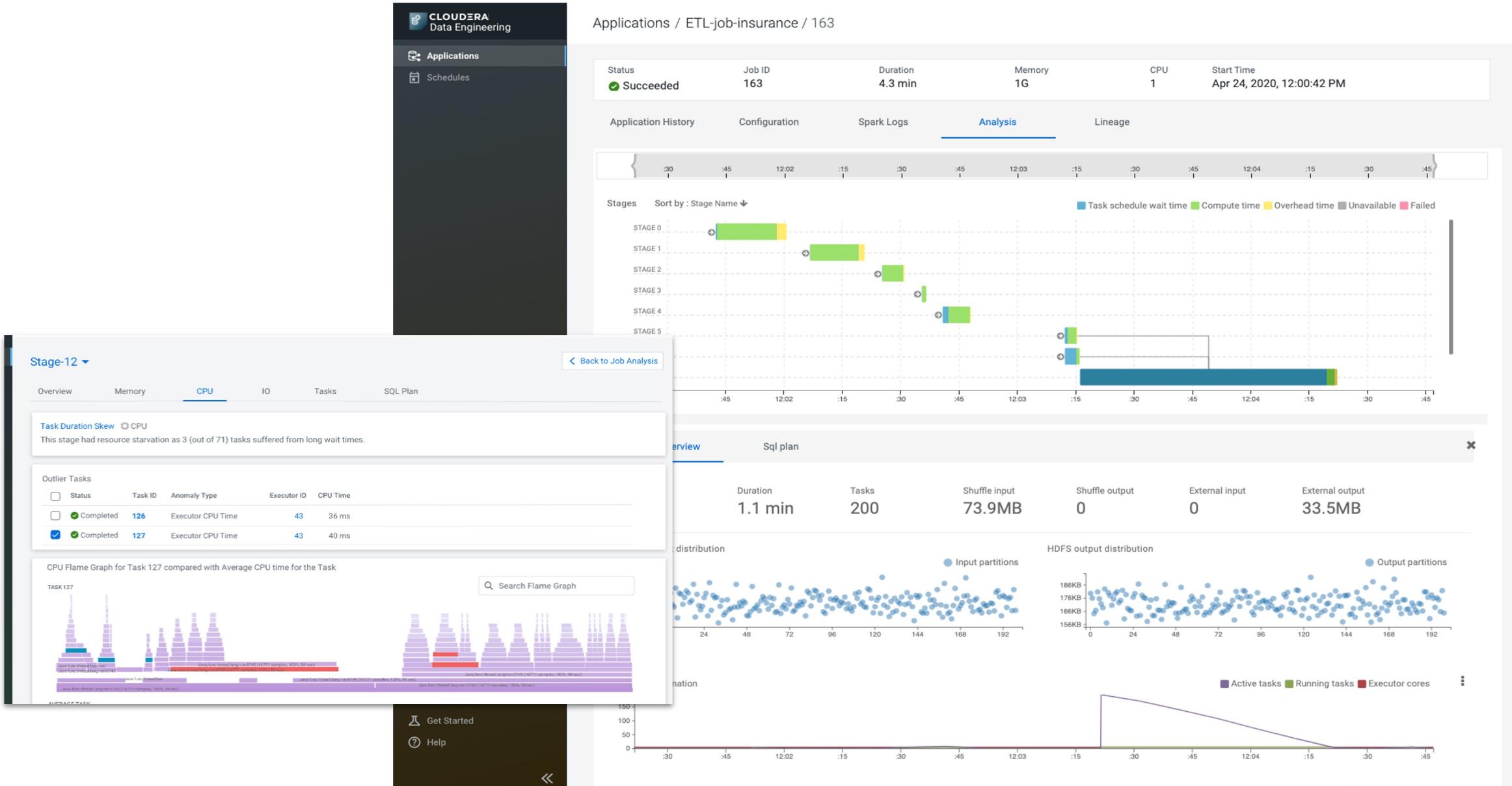
Status	Job	Type	Schedule
⌚	vivek-airflow-job-2	Airflow	*/20 * * * *
⌚	ps-pi	Spark	Ad-Hoc
⌚	tpcds	Spark	Ad-Hoc
⌚	ps-skew	Spark	Ad-Hoc

A red arrow points from the 'Type' column header in the table towards the 'Airflow' entry in the first row.

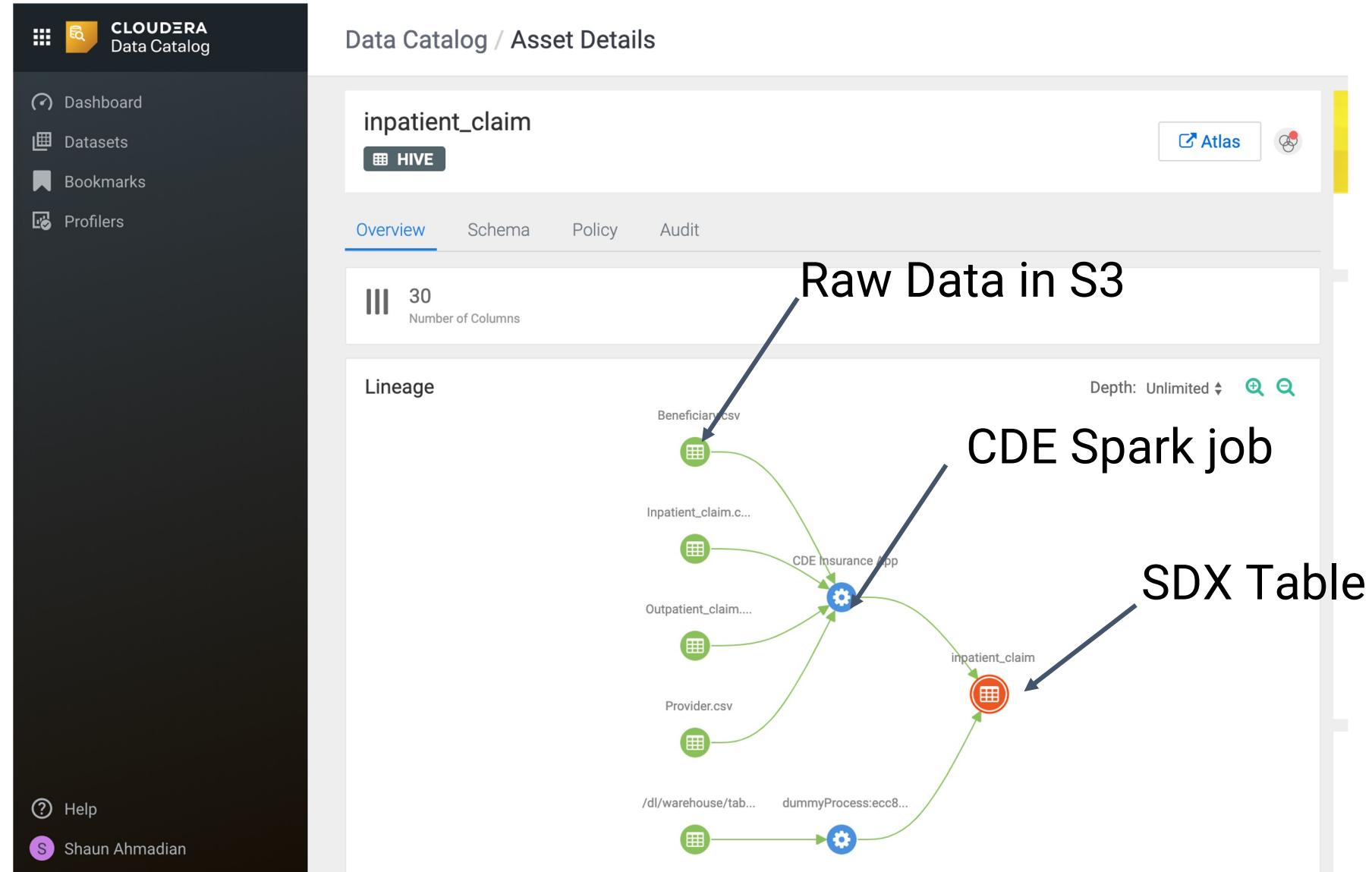
Build Spark And Hive Pipelines With Dependencies



Visual Troubleshooting & Performance Tuning



Automatic Capture Of Data Lineage in SDX



Rich API

The screenshot shows a Swagger UI interface for a RESTful API. The top navigation bar includes the Swagger logo, the file name "doc.json", and a "Explore" button. A version indicator "1.0" and a base URL "[Base URL: s64m5zcl.cde-2gvjdmhl.dex-mow.svbr-nqvp.int.clqr.work/dex/api/v1] doc.json" are also present.

The API is organized into several sections:

- applications**:
 - GET /applications** List all applications
 - POST /applications** Create an application
 - GET /applications/{name}** Describe application
 - DELETE /applications/{name}** Delete application
 - PATCH /applications/{name}** Update application
- applications schedule**:
 - POST /applications/{name}/schedule/clear** Clear application schedule
 - POST /applications/{name}/schedule/mark-success** Mark application schedule as successful
 - POST /applications/{name}/schedule/pause** Pause application schedule
 - POST /applications/{name}/schedule/unpause** Unpause application schedule
- info**:
 - GET /info** Information about the instance
- jobs**:
 - GET /jobs** List jobs
 - POST /jobs** Run a job
 - GET /jobs/{id}** Describe job

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.

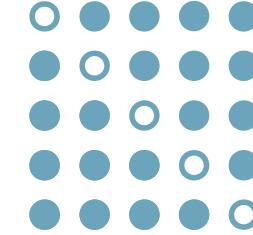
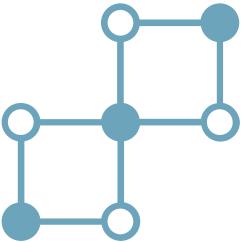
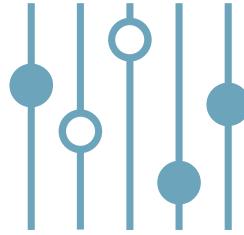


Machine Learning Service

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 organization's 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**

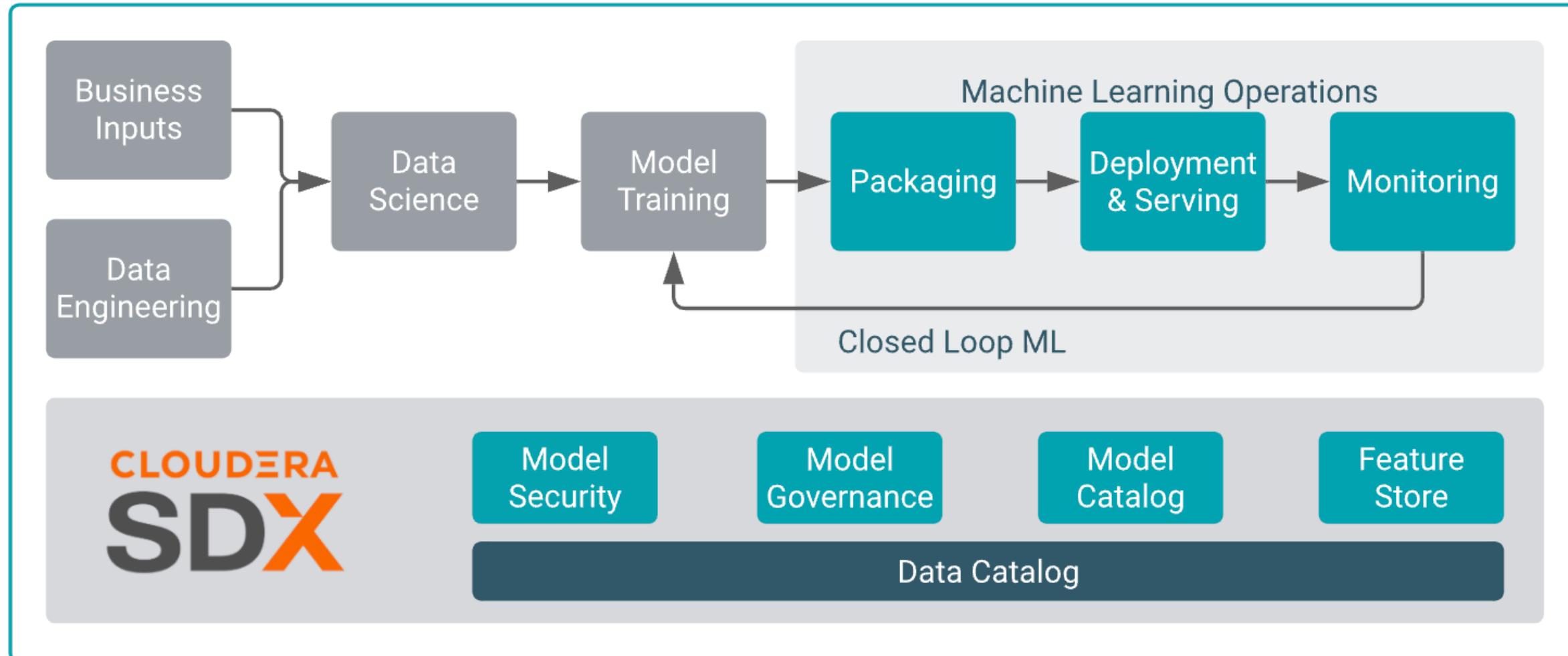
DEPLOY APIs

Deploy and monitor models as APIs to **serve predictions**

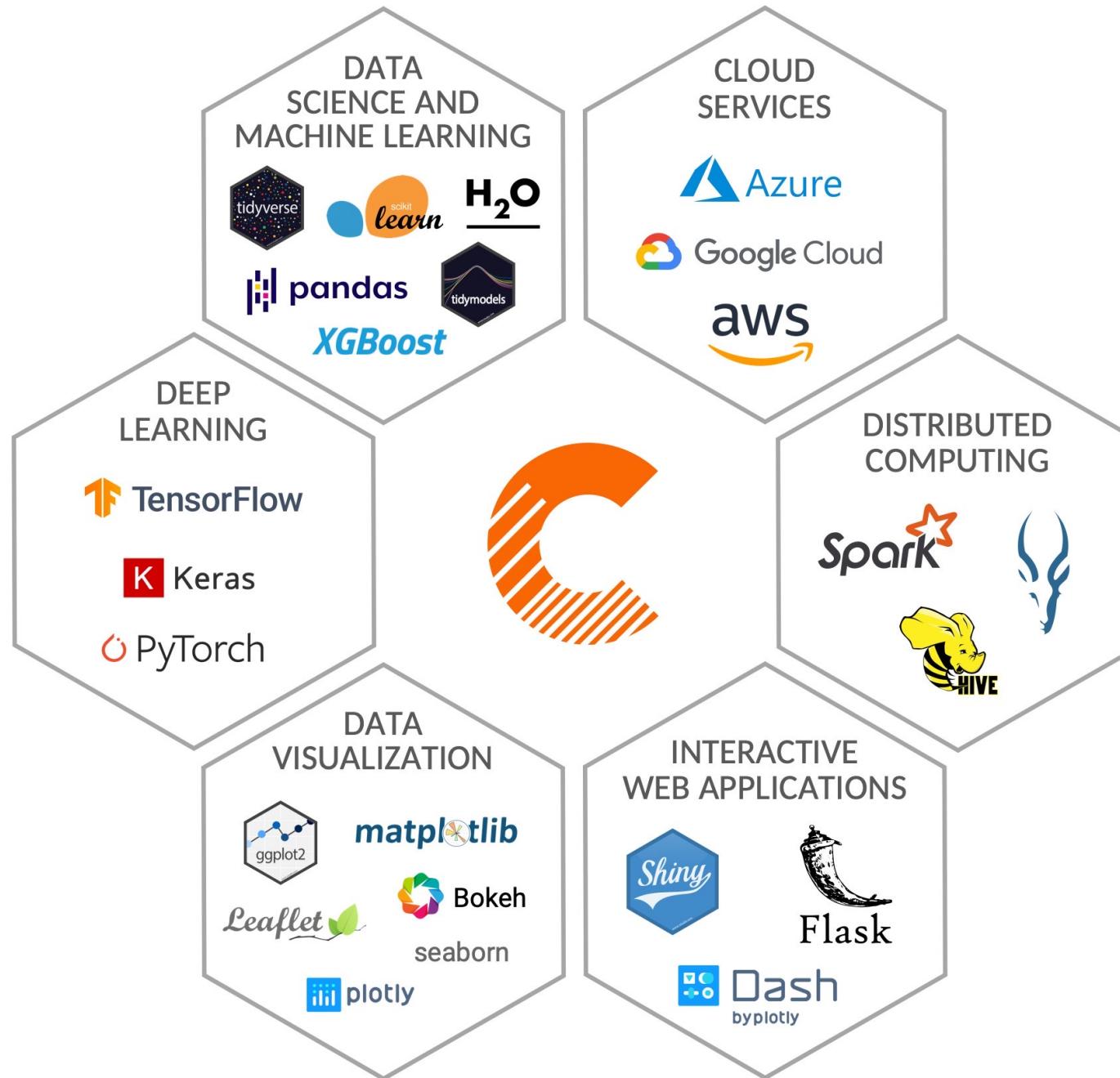
MANAGE SHARED RESOURCES

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

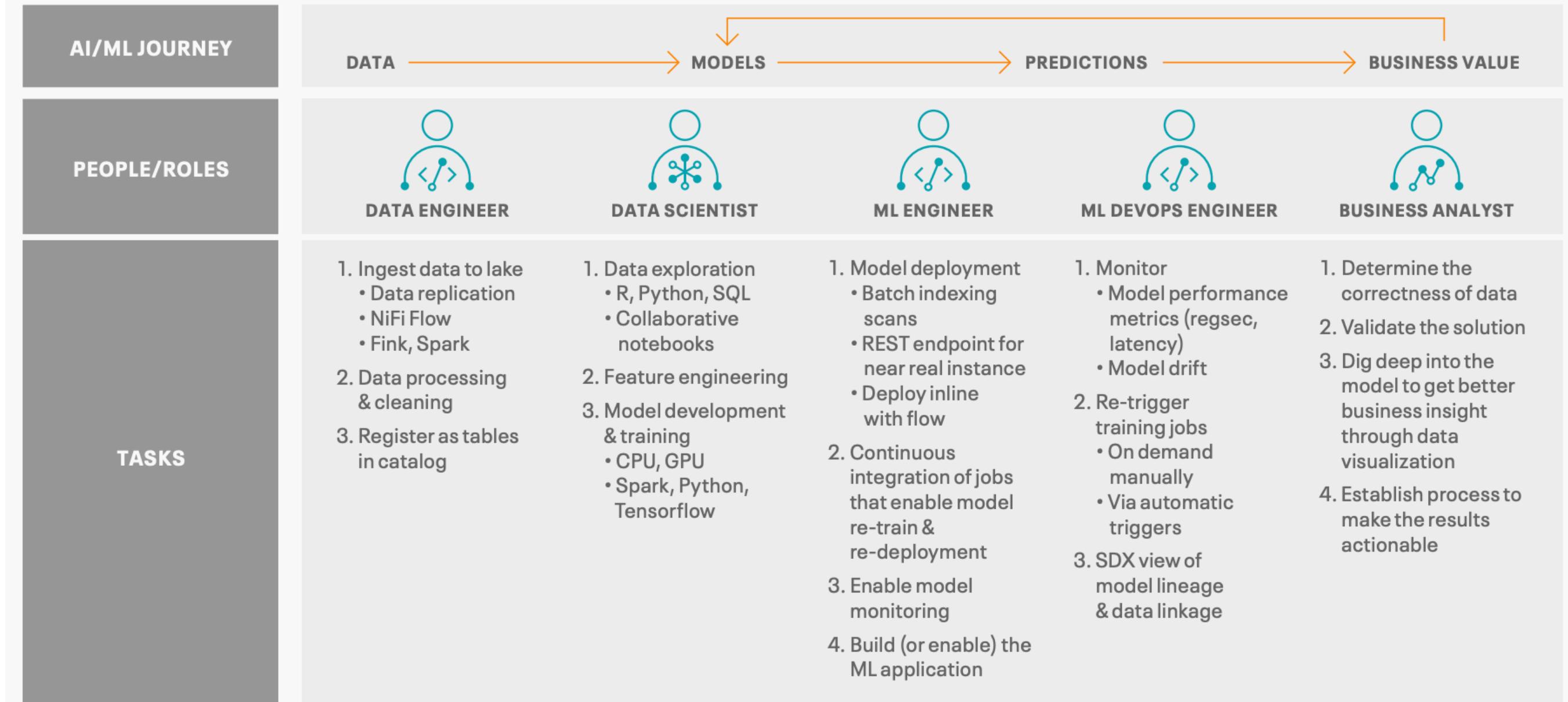
ML Workflow and Operations



What CML Can Help You Do



AI/ML Journey



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)
<i>Close to End-of-Life</i>	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 within the organization

The screenshot shows the Cloudera Management Console interface. On the left is a dark sidebar with navigation links: Dashboard, Environments, Data Lakes, User Management, Data Hub Clusters, Data Warehouses, **ML Workspaces** (which is currently selected), and Classic Clusters. The main area is titled "Machine Learning Workspaces". It features a search bar, a dropdown for "Environment" set to "All", and a large blue button labeled "Provision Workspace". Below these are two tables. The first table lists workspaces with columns for Status (green checkmark or red exclamation mark), Workspace Name, Launch Workspace button, Environment, Creation Date, and Actions (three dots). The second table lists environments with columns for Environment Name, Launch Environment button, and Actions (three dots). At the bottom right, there is a pagination message "Displaying 1 - 10 of 10" and a "25 / page" dropdown.

Status	Workspace Name	Launch Workspace	Environment	Creation Date	Actions
✓	irbcm1	Launch Workspace	ml-demo-env-6jan	01/13/2020 12:38 PM PST	⋮
✓	regtest	Launch Workspace	mlx-dev-prod-env	01/13/2020 10:24 AM PST	⋮
!	ram-CB-4990	Launch Workspace	sense-prod-env	01/10/2020 10:40 AM PST	⋮
✓	ml-demo-wksp-7jan	Launch Workspace	ml-demo-env-6jan	01/07/2020 6:29 AM PST	⋮
✓	eng-cml-cluster	Launch Workspace	mlx-dev-prod-env	01/06/2020 2:21 PM PST	⋮
✓	OASC-test	Launch Workspace	ml-demo-env-3dec	12/10/2019 12:03 PM PST	⋮
!	nallen-demo-workspace	Launch Workspace	testenv	12/06/2019 10:24 AM PST	⋮
✓	bigdatum-MLW	Launch Workspace	bigdatum-environment	12/03/2019 10:43 PM PST	⋮
!	ml-demo-wksp-3dec	Launch Workspace	ml-demo-env-3dec	12/03/2019 11:04 AM PST	⋮
✓	dnarain-ml-aps1	Launch Workspace	lord-of-the-rules-aps1	08/27/2019 11:17 AM PDT	⋮

Environment	Launch Environment	Actions
mlx-dev-prod-env	Launch Environment	⋮
sense-prod-env	Launch Environment	⋮
ml-demo-env-6jan	Launch Environment	⋮
mlx-dev-prod-env	Launch Environment	⋮
ml-demo-env-3dec	Launch Environment	⋮
testenv	Launch Environment	⋮
bigdatum-environment	Launch Environment	⋮
ml-demo-env-3dec	Launch Environment	⋮
lord-of-the-rules-aps1	Launch Environment	⋮

Elastic Kubernetes Cluster

- Each ML Workspace is an elastic Kubernetes cluster

Advanced Options

CPU Settings

* Instance Type

Select an Environment to see options.

Autoscale Range

0 5 10

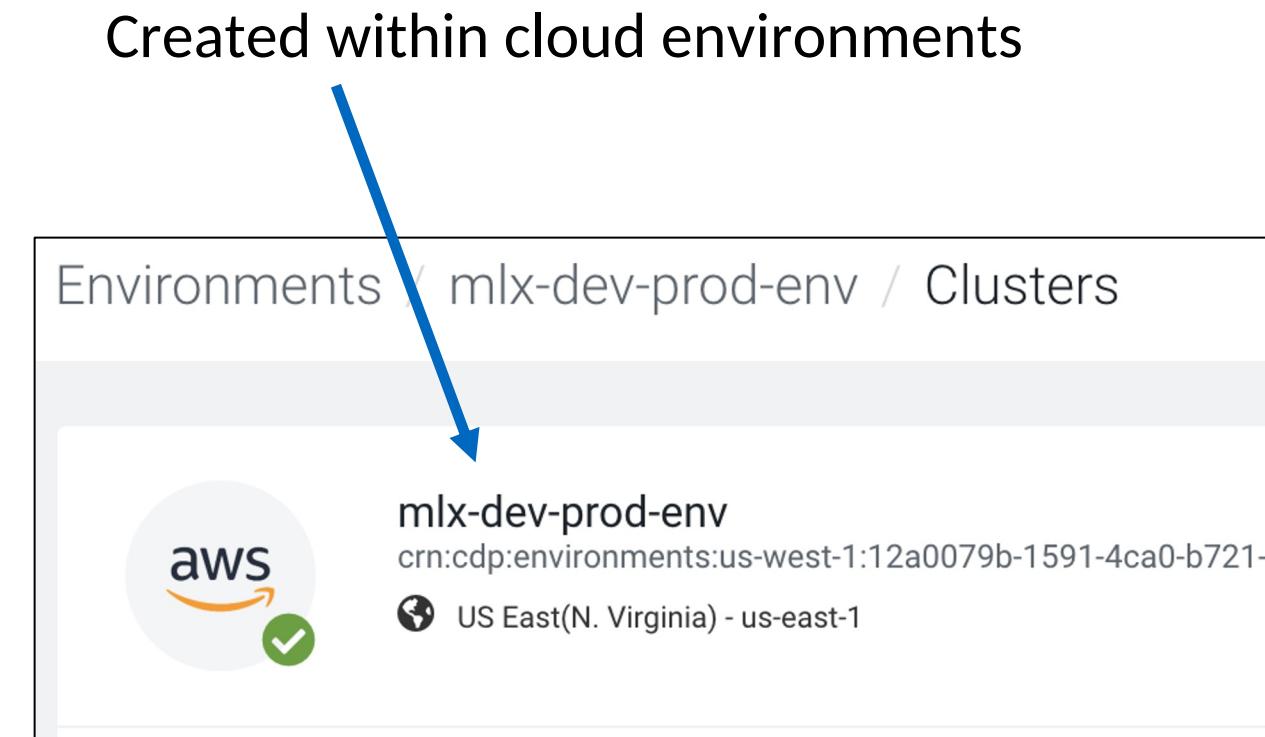
GPU Settings

Instance Type

Select an Environment to see options.

Autoscale Range

0 10



Accessing a Workspace

The screenshot shows the CDP web interface with the following components:

- Header:** "Your Enterprise Data Cloud"
- Top navigation icons:** Data Hub Clusters (pink), DataFlow (purple), Data Engineering (blue), Data Warehouse (light blue), Operational Database (teal), and Machine Learning (green). The Machine Learning icon is highlighted with a red box.
- Sub-navigation:** "Machine Learning Workspaces" with a green sidebar icon.
- Search and filters:** "Search Workspaces" input, "Environment" dropdown set to "All", and "Provision Workspace" button.
- Table:** Displays a single workspace row with the following details:

Status	Workspace	Environment	Region	Creation Date	Cloud Provider	Actions
Ready	edu-cml	bshimmel-dev-us-east-2	us-east-2	06/20/2022 12:52 PM EDT	AWS	⋮
- Pagination:** "Displaying 1 - 1 of 1" and "25 / page" dropdown.

- 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

The screenshot shows the Cloudera Machine Learning interface with the 'Projects' page selected. The left sidebar has a dark theme with white icons and text, listing options like 'Projects', 'Sessions', 'Experiments', 'Models', 'Jobs', 'Applications', 'User Settings', 'AMPS', 'Runtime Catalog', 'Site Administration', and 'Learning Hub'. The main content area has a light background. At the top, there's a search bar labeled 'Project quick find' and a dropdown for 'dev_20_22829'. Below that is a header with 'Active Workloads' and counts for Sessions (0), Experiments (0), Models (1), Jobs (0), and Applications (7). To the right are two resource monitoring sections: 'User Resources' and 'Workspace Resources'. The 'User Resources' section shows CPU (3.5 vCPU available), Memory (7.0 GB available), and GPU (0.0 GPU available). The 'Workspace Resources' section shows similar metrics. Below these are filters for 'Search Projects', 'Scope' (set to 'My Projects'), and 'Creator' (set to 'All'). A 'Sort By' dropdown is set to 'Last Updated'. On the far right, there are buttons for 'New Project' and other navigation. The main area displays seven project cards in a grid:

- CDP on CML**: Created by dev_20_22829, last worked on 17 hours ago.
- Churn Analysis Refact...**: Created by dev_20_22829, last worked on a day ago.
- Continuous Model Mo...**: Created by dev_20_22829, last worked on 3 days ago.
- Evidently Test**: Created by dev_20_22829, last worked on 14 days ago.
- Deep Learning with GP...**: Created by dev_20_22829, last worked on 15 days ago.
- Duocar 20**: Created by dev_20_22829, last worked on 15 days ago.
- Student 20**: Created by dev_20_22829, last worked on 15 days ago.

Pagination controls at the bottom right indicate page 1 of 25.

A Project Contains...

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

```
Seaborn Analysis
analysis.py
File Edit View Navigate Run
analysis.py

1 # Setup
2 # -----
3
4 import pandas as pd
5 import seaborn as sns
6
7 # Basic Data Manipulation
8 # -----
9 #
10 # Use the seaborn tips dataset to generate a best fitting linear regression line
11
12 tips = sns.load_dataset("tips")
13 sns.set(font="DejaVu Sans")
14 sns.jointplot("total_bill", "tip", tips, kind='reg').fig.suptitle("Tips Regression", y=1.01)
15
16 # Examine the difference between smokers and non smokers
17 sns.lmplot("total_bill", "tip", tips, col="smoker").fig.suptitle("Tips Regression - categorized by smoker", y=1.05)
18
19 # Explore the dataframe
20 tips.head()
21
22 # Using IPython's Rich Display System
23 # -----
24 #
25 # IPython has a [rich display system](bit.ly/HHP0ac) for
26 # interactive widgets.
27
28 from IPython.display import IFrame
29 from IPython.core.display import display
30
31 # Define a google maps function.
32 def gmaps(query):
33     url = "https://maps.google.com/maps?q={0}&output=embed".format(query)
34     display(IFrame(url, '700px', '450px'))
35
36 gmaps("Golden Gate Bridge")
37
38 # Worker Engines
39 # -----
40 #
41 # You can launch worker engines to distribute your work across a cluster.
42 # Uncomment the following to launch two workers with 2 cpu cores and 0.5GB
43 # memory each.
44
45 # import cdsw
46 # workers = cdsw.launch_workers(n=2, cpu=0.2, memory=0.5, code="print('Hello from a CDSW Worker')")
```

Line 1, Column 1

★ 47 Lines Python Spaces 2

Managing Project Files

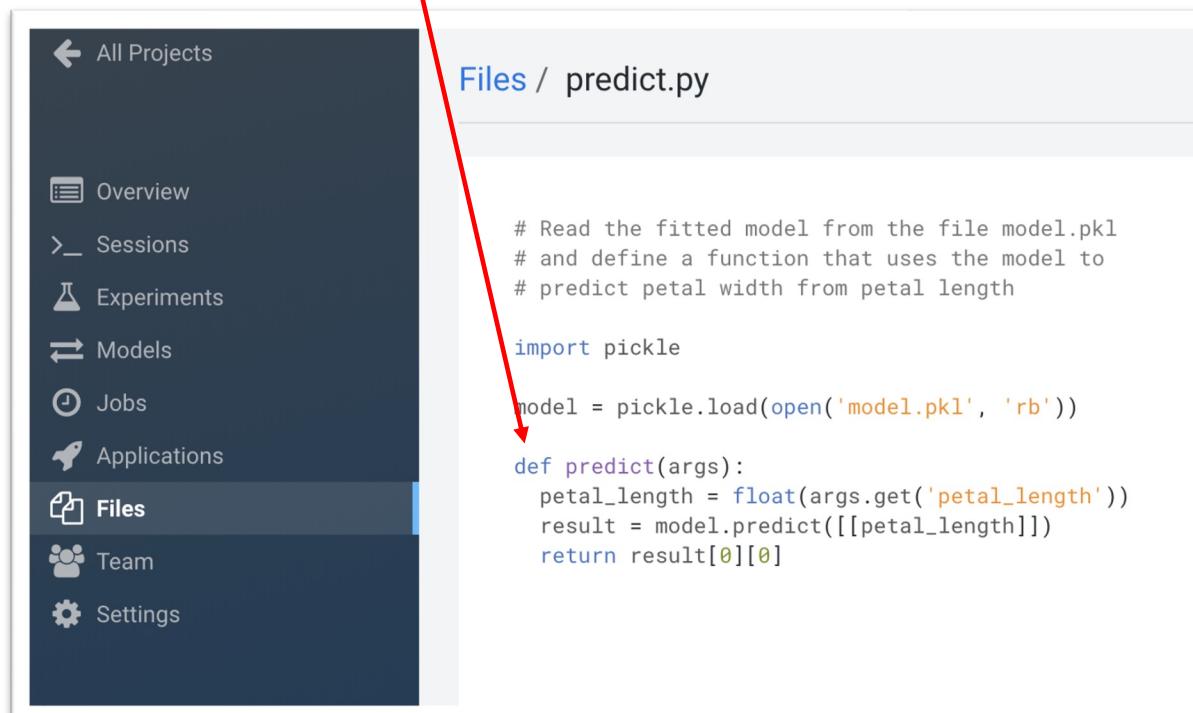
The screenshot shows a dark-themed user interface for managing project files. On the left, a sidebar lists various project management sections: All Projects, Overview, Sessions, Data, Experiments, Models, Jobs, Applications, Files (which is currently selected and highlighted in green), Collaborators, Project Settings, and Help. The main area is titled 'Files' and displays a list of files and directories:

- seaborn-data
- analysis.ipynb
- analysis.py
- cdsw-build.sh
- config.yml
- entry.py
- fit.py
- lineage.yaml
- pi.py
- predict.py
- predict_with_metrics.py
- README.md
- requirements.txt
- use_model_metrics.py

- **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...



All Projects

- Overview
- Sessions
- Experiments
- Models
- Jobs
- Applications
- Files**
- Team
- Settings

Files / predict.py

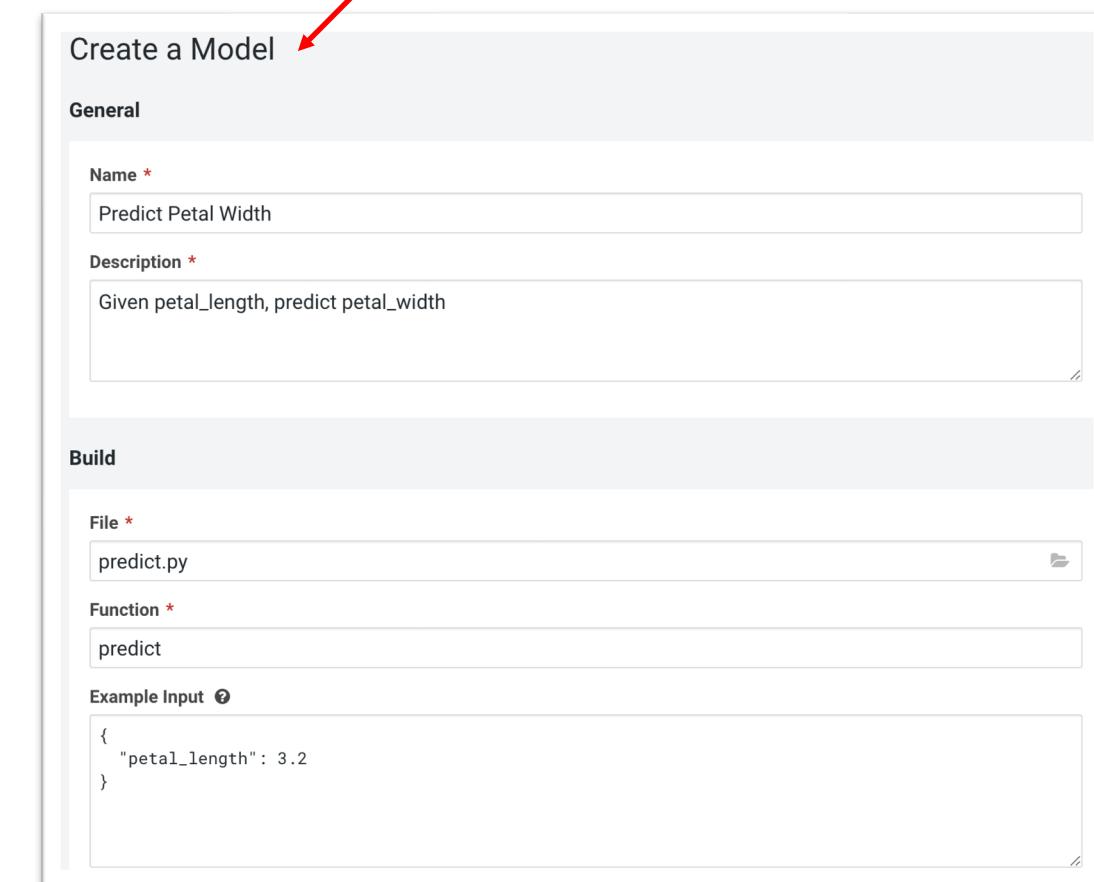
```
# Read the fitted model from the file model.pkl
# and define a function that uses the model to
# predict petal width from petal length

import pickle

model = pickle.load(open('model.pkl', 'rb'))

def predict(args):
    petal_length = float(args.get('petal_length'))
    result = model.predict([[petal_length]])
    return result[0][0]
```

...and then use CML to serve it up



Create a Model

General

Name *

Description *

Build

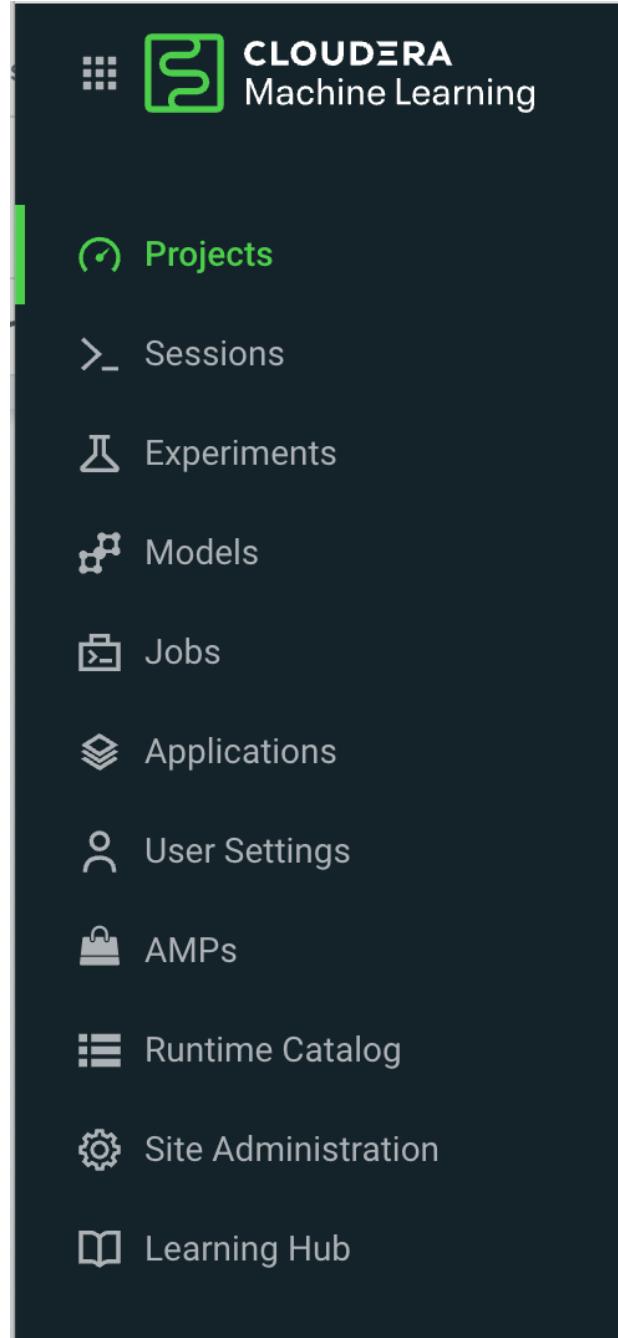
File *

Function *

Example Input ?

```
{ "petal_length": 3.2 }
```

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

Environment

Provides access to CML within a given CDP environment

Workspace

Provides access to a single workspace

- **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 specific 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

The screenshot shows the Cloudera Machine Learning workspace interface for a workspace named 'cml-edu'. The left sidebar has 'Workspaces' selected. The main area displays workspace details: Name (cml-edu), Environment / Region (ps-sandbox-aws / us-east-2), Filesystem ID (fs-00129cad4de821fa2), NFS Protocol version (unknown), CRN (crn:cdp:ml:us-west-1:558bc1d2-8867-4357-8524-311d51259233:workspace...), Workspace ID (ml-8250e328-99f), Cluster Name (liftie-jyr14lyr), Monitoring (Enabled), and TLS (Enabled). A red arrow points to the 'Actions' dropdown menu on the right, which includes options: Manage Access (highlighted), Manage Remote Access, Download Kubeconfig, Open Grafana, Upgrade Workspace, Backup Workspace, and Remove Workspace.

Machine Learning Workspaces / cml-edu

Status ✓ Ready

Details Events & Logs

Actions ▾

Name cml-edu

Environment / Region ps-sandbox-aws / us-east-2

Filesystem ID fs-00129cad4de821fa2

NFS Protocol version unknown

CRN crn:cdp:ml:us-west-1:558bc1d2-8867-4357-8524-311d51259233:workspace...

Workspace ID ml-8250e328-99f

Cluster Name liftie-jyr14lyr

Monitoring Enabled

TLS Enabled

Manage Access

Manage Remote Access

Download Kubeconfig

Open Grafana

Upgrade Workspace

Backup Workspace

Remove Workspace

Updating a User's Roles

- Select the user to add or remove a role

Update Resource Roles for Sheryl Sarokas

X

Resource Roles		
<input type="button" value="-"/> Role	Description	
<input type="checkbox"/> MLWorkspaceAdmin ⓘ	Grants permission to manage all machine learning workloads and settings inside a specific workspace.	
<input type="checkbox"/> MLWorkspaceBusinessUser ⓘ	Grants permission to view shared machine learning applications inside a specific workspace.	
<input checked="" type="checkbox"/> MLWorkspaceUser ⓘ	Grants permission to run machine learning workloads inside a specific workspace.	
<input type="checkbox"/> Owner ⓘ	Grants all permissions on the resource.	

[Cancel](#) [Update Roles](#)

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

The screenshot shows the Cloudera Machine Learning interface. At the top left is the 'CLOUDERA Machine Learning' logo. To its right is a navigation bar with the following items: 'All Projects' (with a back arrow), 'Overview' (highlighted in green), 'Sessions', 'Data', 'Experiments', 'Models', 'Jobs', 'Applications', 'Files', 'Collaborators', and 'Project Settings'. On the right side of the interface, the path 'bshimeldevuseast2_3 / ML_Project' is displayed. The main area is titled 'ML_Project' and includes a lock icon. Below it is the sub-header 'Sample project'. There are three sections: 'Models' (text: 'This project has no models yet. Create a [new model](#).'), 'Jobs' (text: 'This project has no jobs yet. Create a [new job](#) to document your analytics pipelines.'), and 'Files' (list: 'Name ^', 'seaborn-data', 'analysis.ipynb', 'analysis.py', 'cdsw-build.sh').

New Project

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

* Project Name
Customer Churn

Project Description
This primary goal of this project is to build a logistic regression classification model to predict the probability that a group of customers will churn from a telecommunications company.

Project Visibility
 Private - Only added collaborators can view the project
 Public - All authenticated users can view this project.

Initial Setup
[Blank](#) [Template](#) [AMPs](#) [Local Files](#) [Git](#)

Templates include example code to help you get started.
[Python](#) ▾

New Project Runtime Setup

Runtime setup

Basic **Advanced**

Basic configuration adds the most commonly used Editors for the Kernel of your choice. To fine-tune the Editors available in the project, choose the Advanced tab.

Kernel

Python 3.7

Add GPU enabled Runtime variant

These runtimes will be added to the project:

- JupyterLab - Python 3.7 - Standard - 2022.04
- PBJ Workbench - Python 3.7 - Tech Preview - 2022.04
- Workbench - Python 3.7 - Standard - 2022.04

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

Collaborators

This project is **private**. Only collaborators can view and edit this project. [Change Settings](#).

Add Collaborator

Search by name, username, or email... Viewer Add

Collaborator	Permission	Actions
 bshimeldevuseast2_3	Owner	

Granting Admin or Contributor permission to other users may have security impact since it gives them full access to your project files and running sessions.

Chapter Topics

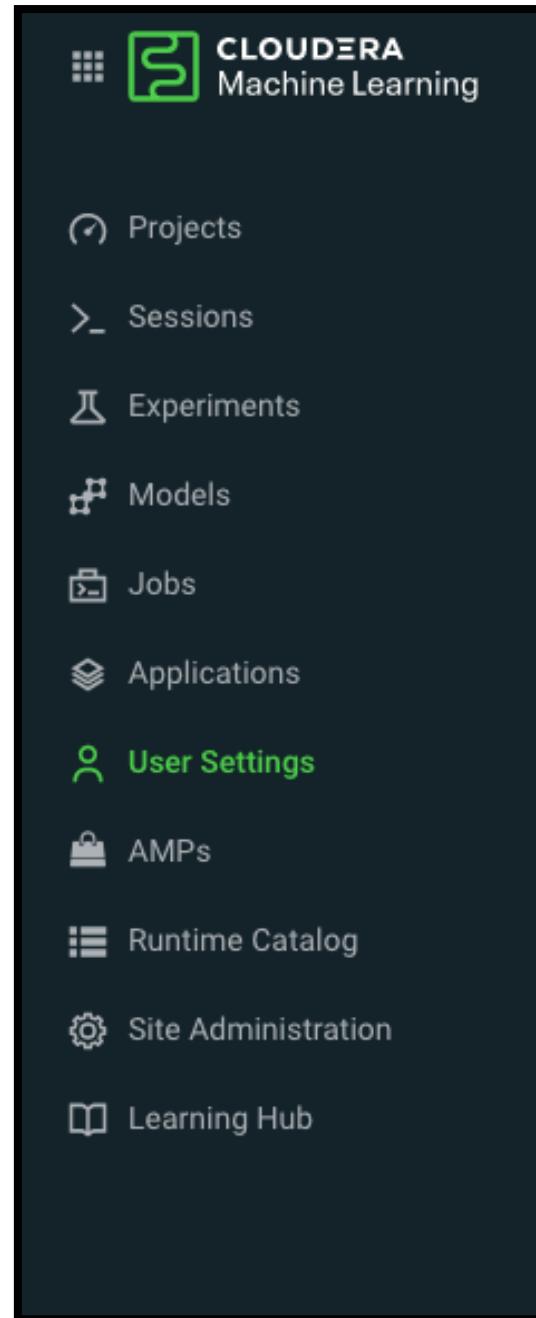
Introduction to CML

- 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

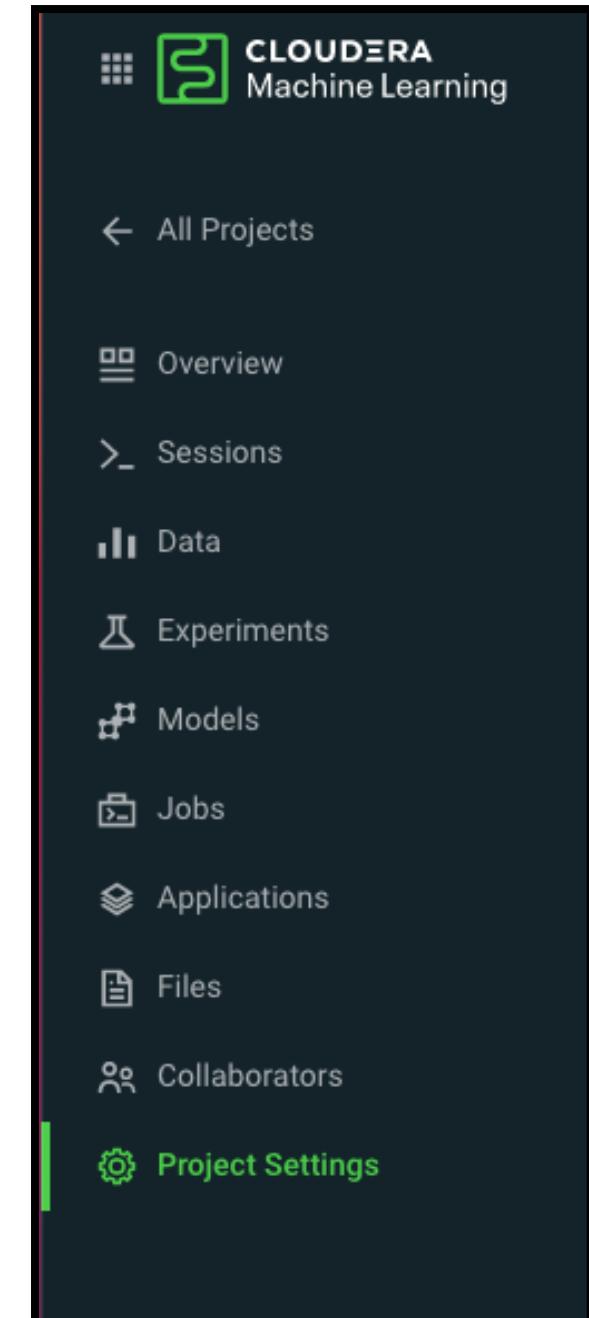
Settings

- Located in CML menu
- Are context sensitive

Workspace



Project



User Settings

■ Specific to the User within the Workspace

User Settings

Profile Teams Outbound SSH API Keys Remote Editing Environment Variables

Full Name
Bruce Shimel

Email
bshimel@cloudera.com

Bio
Bio

Update Account

Project Settings in CML

Project Settings

Options Runtime/Engine Advanced SSH Tunnels Data Connections Prototype Delete Project

Default Engine: ML Runtime ⓘ Legacy Engine ⓘ

Available Runtimes

Sessions and other workloads in this Project can use one of the Runtime variants configured below.

Editor	Kernel	Edition	Version	Jobs / Apps / Models using Runtime	
JupyterLab	Python 3.9	Standard	2022.04	0 / 0 / 0	X
Workbench	Python 3.9	Standard	2022.04	0 / 1 / 0	

Displaying 1 - 2 of 2 < 1 > 25 / page ▾

Site Administration

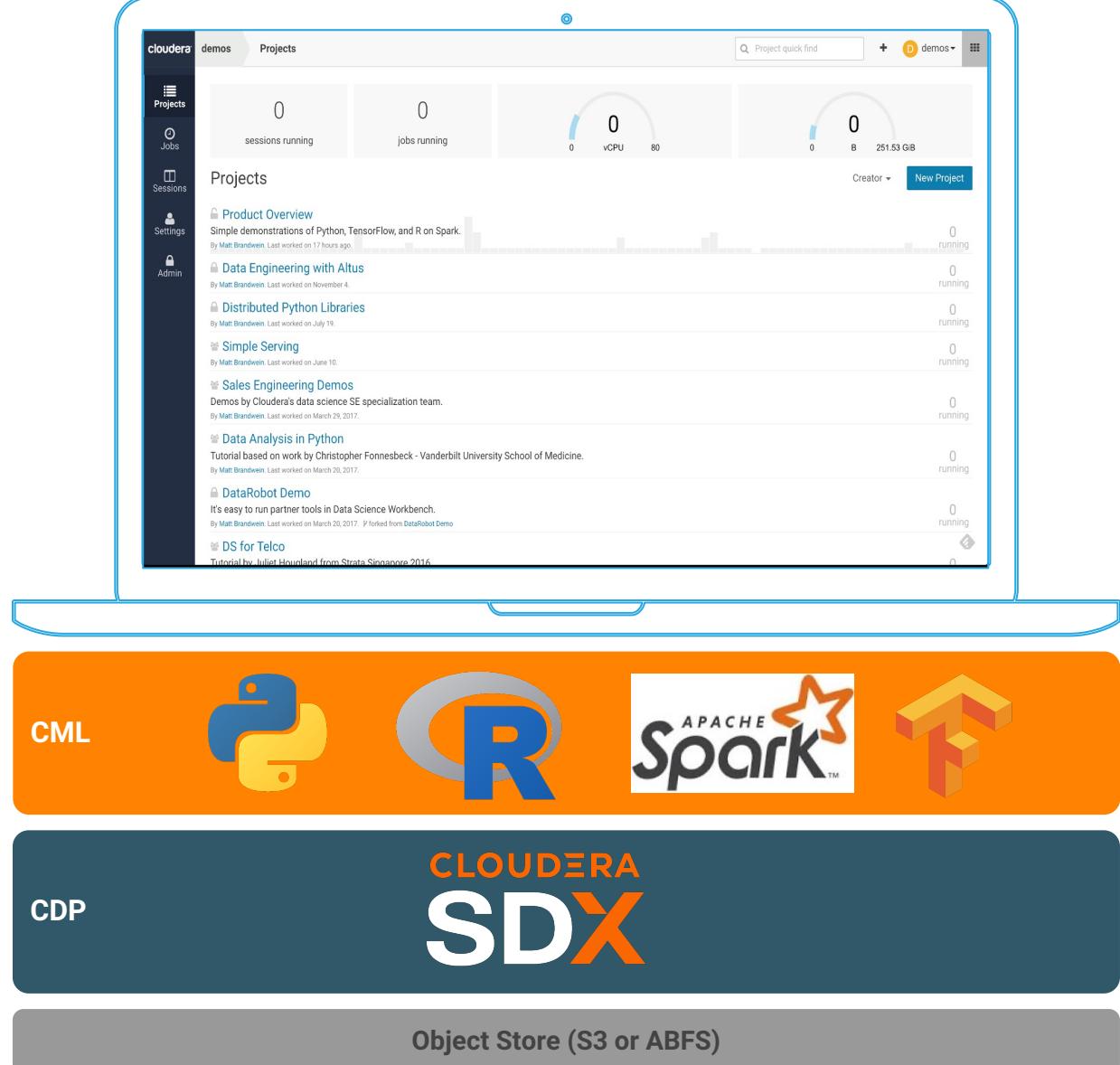
- Site administrators can manage the entire workspace
 - Monitor and manage all user activity across a workspace
 - Add new custom engines
 - Configure certain security settings

The screenshot shows the 'Site Administration / Overview' page. At the top, there is a navigation bar with links for Overview, Users, Teams, Usage, Quotas, Models, Runtime/Engine, Data Connections, Security, AMPs, Settings, and Support. The 'Overview' link is underlined, indicating it is the active page. To the right of the navigation bar is a search bar labeled 'Project quick find'. Below the navigation bar, the page title 'Site Administration' is displayed. Underneath the title, there is a section titled 'Cluster Monitoring' with the sub-instruction 'View cluster usage metrics and trends in custom built Grafana dashboards.' followed by a 'Grafana Dashboard' link. Below this, there is a section titled 'Cluster Metrics Snapshot' containing a table with two rows. The first row has 'Release' in the first column and 'dev' in the second column. The second row has 'Domain' in the first column and 'ml-7a767539-390.bshimel.kfjr-x0dh.cloudera.site' in the second column.

Release	dev
Domain	ml-7a767539-390.bshimel.kfjr-x0dh.cloudera.site

- **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



- **Isolated, containerized** 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

Start A New Session

Session Name
Exploration

Runtime

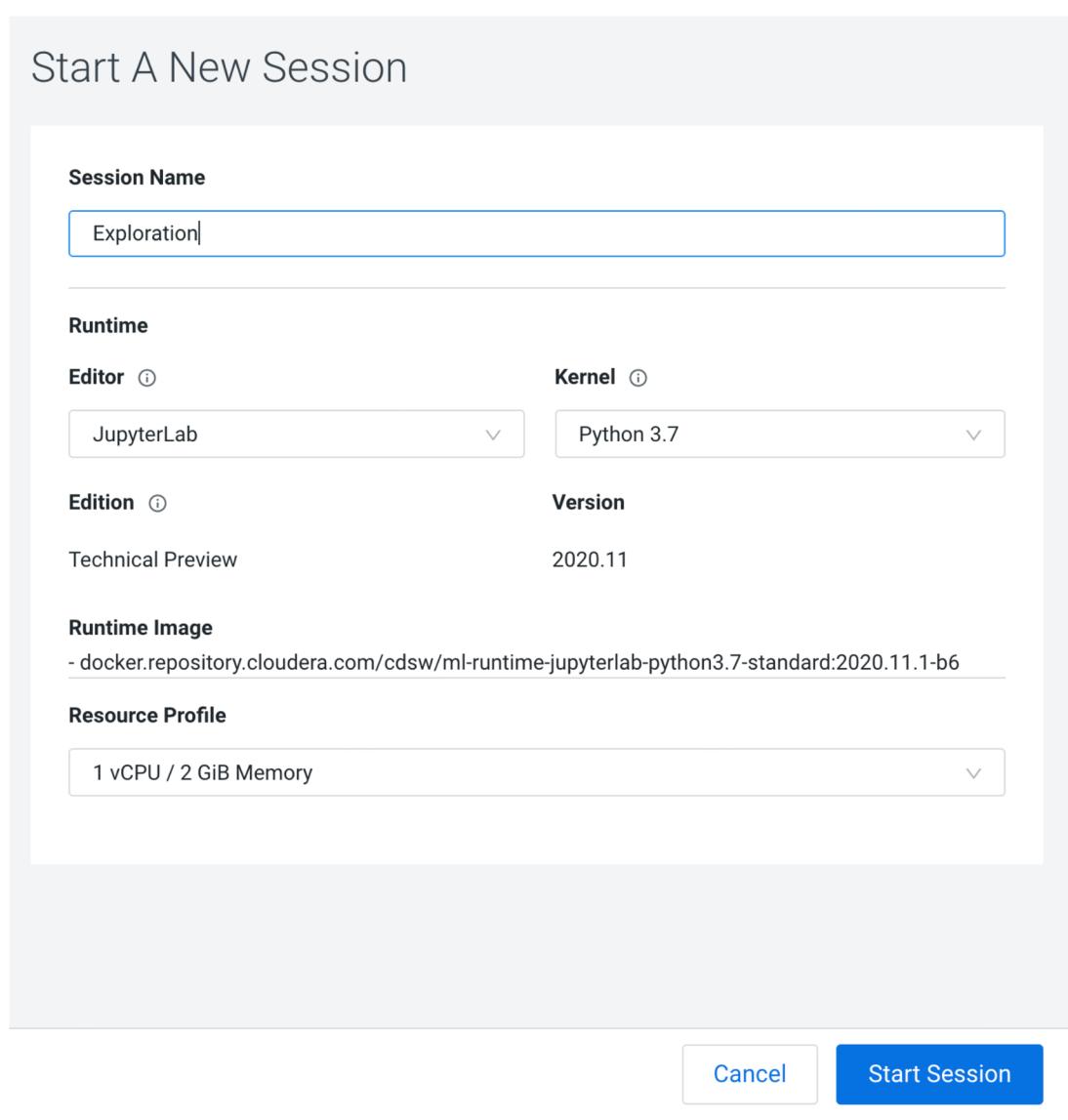
Editor (i) JupyterLab Kernel (i) Python 3.7

Edition (i) Technical Preview Version 2020.11

Runtime Image
- docker.repository.cloudera.com/cdsw/ml-runtime-jupyterlab-python3.7-standard:2020.11.1-b6

Resource Profile
1 vCPU / 2 GiB Memory

Cancel Start Session



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

 CLOUDERA
Machine Learning

- Projects
- Sessions
- Experiments
- Models
- Jobs
- Applications
- User Settings
- AMPs
- Runtime Catalog**
- Site Administration
- Learning Hub

Runtime Catalog

Search Runtime Editor All Kernel All Edition All

Status	Type	Editor	Kernel	Edition
✓	CLOUDERA	Workbench	Cloudera Data Visualization	CDV 6.3.7
✓	CLOUDERA	Workbench	Cloudera Data Visualization	CDV 6.4.1
✓	CLOUDERA	JupyterLab	Python 3.7	Nvidia GPU
✓	CLOUDERA	JupyterLab	Python 3.7	Standard
✓	CLOUDERA	JupyterLab	Python 3.8	Nvidia GPU
✓	CLOUDERA	JupyterLab	Python 3.8	Standard
✓	CLOUDERA	JupyterLab	Python 3.9	Nvidia GPU
✓	CLOUDERA	JupyterLab	Python 3.9	Standard
✓	CLOUDERA	PBJ Workbench	Python 3.7	Tech Preview
✓	CLOUDERA	Workbench	Python 3.7	Nvidia GPU
✓	CLOUDERA	Workbench	Python 3.7	Standard
✓	CLOUDERA	Workbench	Python 3.8	Nvidia GPU

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**

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

Start A New Session

Session Name
Test Session

Runtime

Editor	Kernel	Edition	Version
JupyterLab	Python 3.9	Standard	2022.04

Configure additional runtime options in [Project Settings](#).

Enable Spark [Spark 3.2.0 - CDE 1.15 - HOTFIX-2](#)

Runtime Image
- docker.repository.cloudera.com/cloudera/cdsw/ml-runtime-jupyterlab-python3.9-standard:2022.04.1-b6

Resource Profile
1 vCPU / 2 GiB Memory

[Cancel](#) [Start Session](#)

- **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

Install Data  Success

By Bruce Shimel – Session – 1 vCPU / 2 GiB Memory – 2 days ago

Session Logs    Export PDF

Getting Started

This is your **session**. Your **editor** is on the left and your **input prompt** is on the bottom.

To execute code from the editor, select the code and execute it with **Command-Enter** on Mac or **Ctrl-Enter** on Windows. You can also enter code at the prompt below.

Use **?command** to get help on a particular command.

```
> !chmod 755 cml/setup-data.sh
> !cml/setup-data.sh

S3_ROOT = s3a://cdp-storage-bshimel-456-class-2283/datalake-bshimel-456-class-2283
HIVE_EXT = s3a://cdp-storage-bshimel-456-class-2283/datalake-bshimel-456-class-2283/warehouse/tablespace/external/hive
PRINCIPAL = adm_bshimel_2283
DATALAKE = datalake-bshimel-456-class-2283
Aug 09, 2022 8:01:46 AM org.apache.knox.gateway.shell.KnoxSession createClient
INFO - Using default IAAS configuration
```

"!" executes commands in the shell, outside of the Python interpreter

Session Logs

Install Data  Success

By Bruce Shimel – Session – 1 vCPU / 2 GiB Memory – 2 days ago

Session Logs   

Name	Status
engine	unknown
spark executor 1	failed
spark executor 2	failed
spark executor 3	failed

Completed:

2022-08-09 08:01:39.257 1	INFO	Engine.Init.Root	r8p752wod8bl1hqy	Initial engine startup as UID data = {"uid":8536,"user":"cdsw"}
2022-08-09 08:01:39.257 1	INFO	EngineInit.LiveLog	r8p752wod8bl1hqy	Start creating LiveLog client data = {"user":"cdsw"}
2022-08-09 08:01:39.263 1	INFO	EngineInit.LiveLog	r8p752wod8bl1hqy	Finish creating LiveLog client data = {"user":"cdsw"}
2022-08-09 08:01:39.263 1	INFO	Engine.Init.Root	r8p752wod8bl1hqy	Not chowning /cdn and /output because we're not root data = {"u
2022-08-09 08:01:39.264 1	INFO	EngineInit.AddonSetup	r8p752wod8bl1hqy	Start setting HadoopCLI runtime addon on the engine data = {"user"
2022-08-09 08:01:39.273 1	INFO	EngineInit.Utils	r8p752wod8bl1hqy	Created symlink data = {"dest":"/usr/lib/hadoop","src":"/runtime-a
2022-08-09 08:01:39.276 1	INFO	EngineInit.Utils	r8p752wod8bl1hqy	Created symlink data = {"dest":"/usr/lib/hadoop-mapreduce","src":'
2022-08-09 08:01:39.279 1	INFO	EngineInit.Utils	r8p752wod8bl1hqy	Created symlink data = {"dest":"/usr/lib/hadoop-hdfs","src":"/runi
2022-08-09 08:01:39.284 1	INFO	EngineInit.Utils	r8p752wod8bl1hqy	Created symlink data = {"dest":"/usr/lib/jvm/java-8-openjdk-amd64"
2022-08-09 08:01:39.286 1	INFO	EngineInit.Utils	r8p752wod8bl1hqy	Created symlink data = {"dest":"/usr/lib/libhdfs.so","src":"/runi
2022-08-09 08:01:39.292 1	INFO	EngineInit.Utils	r8p752wod8bl1hqy	Created symlink data = {"dest":"/etc/java-8-openjdk","src":"/runi
2022-08-09 08:01:39.292 1	INFO	EngineInit.AddonSetup	r8p752wod8bl1hqy	Setting environment variables on the path data = {"envs":{ "CDH_M
2022-08-09 08:01:39.292 1	INFO	EngineInit.AddonSetup	r8p752wod8bl1hqy	EI_

Session List

adm_bshimel_2283 / Test_01 / Sessions

Project quick find + B adm_bshimel_2283 ▾

Creator All Show Running Only

Stop Selected Delete Selected

<input type="checkbox"/>	Status	Session	Kernel	Creator	Created At	Duration	
<input type="checkbox"/>	Timeout	Test	(Python 3.9 Workbench Standard)	Bruce Shimel	08/09/2022 4:12 AM	1h 8m 16s	<input type="button" value="Delete"/>
<input type="checkbox"/>	Success	Install Data	(Python 3.9 Workbench Standard)	Bruce Shimel	08/09/2022 4:01 AM	6m 11s	<input type="button" value="Delete"/>
<input type="checkbox"/>	Success	Install Dependencies	(Python 3.9 Workbench Standard)	Bruce Shimel	08/09/2022 3:58 AM	2m 48s	<input type="button" value="Delete"/>

Displaying 1 - 3 < 1 > 25 / page

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

The screenshot shows the 'User Settings' page in Cloudera Manager. The 'Outbound SSH' tab is active. In the 'User Public SSH Key' section, a large portion of the key is redacted in pink. A 'Reset SSH key' button is visible below the key field.

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

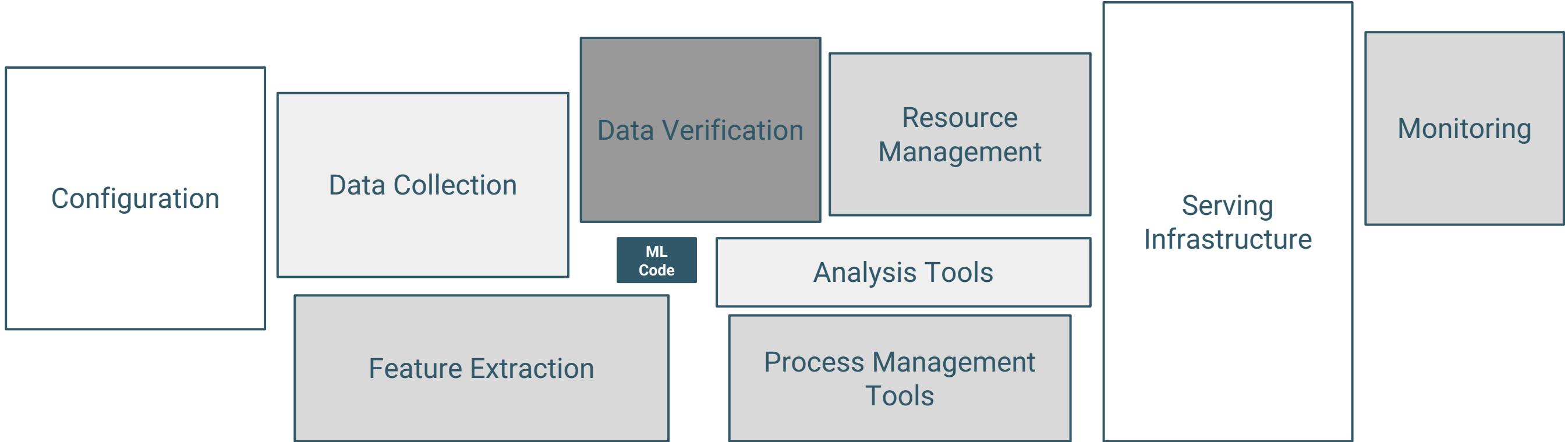
- Begin with https://
- Use for read-only access to public repos
- No extra setup required before using

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

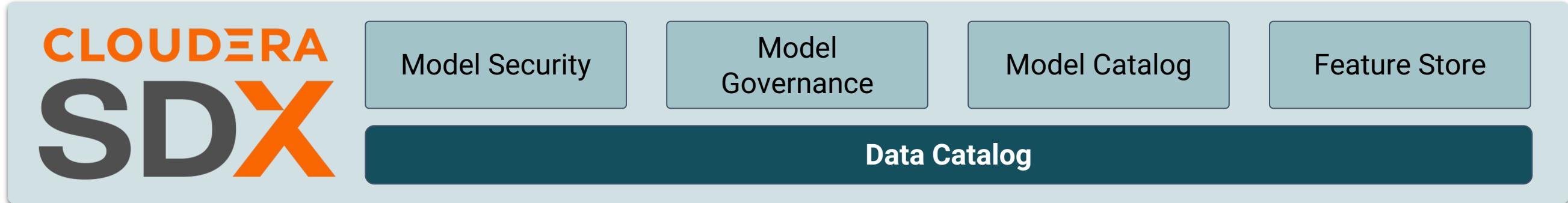
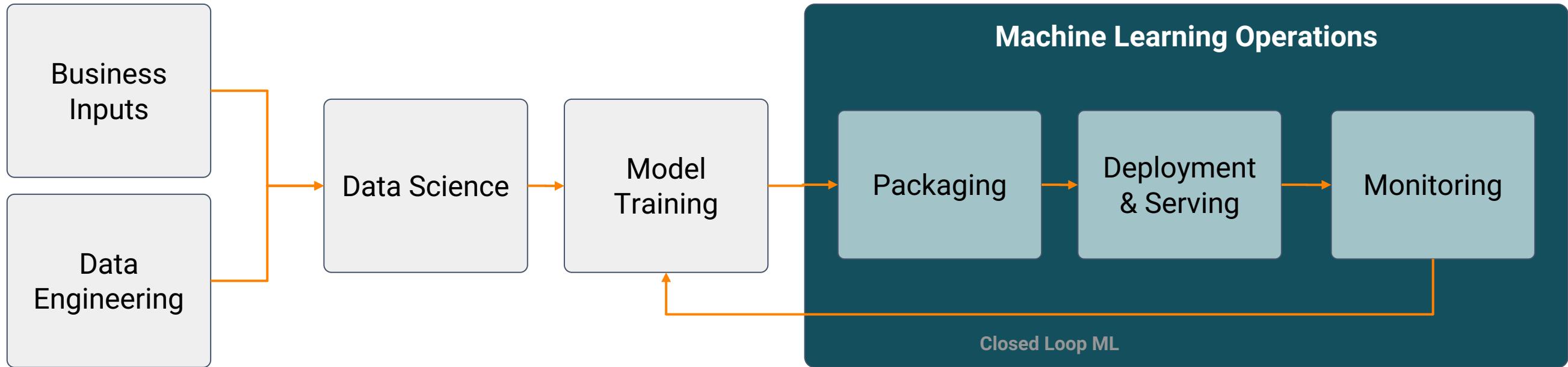
Google Paper



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

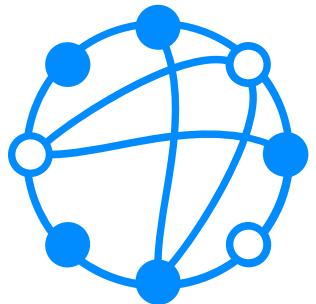
Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle.
The required surrounding infrastructure is vast and complex.

ML Operations Closes the Loop

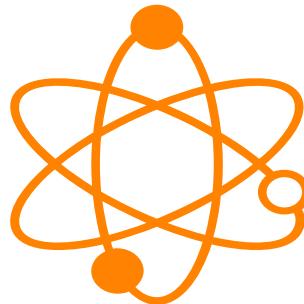


AI Applications

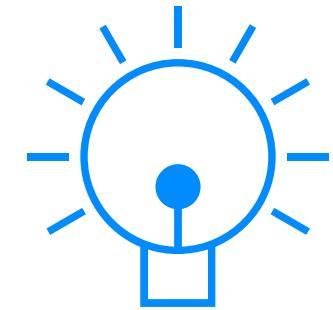
- Power more data science and AI use cases faster across the enterprise



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	<i>Unique name for the application</i>
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.*



ginablaber @ginablaber

The story of enterprise Machine Learning: "It took me 3 weeks to develop the model. It's been >11 months, and it's still not deployed."
@DineshNirmalIBM #StrataData #strataconf

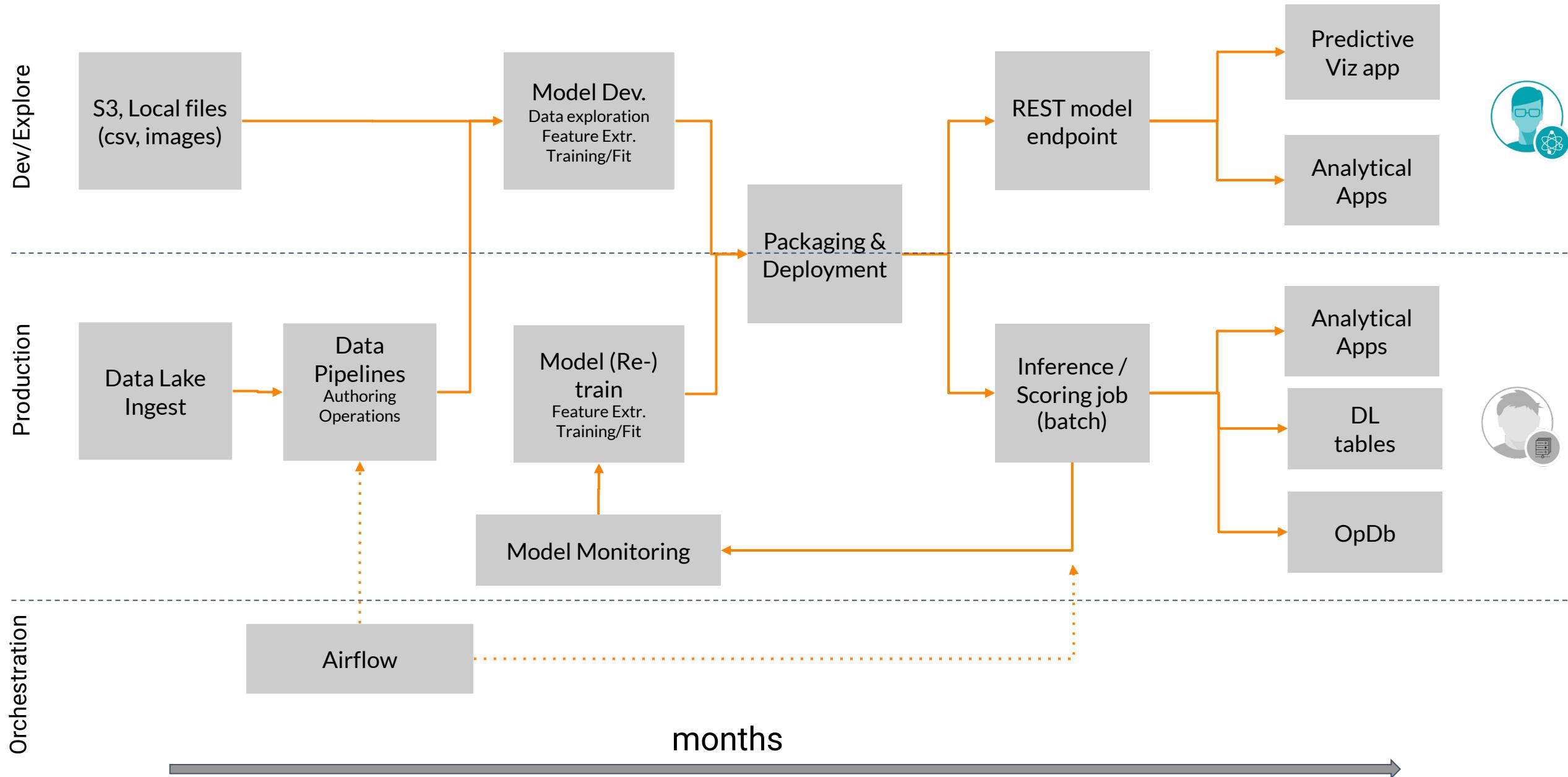


7

22



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

ID	Phone	Device	Device	Phone	TV	Smart	Screen	Phone	Phone	Phone	Phone	Churn
2876	Yes	Fiber optic	No	No	No	No	No	No	No	No	No	No
6380	Yes	Fiber optic	Yes	No	No	No	No	No	No	No	No	No
2909	Yes	Fiber optic	Yes	No	No	No	No	No	No	No	No	No
151	Yes	Fiber optic	Yes	No	No	No	No	No	No	No	No	No
5814	Yes	Fiber optic	No	No	No	No	No	No	No	No	No	No
2197	Yes	Fiber optic	No	No	No	No	No	No	No	No	No	No
2131	Yes	Fiber optic	No	No	No	No	No	No	No	No	No	No
813	Yes	Fiber optic	Yes	No	No	No	No	No	No	No	No	No
2055	Yes	Fiber optic	Yes	No	No	No	No	No	No	No	No	No
2060	Yes	Fiber optic	Yes	No	No	No	No	No	No	No	No	No
4247	Yes	Fiber optic	Yes	No	No	No	No	No	No	No	No	No
2980	No	Dish	No phone service	Yes	Yes	No	No	No	No	No	No	No

Churn Modeling with XGBoost

EXPLAINABILITY XGBOOST

The screenshot shows a user interface for deep learning image analysis. It includes a 'Select Model' dropdown with options like VGG16, VGG19, MOBILENET, and INCEPTIONV3. Below it, a 'Select Layer' dropdown shows 'LAYER 300' with a preview of extracted features. A 'Distance Metric' dropdown offers choices like COSINE, EUC., SQE., and MIN. At the bottom, there's a visualization of embeddings and a search result section showing a yellow car with a 78.3% confidence score and several other car images.

Deep Learning for Image Analysis

COMPUTER VISION IMAGES



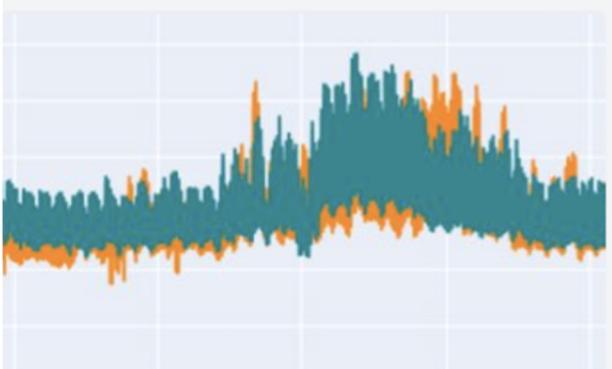
Neural Question Answering

NLP BERT



Airline Delay Prediction

XGBOOST



Structural Time Series

TIME SERIES

Section on Network Intrusion Data										
The KDDCUP99 dataset is a dataset of TCP connections that have been labeled as normal or representative of network attacks. Each TCP connection is represented as a set of attributes (or domain knowledge) pertaining to each connection such as the number of failed logins, connection duration, data bytes from source to destination, etc. The table below is a random sample of 100 rows.										
label	ID	Server Count	Server Error Rate	Server S Error Rate	R Error Rate	Server R Error Rate	Same Server Error Rate	Different Server Error Rate	Server Different Host Rate	Extinct Host Count
1	1	1.001%	0	0	0	0	1	0	0	1
1	1	1.001%	0	0	0	0	1	0	0	1
2	1	1.001%	0	0	0	0	1	0	0	1
2	2	0.001%	0	0	1	0.5	0.23	1	1	2
4	4	0.001%	0	0	0	0	1	0	0	0.0117%
6	6	0.001%	0	0	1	0.5	0.23	1	1	1
6	6	0.001%	1	1	0	0	0.01	0.09	0	1
7	7	0.025%	0	0	0	0	1	0	0.55	1
8	8	0.057%	0	0	1	0.67	0.5	1	1	1

Deep Learning for Anomaly Detection

ANOMALY DETECTION DEEP NEURAL NETS



Fraud Detection

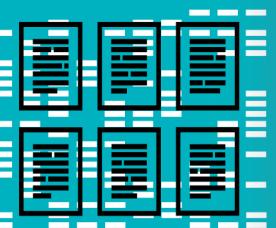
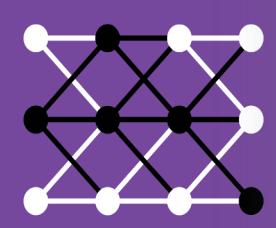
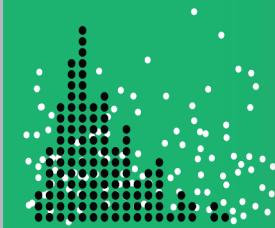
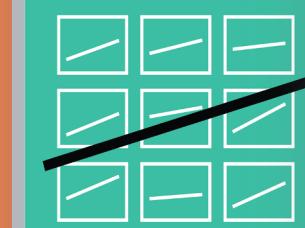
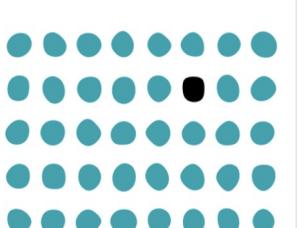
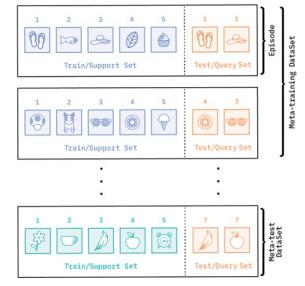
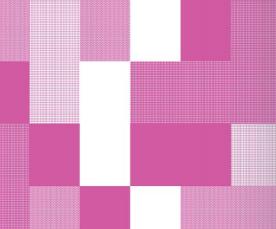
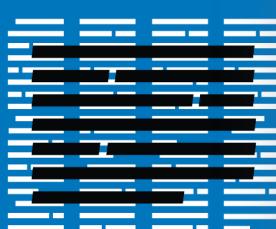
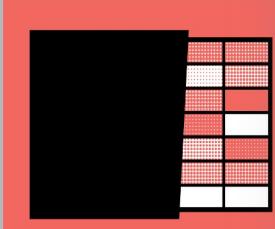
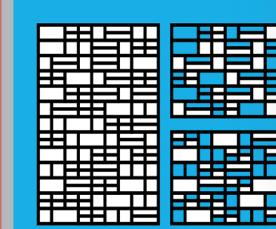
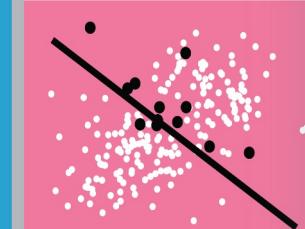
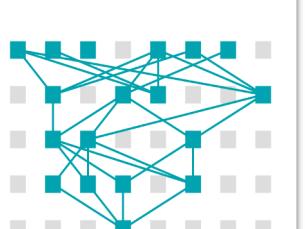
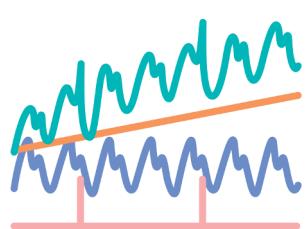
FRAUD DETECTION ANOMALY DETECTION

This interface is for sentiment analysis using Sparklyr and Tensorflow. It has a 'Model Result Output' section with a dropdown menu currently set to 'day'. Below it is a text input field for 'Test Sentence' and a text output field showing 'The model is 99.87 % confident that is Negative'. There are also 'SENTIMENT ANALYSIS' and 'SPARK' buttons at the bottom.

Sentiment Analysis with Sparklyr and Tensorflow

SENTIMENT ANALYSIS SPARK

Powered by Cloudera Fast Forward Labs Research

<p>Natural Language Generation</p> 	<p>Deep Learning: Image Analysis</p> 	<p>Probabilistic Programming</p> 	<p>Semantic Recommendations</p> 	<p>Federated Learning</p> 	<p>Transfer Learning for Natural Language Processing</p> 	<p>Deep Learning for Anomaly Detection</p> 	<p>Meta-Learning</p> 
<p>Probabilistic Methods for Realtime Streams</p> 	<p>Summarization</p> 	<p>Interpretability</p> 	<p>Multi-Task Learning</p> 	<p>Learning with Limited Labeled Data</p> 	<p>Deep Learning for Image Analysis 2019 Edition</p> 	<p>Causality for Machine Learning</p> 	<p>Structural Time Series</p> 

[Preview all of our research here](#)

How to Use AMPs?

■ Just pick one...

The screenshot displays the Cloudera Machine Learning interface, specifically the 'Applied ML Prototypes' section. The left sidebar includes links for Projects, Sessions, Experiments, Models, Jobs, Applications, User Settings, AMPs (selected), Runtime Catalog, Site Administration, and Learning Hub. The main area is a grid of 20 prototypes:

- Churn Modeling with scikit-learn**: CHURN PREDICTION, LOGISTIC REGRESSION.
- Deep Learning for Image Analysis**: COMPUTER VISION, IMAGE ANALYSIS.
- Deep Learning for Anomaly Detection**: ANOMALY DETECTION, TENSORFLOW.
- NeuralQA**: QUESTION ANSWERING, BERT.
- Structural Time Series**: TIME SERIES, PROPHET.
- Question Answering with Wikipedia**: WIKIPEDIA, The Free Encyclopedia.
- Explaining Models with LIME and SHAP**: INTERPRETABILITY, EXPLAINABILITY.
- Active Learning**: ACTIVE LEARNING, LEARNING WITH LIMITED LABELED DATA.
- MLFlow Tracking**: EXPERIMENT TRACKING.
- Few-Shot Text Classification**: NLP, FEW-SHOT LEARNING.
- Canceled Flight Prediction**: BINARY CLASSIFICATION, XGBOOST.
- Streamlit**: STREAMLIT, APPLICATIONS.
- Object Detection Inference Visualized**: COMPUTER VISION, OBJECT DETECTION.
- Getting Started with the CML API**: API, CML.
- TPOT + DASK**: Includes a logo for TPOT (Automated Machine Learning) and DASK (Distributed Python Computing).
- GENSI**: Includes a logo for GENSI (Generalized Semantic Indexing).

Click On It

- Sit back and enjoy the show!

The screenshot shows the Cloudera Machine Learning interface. The left sidebar has a dark theme with icons for All Projects, Overview, Sessions, Experiments, Models, Jobs, Applications, Files, and Team. The main area shows a project titled "Churn Modeling with XGBoost - 1..." under "Prototype Status". The title is "Executing Prototype Setup Steps for new Project" with a subtitle "Prototype Name: ML Churn Demo (v1)" and a description "Prototype to demonstrate building a churn model on CML". A message says "Completed all steps". Below are five step cards:

- Step 3** Job to train models. [View details](#) completed 10/19/2020 11:32 AM
- Step 4** Run model training job. [View details](#) completed 10/19/2020 11:32 AM
Running the job to train models.
- Step 5** Create the churn model prediction api endpoint [View details](#) completed 10/19/2020 11:32 AM
- Step 6** Build model [View details](#) completed 10/19/2020 11:37 AM

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**

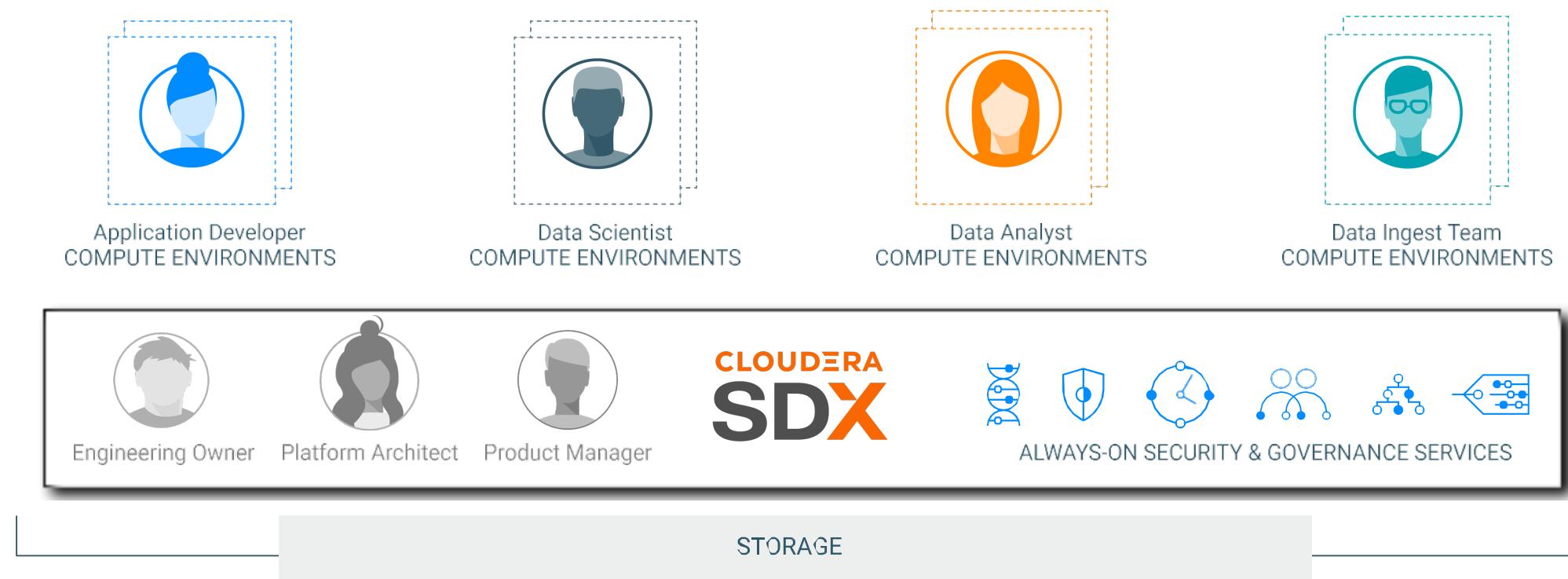


Data Access and Lineage

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 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 Ranger

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**



Apache **Atlas**

Data Catalog

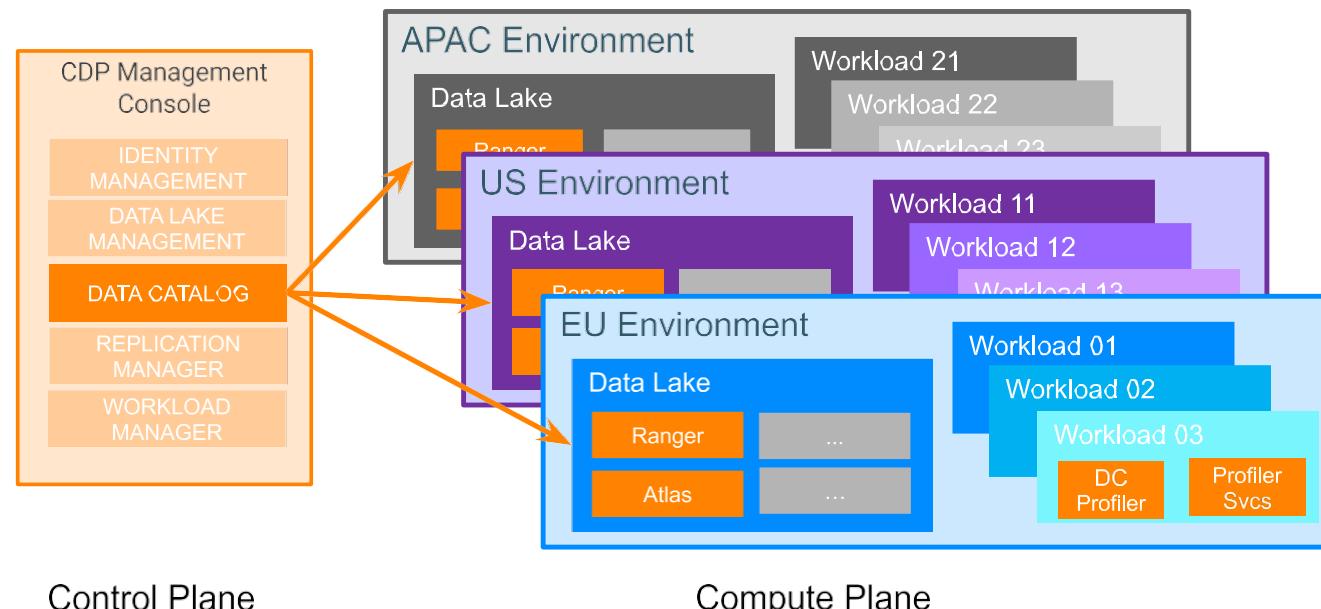
Data Catalog is a service in CDP for managing, securing, and governing data assets

- **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

The screenshot shows the Cloudera Data Catalog interface. On the left, there's a sidebar with 'Page links' (Datasets, Bookmarks, Profilers, Atlas Tags), 'Get Started', 'Help', and a search bar. The main area is titled 'Data Catalog / Search' and shows a table of 'Data Lakes'. The table has columns for Type, Name, Location, Owner, and Source. A red box highlights the 'Data Lakes' section, another highlights the search bar, and a third highlights the 'Tool links' (Atlas and Ranger) in the top right.

Type	Name	Location	Owner	Source
Spark Process	execution-2	/application_1596831586309_0...	-NA-	spark
Spark Process	execution-2	/application_1596831586309_0...	-NA-	spark
HDFS Path	/user/hdfs/uszips-target.csv	/hdfs://case702725-sparktmp...	-NA-	hdfs
Spark Application	Case 702725 application_1...	/application_1596831586309_0...	-NA-	spark
Spark Application	Case 702725 application_1...	/application_1596831586309_0...	-NA-	spark
Spark Process	execution-2	/application_1596831586309_0...	-NA-	spark
AWS S3 Pseudo Dir	/case702725/	/s3://repro-aws-sup-cdp-bucket...	-NA-	aws
Spark Process	execution-2	/application_1597076277511_0...	-NA-	spark
Spark Application	Case 702725 application_1...	/application_1596831586309_0...	-NA-	spark
Spark Application	Case 702725 application_1...	/application_1596831586309_0...	-NA-	spark

- 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

Data Catalog / Search		
<input type="text"/> Search		
Data Lakes	Type	Name
<input type="checkbox"/>	Hive Table	us_customers
<input type="checkbox"/>	Hive Table	txn_stg
<input type="checkbox"/>	Hive Table	txn_final
<input type="checkbox"/>	Hive Table	txn_hist
<input type="checkbox"/>	Hive Table	kudu_txn_final
<input type="checkbox"/>	Hive Table	ext_hist
<input type="checkbox"/>	Hive Table	prov_view
<input type="checkbox"/>	Hive Table	prov_view2
<input type="checkbox"/>	Hive Table	provider_summary
<input type="checkbox"/>	Hive Table	claims_view
<input type="checkbox"/>	Hive Table	claim_savings
<input type="checkbox"/>	Hive Table	consent_data
<input type="checkbox"/>	Hive Table	eu_countries
<input type="checkbox"/>	Hive Table	eventbatchop
<input type="checkbox"/>	Hive Table	telco_churn
<input type="checkbox"/>	Hive Table	employees_masked
<input type="checkbox"/>	Hive Table	employees
<input type="checkbox"/>	Hive Table	uk_employees
<input type="checkbox"/>	Hive Table	eu_employees
<input type="checkbox"/>	Hive Table	tax_2009
<hr/>		
Filters		
TYPE	Clear	
<input checked="" type="checkbox"/>	Hive Table	
<input type="checkbox"/>	HBase Table	
+ Add New Value		
OWNERS	Clear	
DATABASE	Clear	
ENTITY TAG	Clear	
+ Add New Value		
COLUMN TAG	Clear	
+ Add New Value		
CREATED WITHIN	Clear	
<input type="radio"/>	Last 7 days	
<input type="radio"/>	Last 15 days	

Asset Details: Schema

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

Unique values*

Minimum value

Maximum value

Null values

Mean of values

- Use the Schema tab for the Hive table page

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

Data Catalog / Asset Details

Name	Type	Data Lake	Dataset	⋮	Atlas
ww_customers	HIVE TABLE	demo-gov-ekar-datalake	0		

Overview Schema Policy Audit

Access Type: ALL Result: ALL [⟳](#)

Policy ID	Event Time	User	Resource Type	Access Type	Result	Access Enforcer	Client IP
8	09/15/2020 13:43:30 GMT	hive	@table	METADATA OPERATION	ALLOWED	ranger-acl	10.10.1.131
8	09/14/2020 20:42:59 GMT	hive	@table	METADATA OPERATION	ALLOWED	ranger-acl	10.10.1.131
13	09/14/2020 20:14:57 GMT	ekarnowski	@url	READ	ALLOWED	ranger-acl	10.10.1.131
8	09/14/2020 20:14:57 GMT	ekarnowski	@table	CREATE	ALLOWED	ranger-acl	10.10.1.131

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

A screenshot of the Apache Ranger Service Manager interface. The top navigation bar includes 'Ranger', 'Access Manager', 'Audit', 'Security Zone', and 'Settings'. A user dropdown shows 'adm_bshimel_22829'. The main area is titled 'Service Manager' and displays a grid of service configurations. Each configuration row has a folder icon, the service name, and a list of items. To the right of each item are three small icons: a magnifying glass, a pencil, and a trash can. The services listed are: HDFS (cm_hdfs), HBASE (cm_hbase), HADOOP SQL (Hadoop SQL), YARN (cm_yarn, dh_cml_edu_22829_yarn), KNOX (cm_knox), SOLR (cm_solr), KAFKA (cm_kafka), NIFI (NIFI), NIFI-REGISTRY, ATLAS (cm_atlas), ADLS (cm_adls), KUDU (cm_kudu), OZONE (cm_ozone), SCHEMA-REGISTRY, S3 (cm_s3), and KAFKA-CONNECT (cm_kafka_connect).

Service	Item	Action Icons
HDFS	cm_hdfs	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
YARN	cm_yarn	[Magnifying Glass] [Pencil] [Trash Can]
	dh_cml_edu_22829_yarn	[Magnifying Glass] [Pencil] [Trash Can]
KAFKA	cm_kafka	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
ATLAS	cm_atlas	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
OZONE	cm_ozone	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
S3	cm_s3	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
HBASE	cm_hbase	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
KNOX	cm_knox	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
NIFI		[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
ADLS	cm_adls	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
SCHEMA-REGISTRY		[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
HADOOP SQL	Hadoop SQL	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
SOLR	cm_solr	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
NIFI-REGISTRY		[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
KUDU	cm_kudu	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]
KAFKA-CONNECT	cm.kafka_connect	[Magnifying Glass] [Pencil] [Trash Can]
		[Magnifying Glass] [Pencil] [Trash Can]

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)

The screenshot shows the Ranger UI interface for managing Hadoop SQL Policies. The top navigation bar includes links for Access Manager, Audit, Security Zone, and Settings, along with a user profile for 'adm_bshimel_22829'. The current page is 'Hadoop SQL Policies' under the 'Service Manager' section. A search bar at the top allows users to search for policies. Below the search bar is a table listing 17 policies, each with columns for Policy ID, Policy Name, Policy Labels, Status, Audit Logging, Roles, Groups, Users, and Action.

Policy ID	Policy Name	Policy Labels	Status	Audit Logging	Roles	Groups	Users	Action
8	all - global	--	Enabled	Enabled	--	_c_ranger_admins_52763591	hive beacon dpprofiler hue + More..	<input type="button"/> <input type="button"/> <input type="button"/>
9	all - database, table, column	--	Enabled	Enabled	--	_c_ranger_admins_52763591	hive beacon dpprofiler hue + More..	<input type="button"/> <input type="button"/> <input type="button"/>
10	all - database, table	--	Enabled	Enabled	--	_c_ranger_admins_52763591	hive beacon dpprofiler hue + More..	<input type="button"/> <input type="button"/> <input type="button"/>
11	all - storage-type, storage-url	--	Enabled	Enabled	--	_c_ranger_admins_52763591	hive beacon dpprofiler hue + More..	<input type="button"/> <input type="button"/> <input type="button"/>
12	all - database	--	Enabled	Enabled	--	_c_ranger_admins_52763591 public	hive beacon dpprofiler hue + More..	<input type="button"/> <input type="button"/> <input type="button"/>
13	all - hiveservice	--	Enabled	Enabled	--	_c_ranger_admins_52763591	hive beacon dpprofiler hue + More..	<input type="button"/> <input type="button"/> <input type="button"/>
14	all - database, udf	--	Enabled	Enabled	--	_c_ranger_admins_52763591	hive beacon dpprofiler hue + More..	<input type="button"/> <input type="button"/> <input type="button"/>
15	all - url	--	Enabled	Enabled	--	_c_ranger_admins_52763591	hive beacon dpprofiler hue + More..	<input type="button"/> <input type="button"/> <input type="button"/>
16	default database tables columns	--	Enabled	Enabled	--	public	--	<input type="button"/> <input type="button"/> <input type="button"/>
17	Information_schema database tables columns	--	Enabled	Enabled	--	public	--	<input type="button"/> <input type="button"/> <input type="button"/>

Policy Conditions

- Policies are defined using allow and deny conditions

Allow Conditions :

Select Role	Select Group	Select User	Policy Conditions	Permissions	Delegate Admin	
<input type="checkbox"/> Admins	Select Groups	<input type="checkbox"/> hive <input type="checkbox"/> rangerlookup <input type="checkbox"/> impala <input type="checkbox"/> admin	Add Conditions +	<input type="checkbox"/> select <input type="checkbox"/> update <input type="checkbox"/> Create <input type="checkbox"/> Drop <input type="checkbox"/> Alter <input type="checkbox"/> Index <input type="checkbox"/> Lock <input type="checkbox"/> All <input type="checkbox"/> Read <input type="checkbox"/> Write <input type="checkbox"/> ReplAdmin <input type="checkbox"/> Service Admin <input type="checkbox"/> Temporary UDF Admin <input type="checkbox"/> Refresh	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Select Roles	Select Groups	<input type="checkbox"/> {OWNER}	Add Conditions +	<input type="checkbox"/> All	<input checked="" type="checkbox"/>	<input type="checkbox"/>
+ <input type="button" value="Add"/>						

Exclude from Allow Conditions :

Select Role	Select Group	Select User	Policy Conditions	Permissions	Delegate Admin	
Select Roles	Select Groups	Select Users	Add Conditions +	Add Permissions +	<input type="checkbox"/>	<input type="checkbox"/>
+ <input type="button" value="Add"/>						

Deny All Other Accesses : False

View Policies

- List of policies provides overview

The screenshot shows a table with the following data:

Policy ID	Policy Name	Policy Labels	Status	Audit Logging	Roles	Groups	Users	Action
75	all - database, table, column	--	Enabled	Enabled	Admins	--	hive rangerlookup impala admin + More...	
76	all - database, udf	--	Enabled	Enabled	Admins	--	hive rangerlookup impala {OWNER}	
77	access: us_customers_table	--	Enabled	Enabled	Admins	us_employee dpo public	hive	
78	access: ww_customers	--	Enabled	Enabled	--	us_employee eu_employee etl	hive etl_user impala	
79	access: eu_countries	--	Enabled	Enabled	--	public eu_employee	--	
80	prohibit zipcode, insuranceId, bl...	--	Enabled	Enabled	--	analyst	--	
81	prevent UDF create/drop	--	Enabled	Enabled	--	us_employee	--	
90	access: Information Schema pol...	--	Disabled	Enabled	--	us_employee etl	hive	

- 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)

The screenshot shows a database query interface with the following details:

- Query:**

```
1 SELECT surname, streetaddress, country,
2    | age, password, nationalid, ccnumber, mrn, birthday
3 FROM worldwidebank.us_customers
4 LIMIT 50
```
- Results:** A table titled "Results (50)" with columns: surname, streetaddress, country, age, password.
- Annotations:**
 - A box labeled "Street address masked (redacted)" covers the "streetaddress" column.
 - A box labeled "Password masked (hashed)" covers the "password" column.
 - Two orange arrows point from these boxes to the respective column headers in the results table.

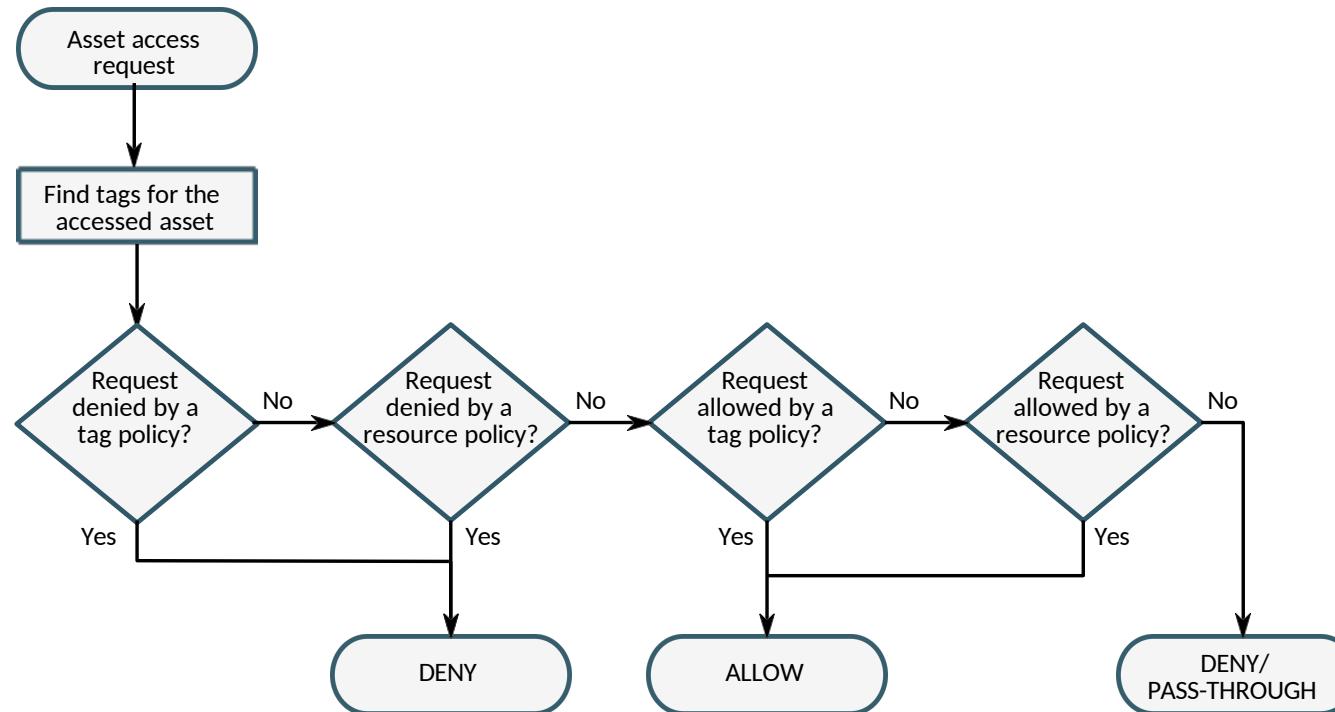
	surname	streetaddress	country	age	password
1	Powers	nnnn XXXXXX XXXXX	US	52	76a3fe33eb676cb12b99f00400772e7f5e5abc00950f432a7428d8e356
2	Whitman	nnnn XXXXXX XXXXX	US	55	6103bc600fc877e5cdf250521e422fc738544bff75165f335e5d2437e67
3	Marone	nnn XXXX XXXXX	US	47	da01407d5922624b1dd67c22e849cb9ee0ba0fbf3b25ab6e3f78ed234a
4	Harp	nnnn XXX XXX XXXXX	US	26	717051c13b6bda179f3f1fca42d415524c7e373dcfec265c0973badf28
5	Pereira	nnnn XXXXXXXX XXXXXXX	US	80	47d2ebe1a1c1146698e22d126282356ac2ccbe0ca60a9d4d6d4f757cd6

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

The screenshot shows the 'Access' tab in the Cloudera Manager interface. A search bar at the top contains the query 'USER: joe_analyst'. Below it, a checkbox labeled 'Exclude Service Users' is unchecked. The main table displays two rows of access logs:

Policy ID	Policy Version	Event Time	Application	User	Service	Name / Type	Permission	Result	Action
80	1	09/09/2020 05:47:42 AM	hiveServer2	joe_analyst	Hadoop SQL Hadoop SQL	worldwidebank/ww_cust... @column	SELECT	select Denied	ra...
89	1	09/09/2020 05:47:42 AM	hiveServer2	joe_analyst	Hadoop SQL Hadoop SQL	worldwidebank/ww_cust... @table	ROW_FILTER	select Allowed	ra...

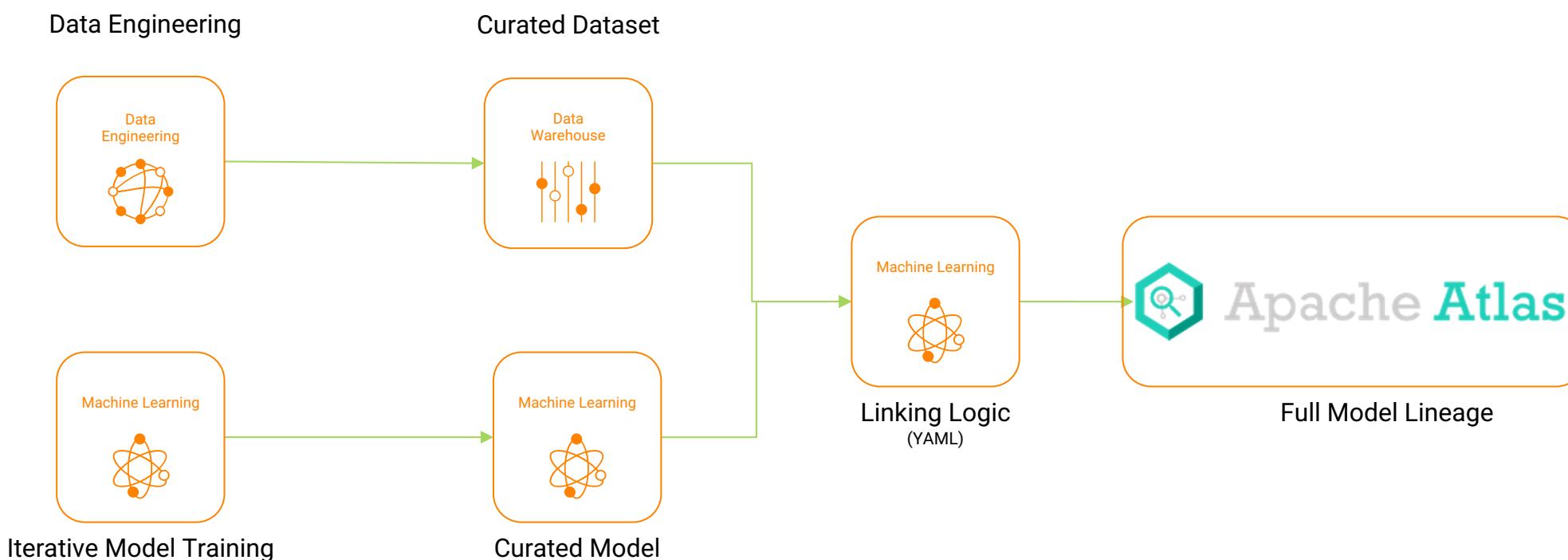
A modal window titled 'Hive Query' is open over the second row, showing the executed SQL query:

```
select zipcode, insuranceid, bloodtype
from worldwidebank.ww_customers
limit 10
```

- 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)

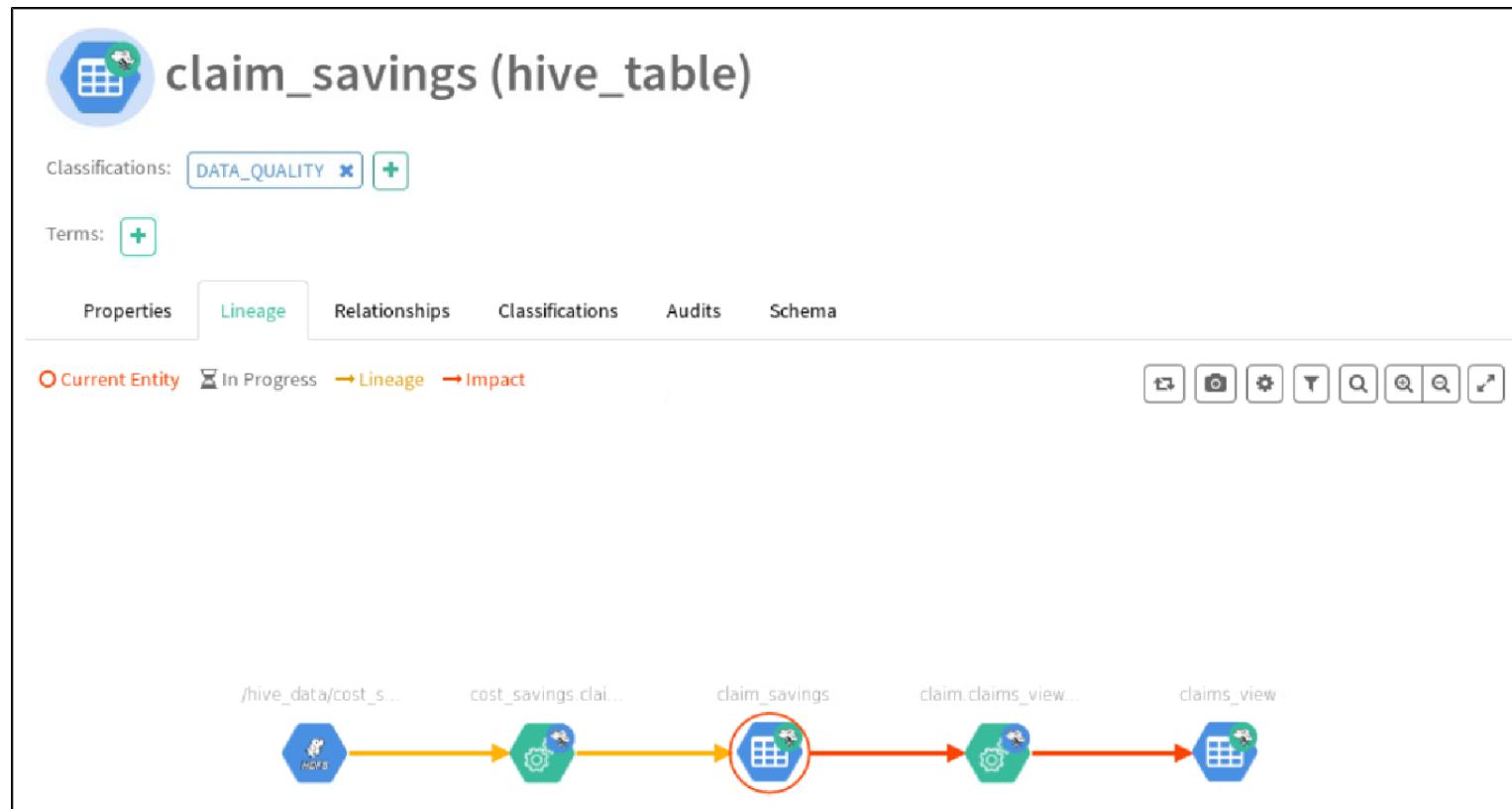
Inspecting Data using Apache Atlas

- 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



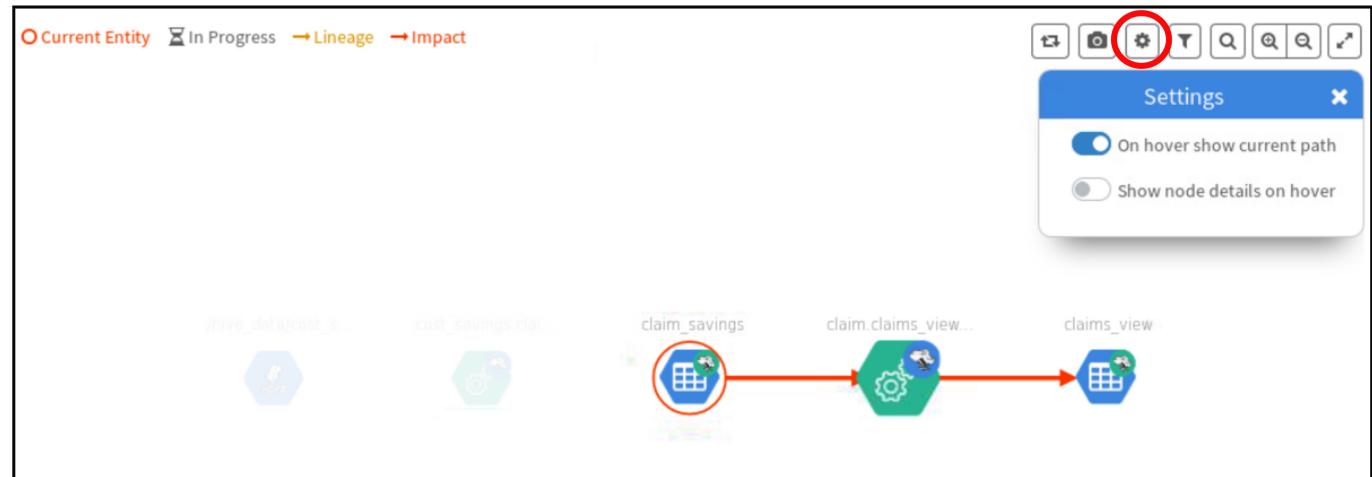
Viewing an Object's Lineage in Apache Atlas

- 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

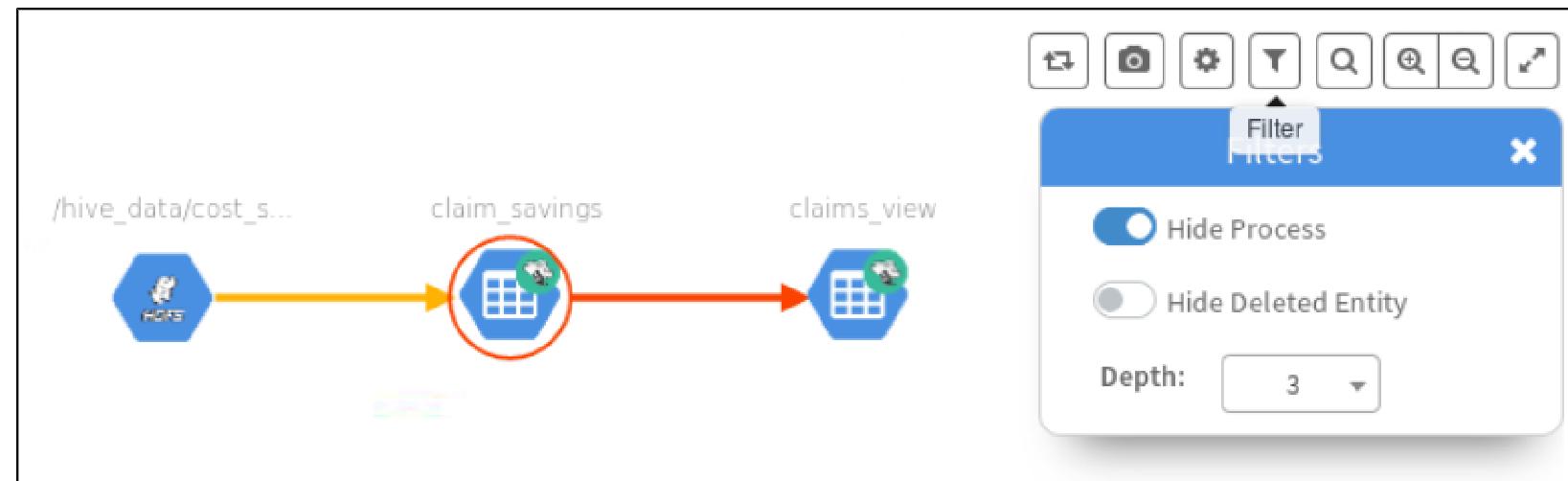
- Click on a node for more detail
- Includes the guid as a link to that node's entity page
- Also shows classifications and terms attached to the entity, if any

Key	Value
guid	e80267cc-8590-4c72-99f9 -bd28ddd1e87a
typeName	hive_process
name	cost_savings.claim_savi ngs@cm:1598027871000
qualifiedName	cost_savings.claim_savi ngs@cm:1598027871000
status	ACTIVE
classifications	N/A
term	N/A

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**

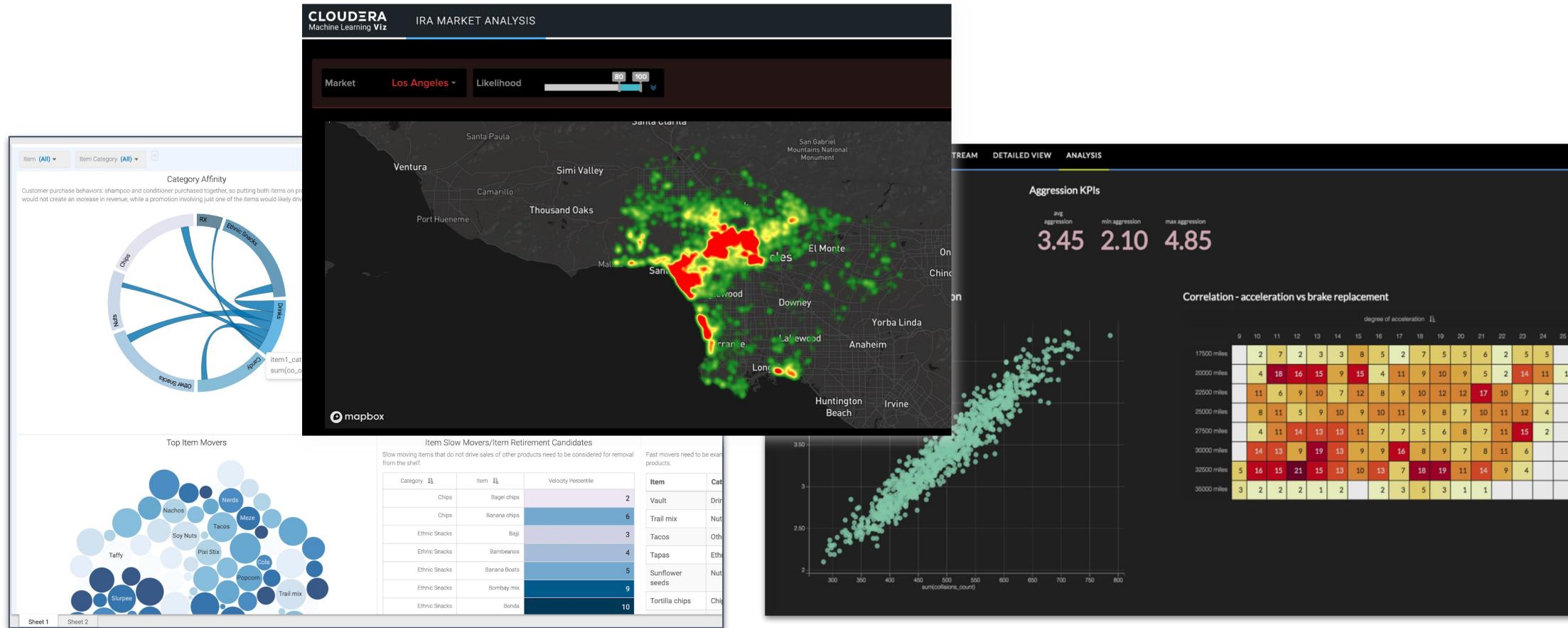


Data Visualization

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

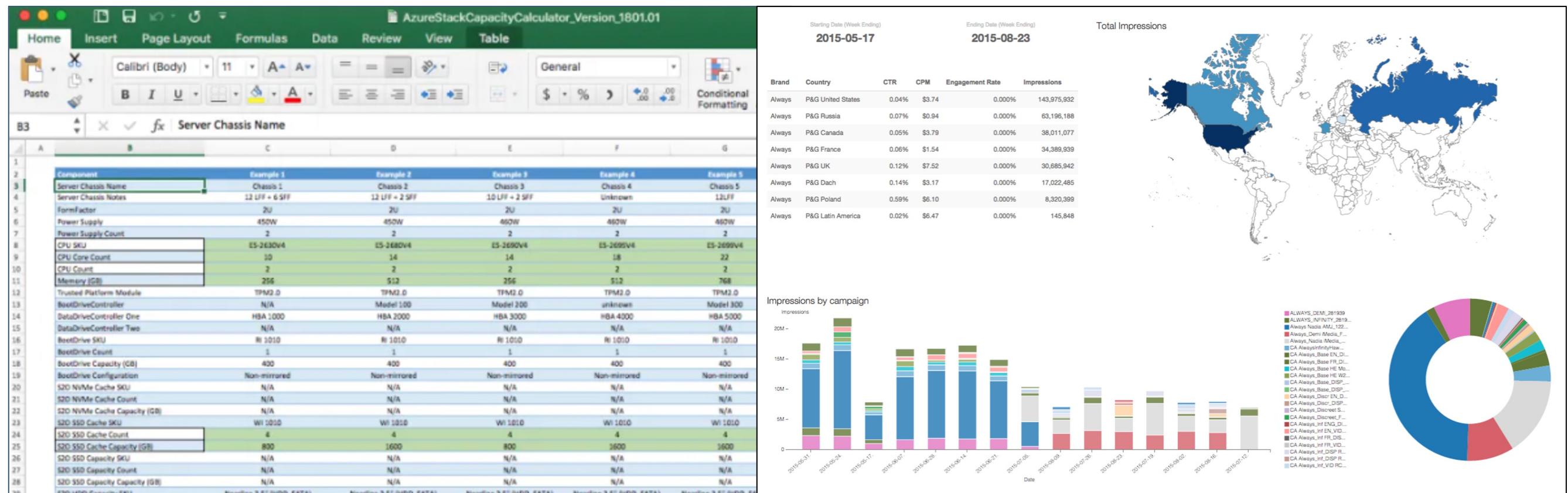
What is CDP Data Visualization?



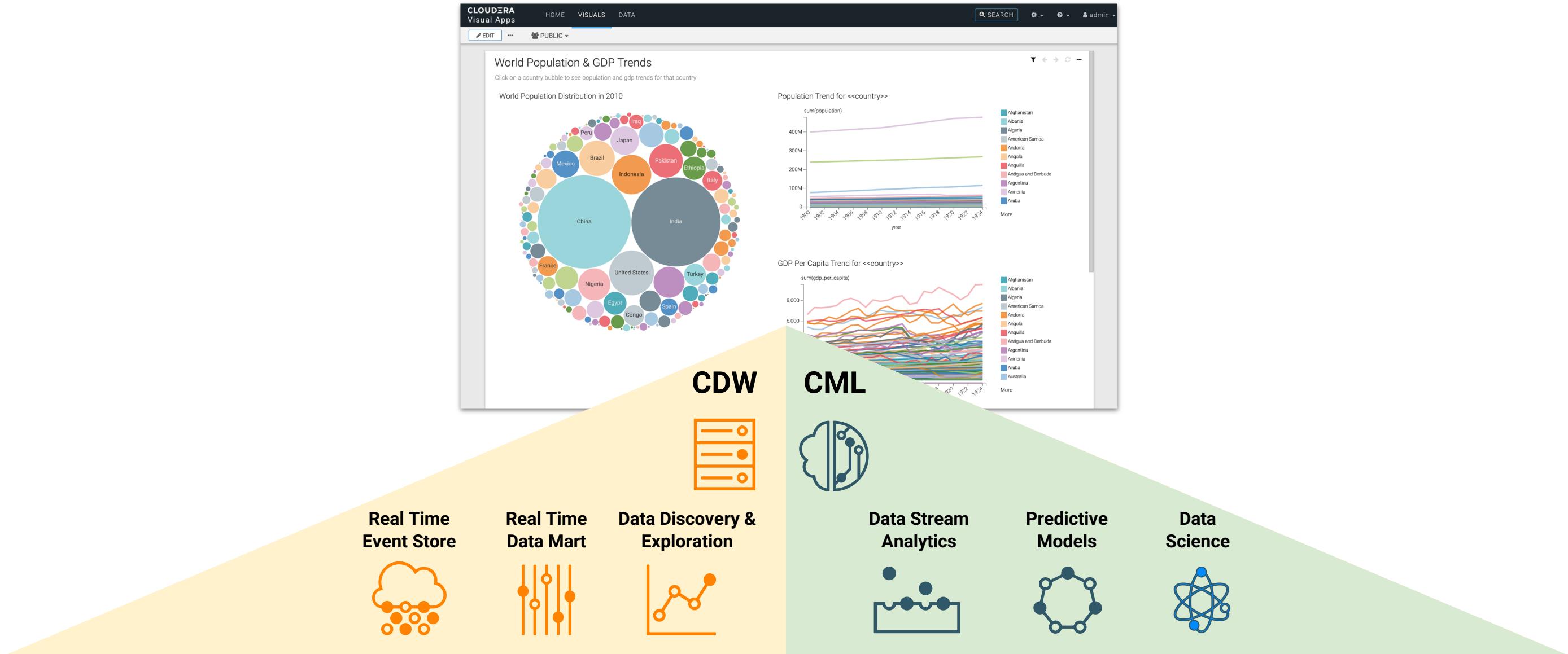
CDP Data Visualization (CDV) enables data engineers, business analysts, and data scientists to **quickly and easily explore data, collaborate, and communicate explainable insights across the data lifecycle.**

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



Bring All Your Data Together



Native Data Visualizations in CML

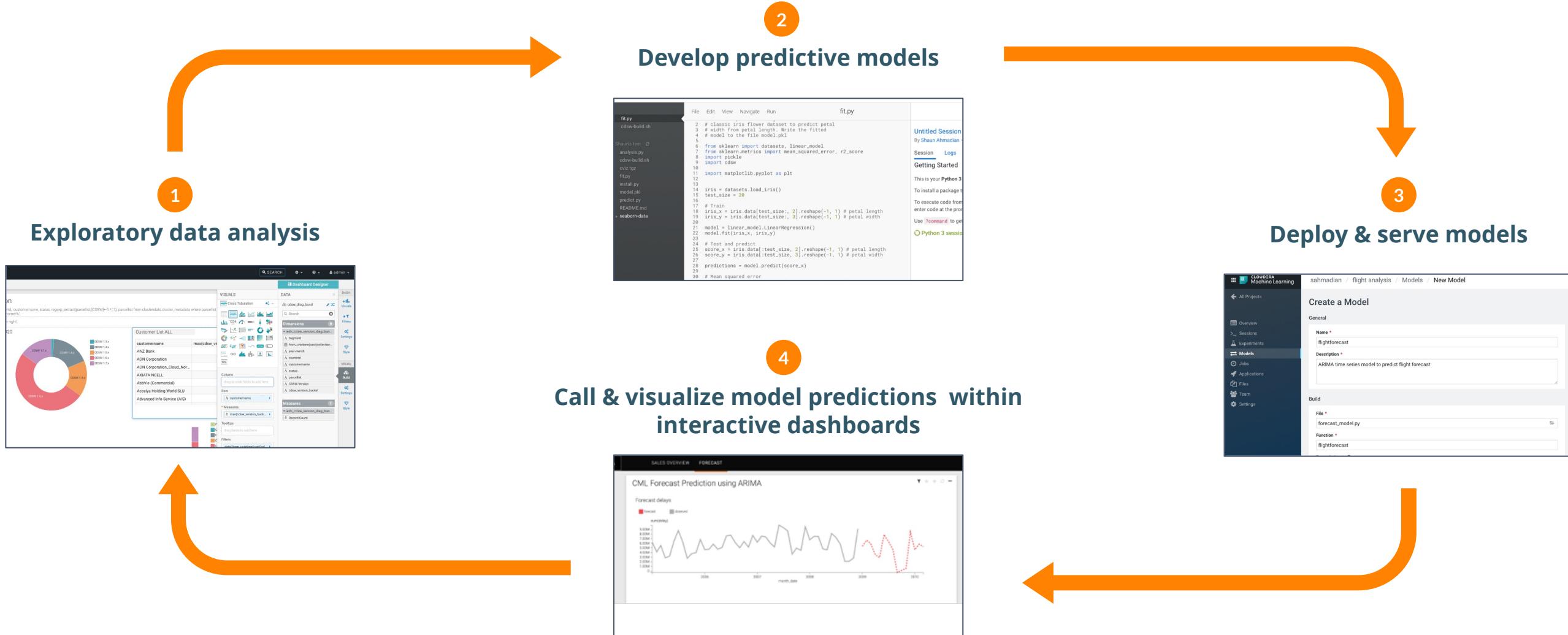
The screenshot displays the Cloudera Machine Learning (CML) interface, specifically the 'Visuals' section. The top navigation bar includes links for HOME, SQL, VISUALS, and DATASETS, along with a search bar and user authentication information. On the left, a sidebar lists project management options like Overview, Sessions, Data (which is selected), Experiments, Models, Jobs, Applications, Files, Collaborators, and Project Settings. The main area is a grid of 20 preview cards, each showing a different type of data visualization:

- Ride Dashboard
- Deficiency Details: <<county:Queens>>
- State of NYC
- Sample App
- Store Details<<owner_name:>>
- Cereal Comparisons
- Earthquakes Around the World
- Life Expectancy Dashboard
- World Population & GDP Trends
- Animated world population - GDP vs life
- US State Population Trends
- Census Dashboard
- Global Threats
- Time & Industry Threat View
- Inspector View
- Consumer View
- Iris species w/ images
- Taxi rides application

Each card contains a small thumbnail of the visualization and a 'View' button.

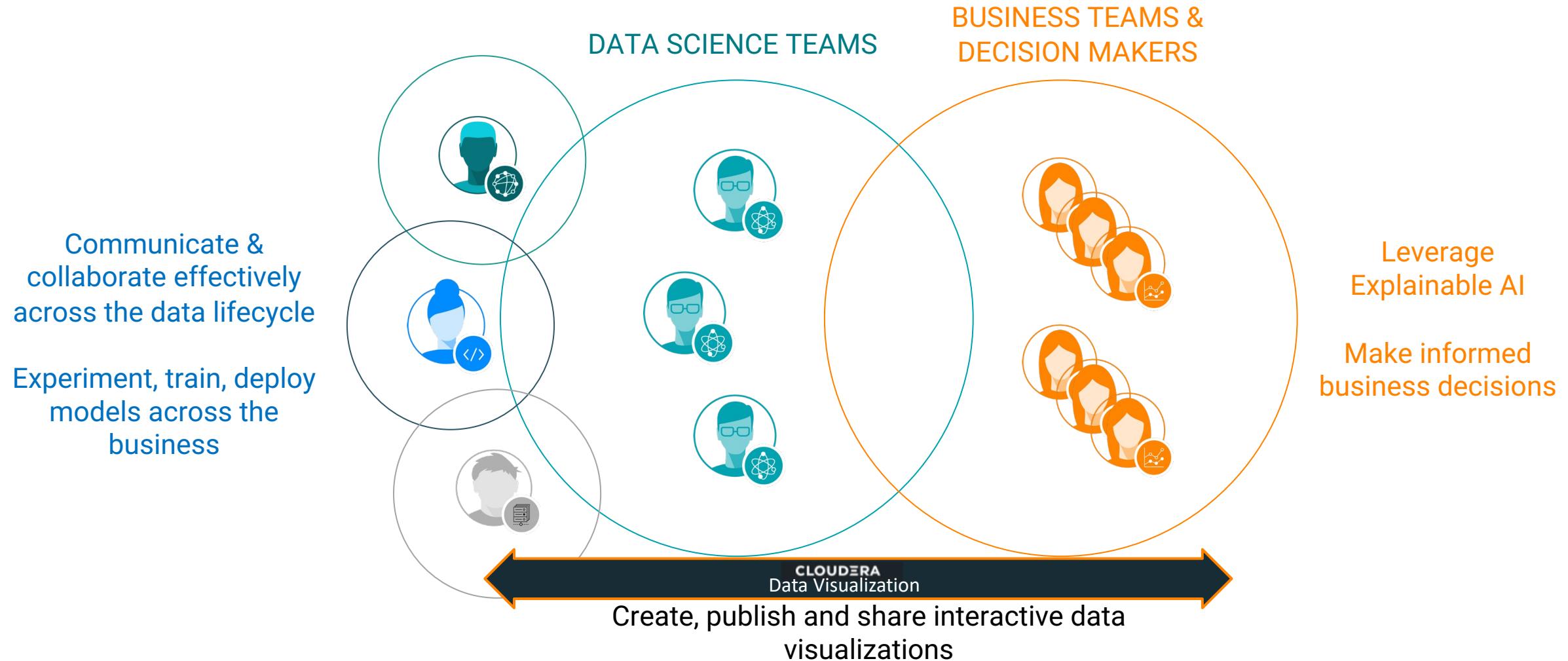
Integrated Data Visualizations for ML Workflows

Create fast, explainable insights from ML models



Visual Collaboration for Getting to Production

Accelerating Production ML Workflows from Raw Data to Business Impact



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

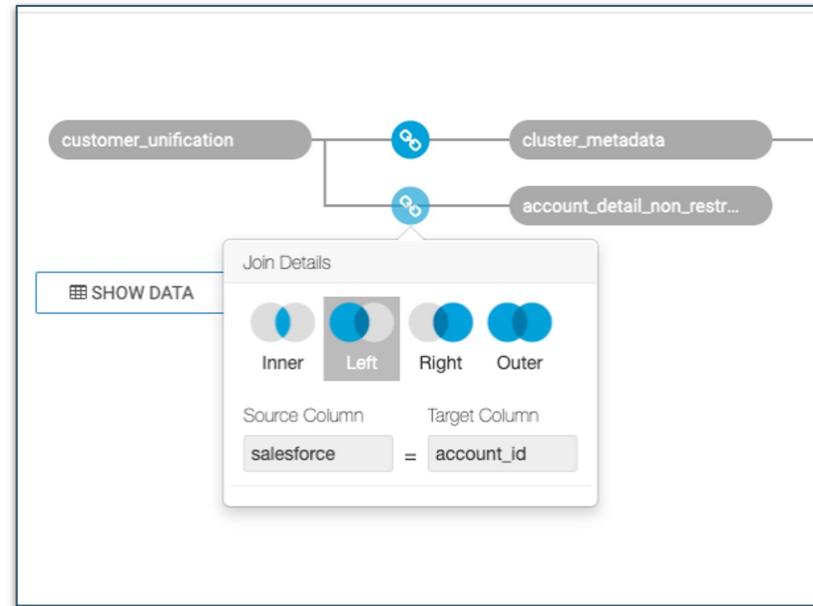
- Real time custom dashboards natively in CDP
- Visualize across the data lifecycle to discover optimization opportunities

Visualization Benefits

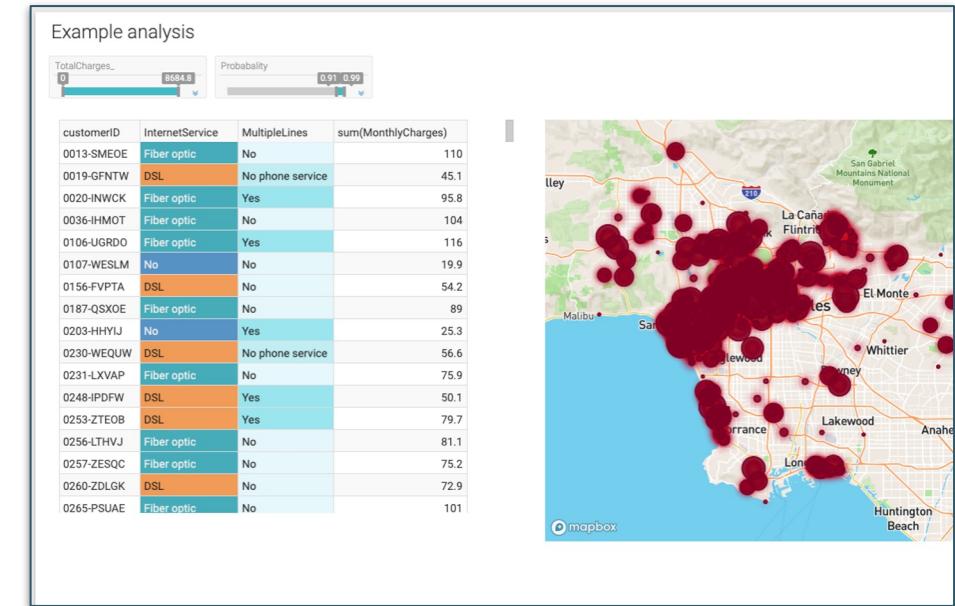
- Work within the **workflow of ML**
- Simple drag & drop interface
- Interactive, dynamic dashboard
- Completely **web based** for integration through links, embedding, HTML/JS
- Many ways to **share and collaborate**
 - Bookmark link
 - Emailed reports
 - Customized applications
- Access to all of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



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

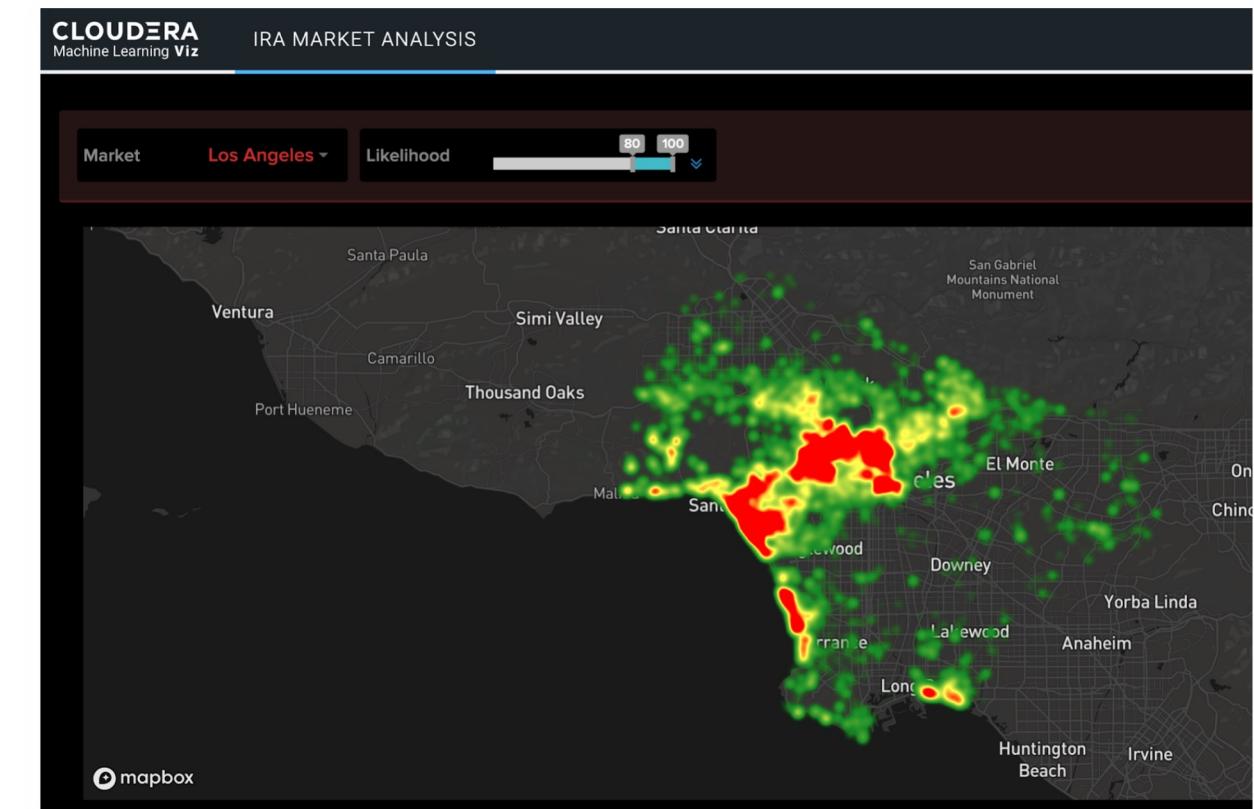
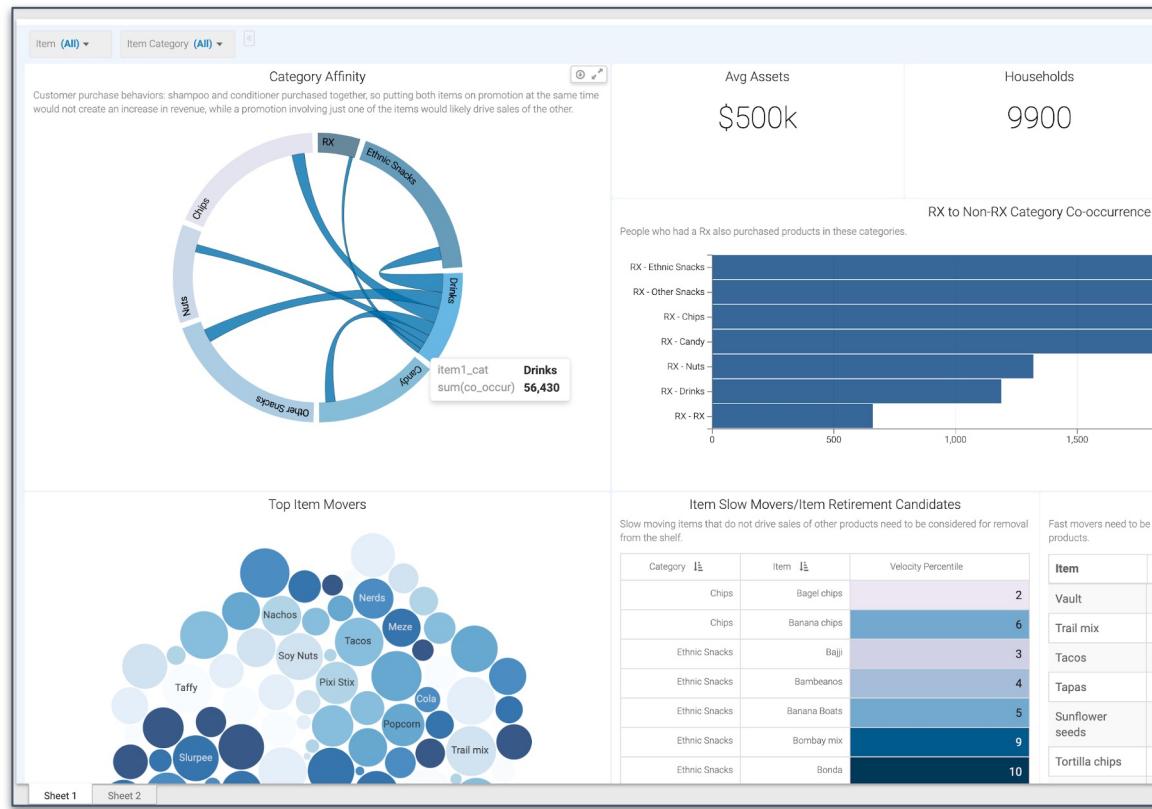
Visuals

- 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



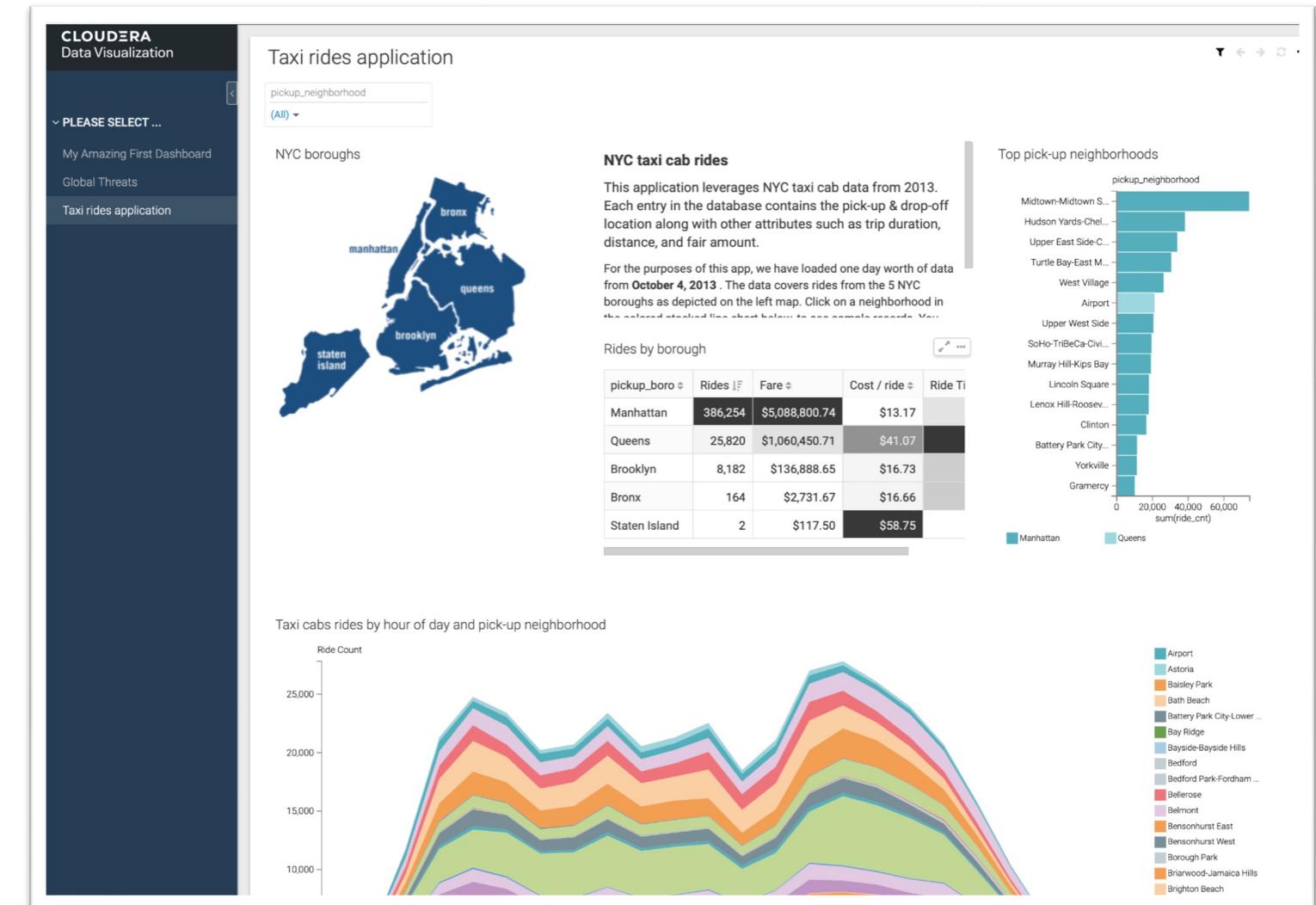
Dashboards

- 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



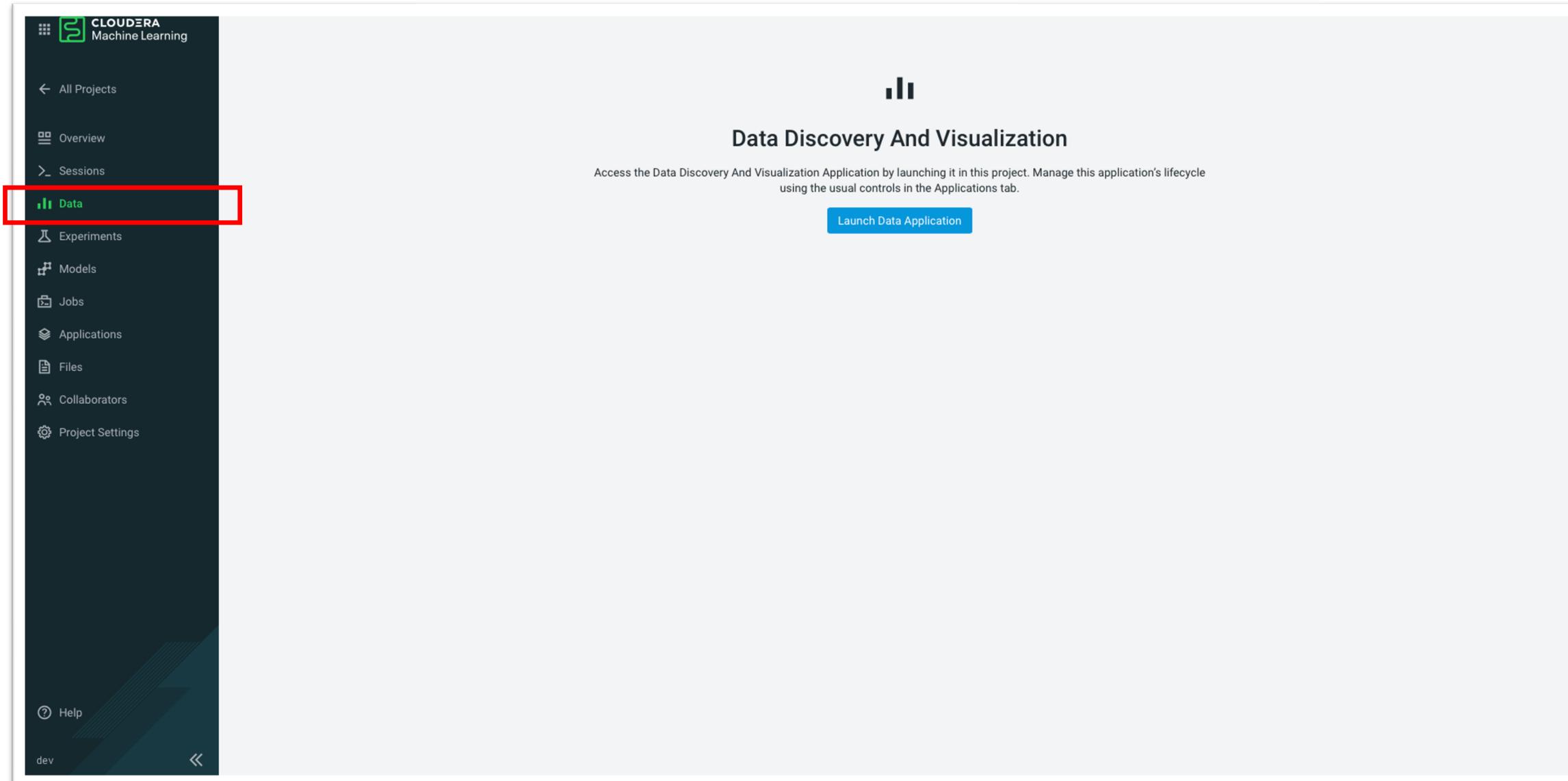
Applications (Apps)

- 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

The screenshot shows the Cloudera Machine Learning (CDV) Home Page. The left sidebar has a dark theme with white text and icons. The 'Data' option is highlighted with a red box. The main content area is titled 'Get Started' and contains five numbered steps:

- Sync Connections: Shows a comparison between the Cloudera Management Console (left) and the User Settings page in the ML workspace (right). Step 1 is highlighted with a red box around the 'Cloudera User' button. Step 3 is highlighted with a red box around the 'User Settings' tab. Step 4 is highlighted with a red box around the 'WORKLOAD_PASSWORD' field.
- Explore with SQL
- Create a Dashboard
- CML Notebook
- What's Next?

Below the steps, there are summary statistics:

- 1 QUERIES
- 17 DASHBOARDS
- 1 APPS
- 13 DATASETS
- 2 CONNECTIONS

At the bottom, there is a 'Recent Queries' section.

Home Page Views

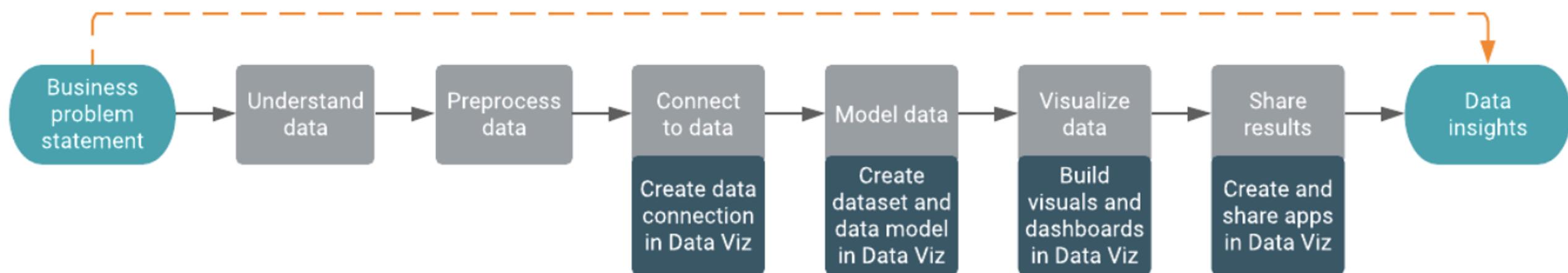
- The interface of Data Visualization has three views:

- HOME
- VISUALS
- DATA

The screenshot shows the Cloudera Data Visualization interface. At the top, there is a navigation bar with three tabs: HOME (highlighted with a red box), VISUALS, and DATA. A yellow callout bubble labeled "Views" points to the navigation bar. Below the navigation bar is a statistics banner with the following information: 16 DASHBOARDS, 1 APPS, 12 DATASETS, 0 QUERIES, and 0 TOTAL VIEWS. A yellow callout bubble points to this banner with the text: "Statistics banner shows the number of Dashboards, Apps, Datasets, Queries and Total Views that you can access". The main content area is divided into sections: "Last Viewed by You" and "Recently Created by You", each displaying a grid of visual preview cards. A yellow callout bubble points to this area with the text: "Visuals preview area provides quick access to the existing visuals and dashboards". On the right side, there is a sidebar with options like "NEW DASHBOARD", "NEW APP", and "LEARN". The "LEARN" section includes links to "Get Started", "What's New in 6.3.6", and "Documentation".

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



Data Warehouse Service

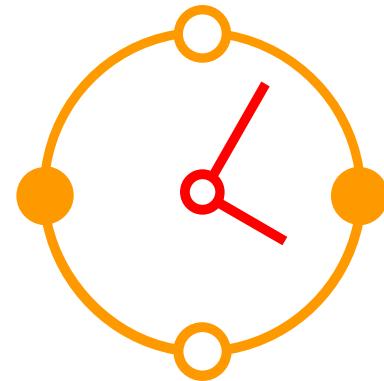
WHY CLOUDERA DATA WAREHOUSE?

Self-service analytics



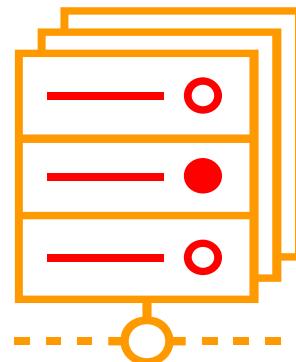
Businesses demand
data democracy

Faster Insights for
faster decisions



LOBs cannot wait

Access to modern data
use cases



Need data to build
competitive
advantage

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

The screenshot shows the Cloudera Data Warehouse (DWX) interface. The top navigation bar includes the DWX logo and the text "DWX Version - 1.0.0.0-501". On the left, a sidebar menu lists "Overview", "Database Catalogs", and "Virtual Warehouses", with a "More..." link. A vertical sidebar on the right shows "Environments | 6 More...".

The main "Overview" section displays two database catalogs:

- it-demo-3-new**: Running, warehouse-1566334661-pnts, It-demo-3. Metrics: DATABASES 2, MEMORY 16 GB, VIRTUAL WAREHOUSES 1.
- It-demo-3-default**: Running, warehouse-1566262402-glw8, It-demo-3. Metrics: DATABASES 2, MEMORY 16 GB, VIRTUAL WAREHOUSES 2.

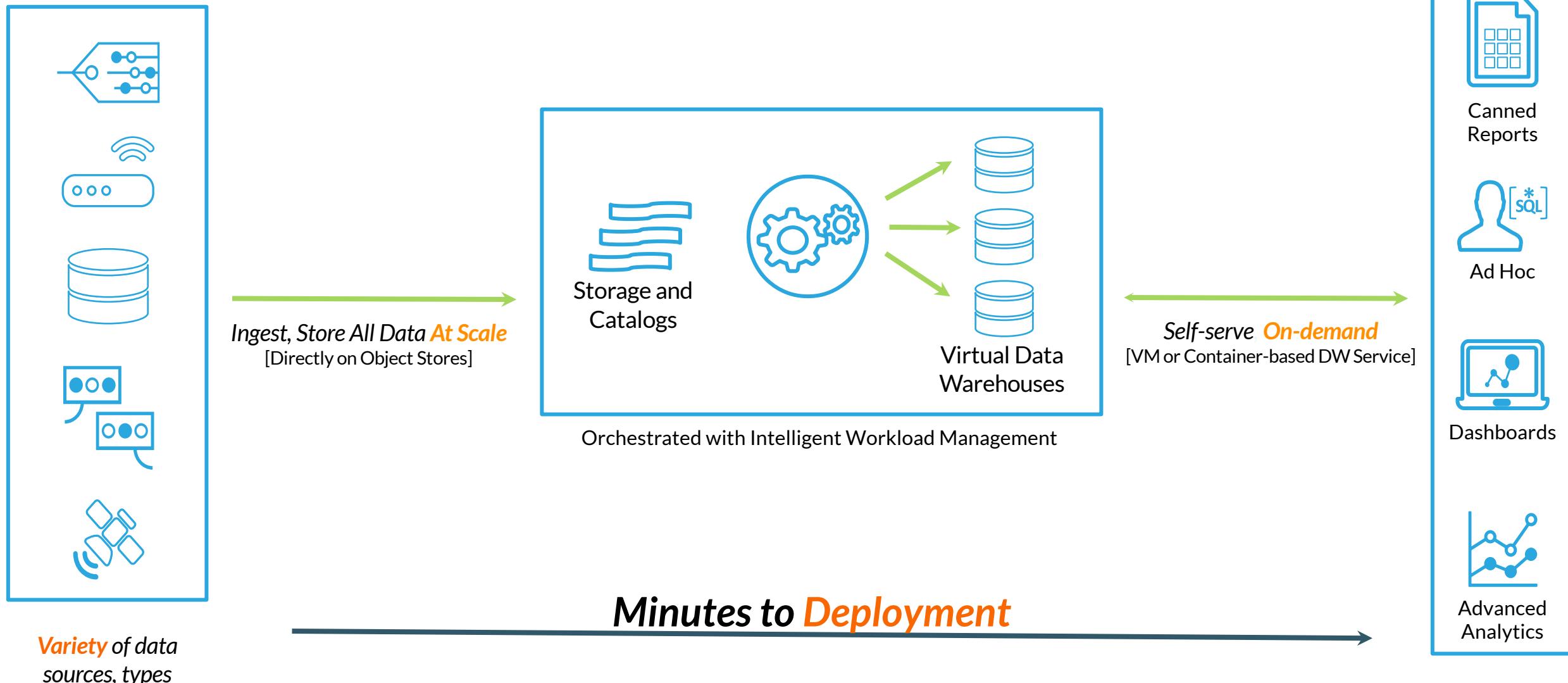
Below these are sections for "Virtual Warehouses | 3" and "Compute Environments | 2".

Name	Type	Status	Compute Environment
prasanth	HIVE	Stopped	compute-1566406600-nh5h
It-warehouse-new	HIVE	Stopped	compute-1566336367-dbk5
It-warehouse	HIVE	Stopped	compute-1566283788-pnks

Each row in the table provides metrics for Node Count, Total Cores, and Total Memory (40 GB), along with a "TYPE" column showing HIVE or COMPACT.

Elegant: Modern Data Warehouse

Guarantees faster time to value



Ease of Use

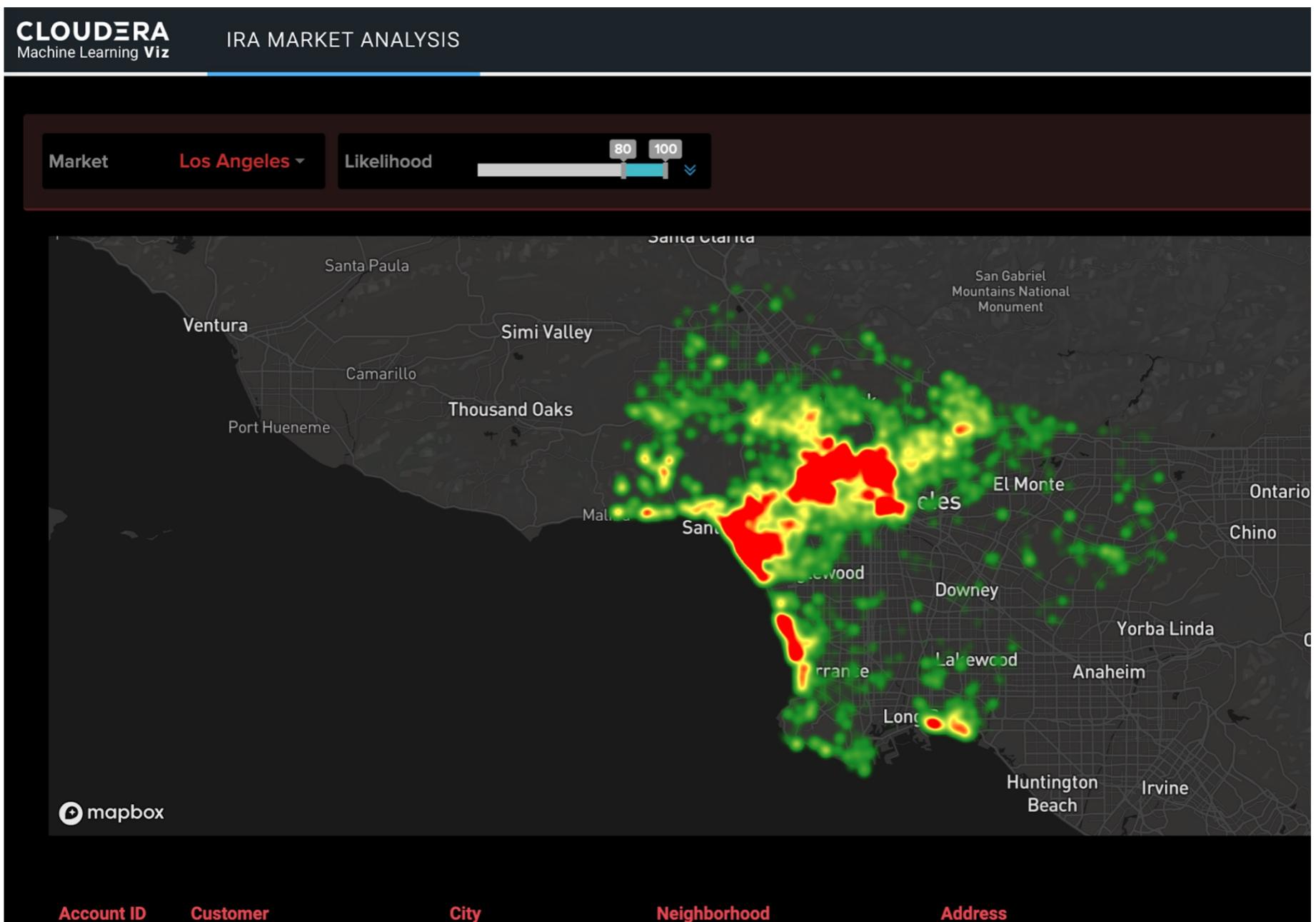
Zero to Query in 10 minutes or less

- **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

The screenshot shows the Cloudera Data Catalog interface. At the top left is the logo and title "CLOUDERA Data Catalog". Below it is a dark sidebar with a "Dashboard" button. The main area is titled "Data Catalog / Dashboard". It features a search bar with the placeholder "lt-demo-3" and a "Search" icon. A "Filters" section contains dropdown menus for "OWNER", "DATABASE", "TABLE TAG", and "CREATED WITHIN". The "OWNER" dropdown lists "hive" (53), "csso_eqbr" (25), "csso_eqbr@CLOUDERA.SITE" (5), and "impala" (1). The "CREATED WITHIN" dropdown has options for "Last 7 days" and "Last 15 days". To the right is a table listing six Hive Tables:

Table Name	Type	Path	Created	Owner	hive	
columns_v2	Hive Table	/sys	Created	Mon Aug 19 2019	Owner	hive
bucketing_cols	Hive Table	/sys	Created	Mon Aug 19 2019	Owner	hive
cds	Hive Table	/sys	Created	Mon Aug 19 2019	Owner	hive
part_col_privs	Hive Table	/sys	Created	Mon Aug 19 2019	Owner	hive
db_privs	Hive Table	/sys	Created	Mon Aug 19 2019	Owner	hive
global_privs	Hive Table	/sys	Created	Mon Aug 19 2019	Owner	hive