CLOUDERA

CDP Migration

CDP Data Services Workshops

Workshop	Date	Duration
CDP Migration Workshop	03-23-2022	2hrs
Cloudera Data Flow	03-31-2022	3hrs
Cloudera Data Engineering	04-07-2022	3hrs
Cloudera Data Warehouse	04-14-2022	3hrs
Cloudera Machine Learning	04-21-2022	3hrs

Quiz at the end of each workshop
Prizes to be won

NDA

The information in this document is proprietary to Cloudera. No part of this document may be reproduced or disclosed without the express prior written permission of Cloudera.

The information in this document is our currently intended developments and functionalities of Cloudera products which may change without notice at Cloudera's discretion. Cloudera makes no commitments about any future developments or functionality in any Cloudera product.

The development, release, and timing of release of any software features or functionality described in this document remains at the discretion of Cloudera and you should not rely on any statements about development plans or anticipated functionality in this document when making any purchasing decisions.

AGENDA

Cloudera at Glance

The path to CDP

- CDP Private Cloud
- CDP Public Cloud
- Tools to Expedite CDP Migration
- Additional Details on Migration
- Customer Story
- Cloudera PS & Training

Key Resources

CLOUDERA AT A GLANCE

CLOUDERA

THE ENTERPRISE DATA CLOUD COMPANY

Any Cloud

Data Lifecycle

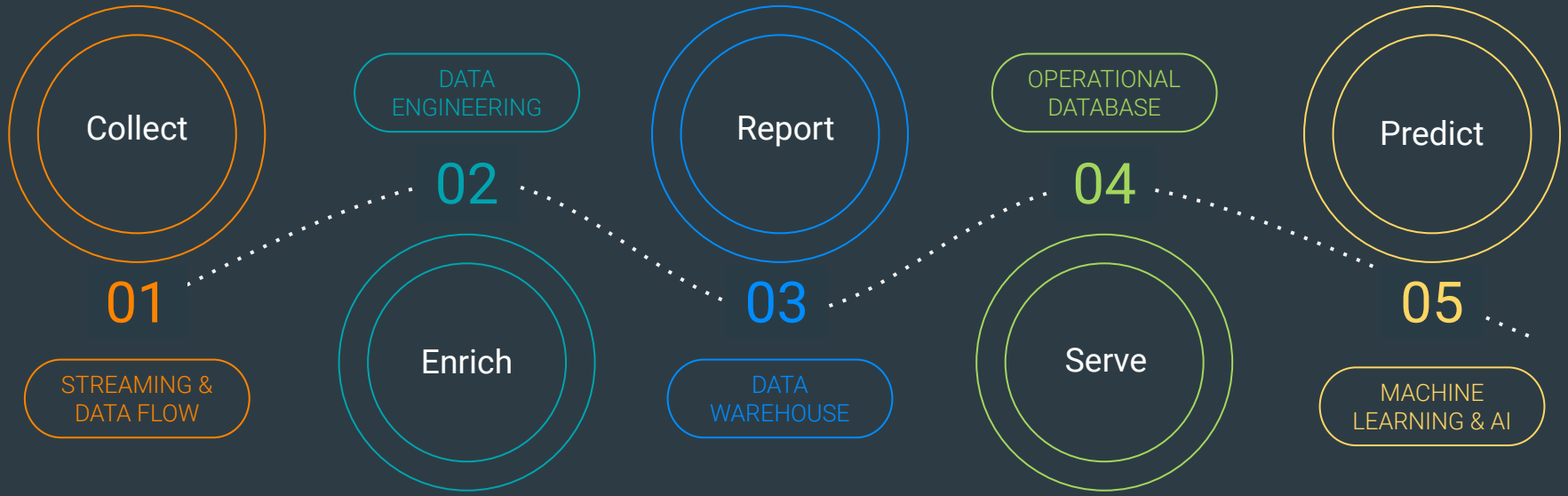
CLOUDERA SDX

Secure & Governed

Open

CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

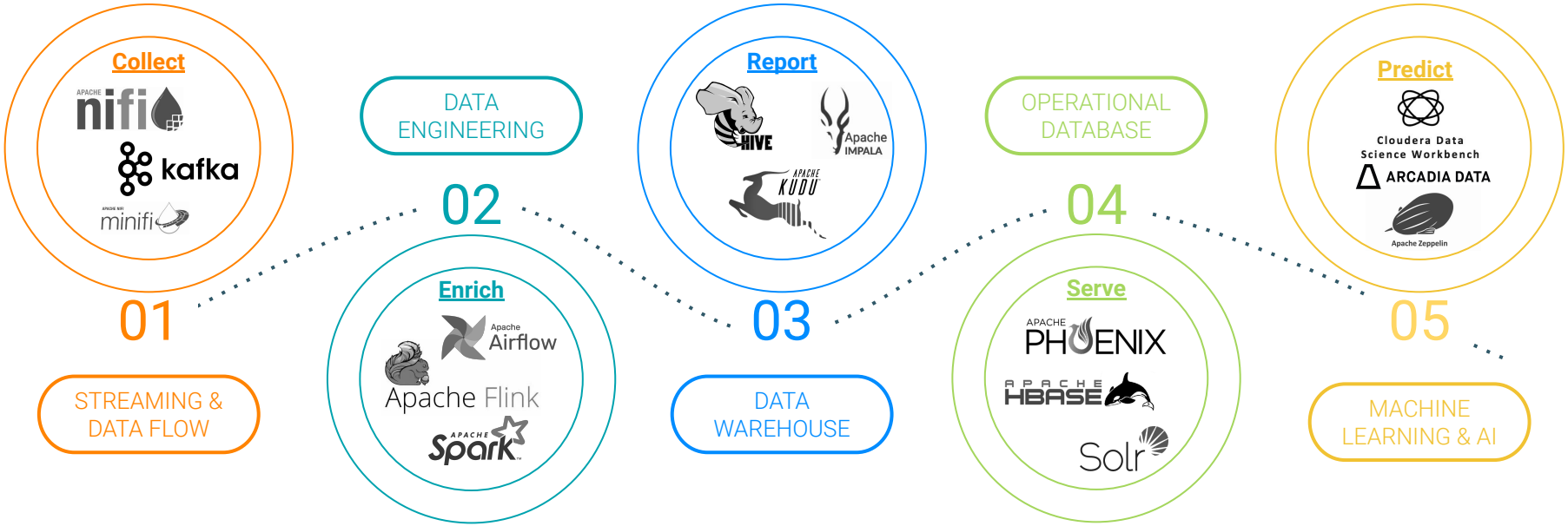
Manage and secure the data lifecycle in any cloud or datacenter



CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

What is the industry's best enterprise-grade blend of data management framework?

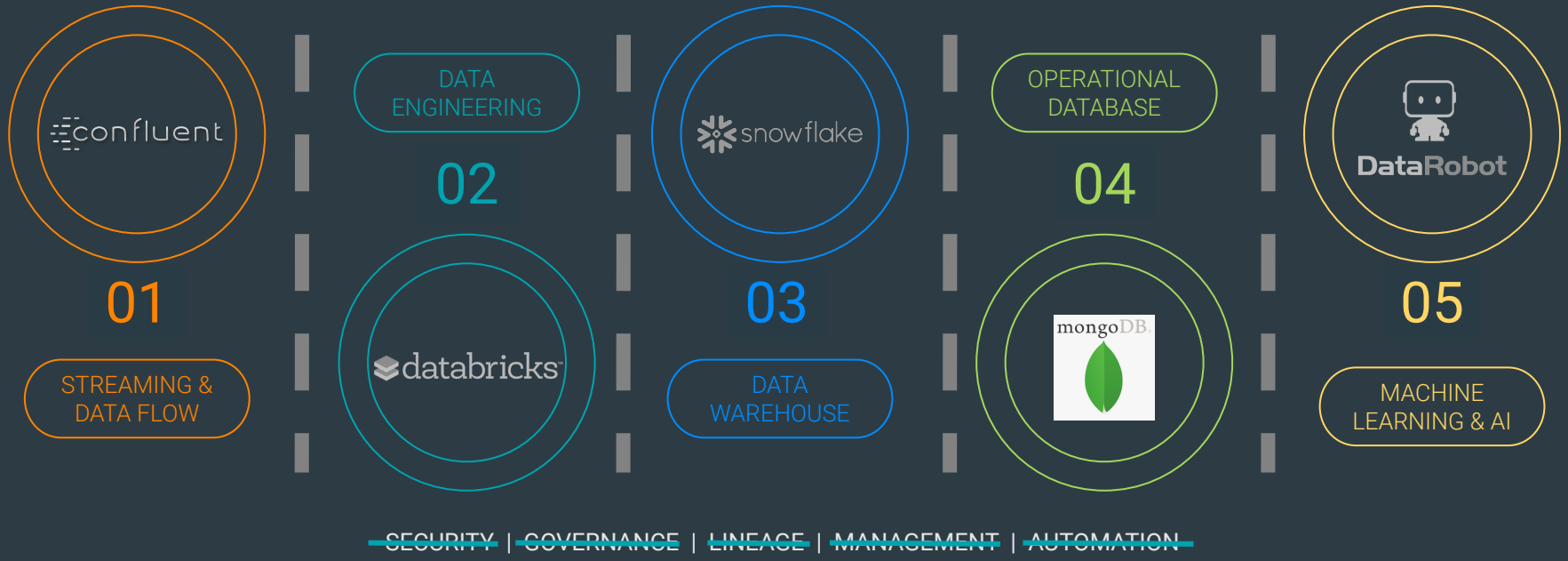


SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

Apache Ranger Apache Atlas

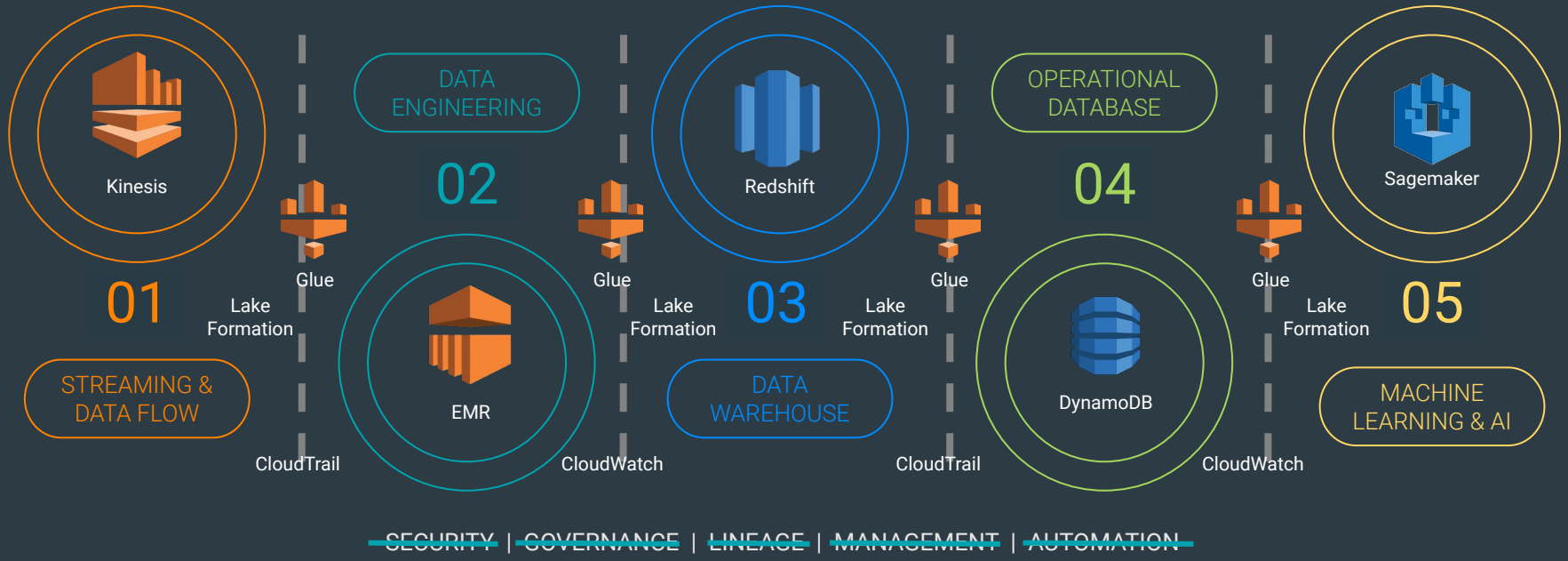
POINT SOLUTIONS HAVE AN INTEGRATION TAX

Security & governance is an afterthought



POINT SOLUTIONS FROM PUBLIC CLOUD PROVIDERS

Building blocks



HOW DO CUSTOMERS USE CLOUDERA?

Every business use case is a data lifecycle use case

	BANKING	TECHNOLOGY	TELCO	LIFE SCIENCES	MANUFACTURING
USE CASES	• Fraud detection • Anti-money laundering • Spend analytics	• Customer analytics • Threat detection • Predictive support	• Churn analysis • Customer care • Network optimization	• Patient care (IoT) • Genomics research • Regulatory compliance	• Predictive maintenance (IoT) • Supply chain optimization • Remote monitoring
KEY CUSTOMERS	• Barclays • Citi • Santander UK	• Cisco • Intel • Reef Technology	• Globe Telecom • Deutsche Telecom • Robi Axiata	• GlaxoSmithKline • Clearsense • Cerner	• Navistar • Micron • Sikorsky

INDUSTRY ANALYST RECOGNITION

Enterprise Data Cloud

Enterprise Data Platform

Ventana Research — *January 2021*

Cloud Data Ecosystems

Gartner — *January 2020*

Enterprise Intelligence Platforms

451 Research — *December 2019*

Cloudera Data Platform (CDP)

FEIBUSTECH — January 18, 2021

"...To realize the full potency of hybrid cloud, organizations really need a holistic approach to the entire data lifecycle. Before CDP, they had to assemble the pieces themselves — a costly, time-consuming undertaking with potential gotchas lurking at every turn..."

Ventana Research — January 22, 2021

"...CDP is an enterprise data platform built on open-source software...that offers key data analytics and artificial intelligence functionality. CDP can leverage all data types, including structured and unstructured data, relational data and streaming data from any point in the data lifecycle...."

ENTERPRISE DATA CLOUD

ENTERPRISES ARE EMBRACING PRIVATE CLOUD

IDC Research - Cloud Growth, Migration, and Repatriation Continue to Gain Momentum

67%

Of enterprise workloads run on public and private cloud implementations

84%

Of enterprises report repatriating some workloads from public cloud

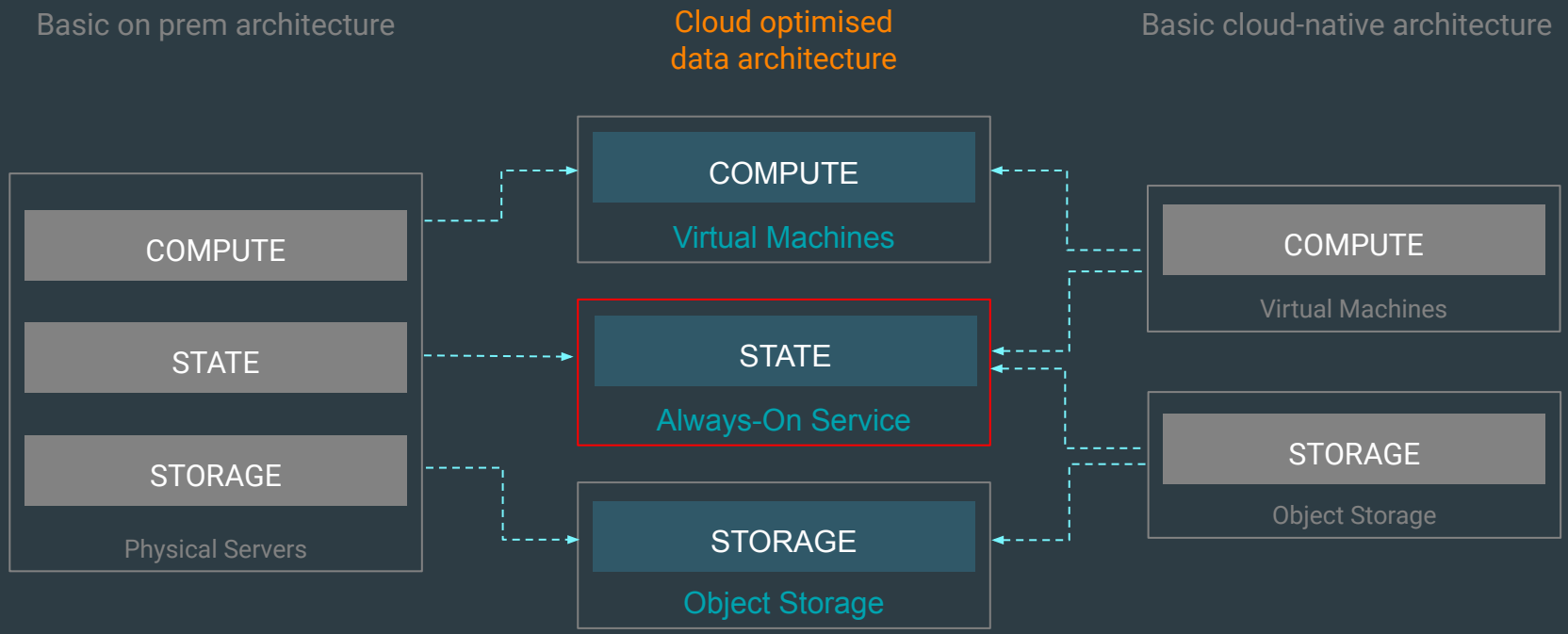
52%

Of repatriated workloads move to private clouds

"As enterprise customers gain cloud expertise they're placing investments in private cloud solutions for increased security, compliance, performance, control and cost savings. Private clouds often act as a stepping-stone in the hybrid cloud journey."

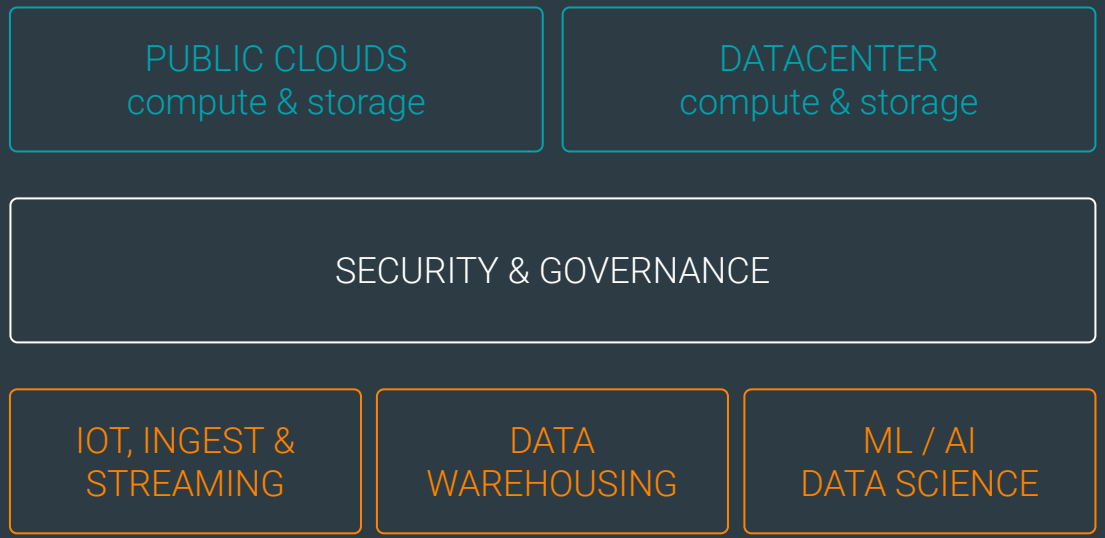
IDC, Cloud Growth, Migration, and Repatriation Continue to Gain Momentum, Michelle Bailey, Chris Kanthan, March 2020
IDC, Cloud Pulse 1Q20 Survey Findings, Doc # US46396720, May 2020

SEPARATE STORAGE AND COMPUTE *AND STATE*



ENTERPRISE DATA CLOUD DESIGN PRINCIPLES

- Hybrid and multi-cloud
- Secure and governed
- Multi-function analytics
- Open platform

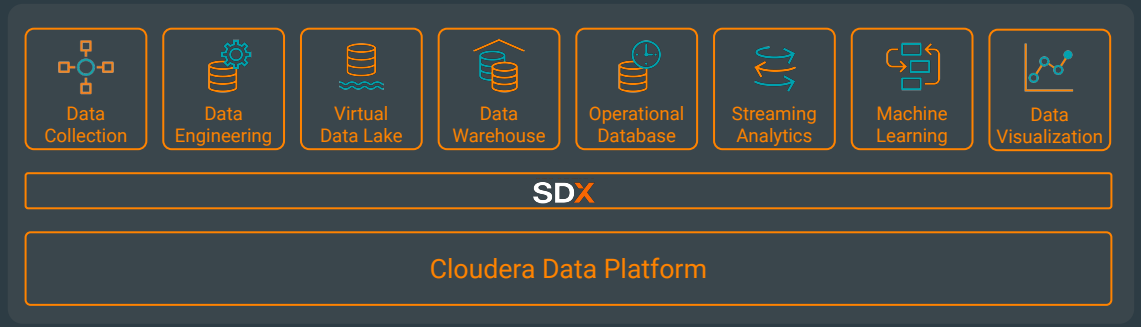


CLOUDERA DATA PLATFORM

A HYBRID / MULTI-CLOUD DATA PLATFORM *AND* AN INTEGRATED SUITE OF SECURE ANALYTIC APPS

Real-time
Batch
Structured
Unstructured

Data Sources



Analysts
Engineers
Scientists
Developers

Data Users

Data Lifecycle
integration for better user productivity and faster time to value

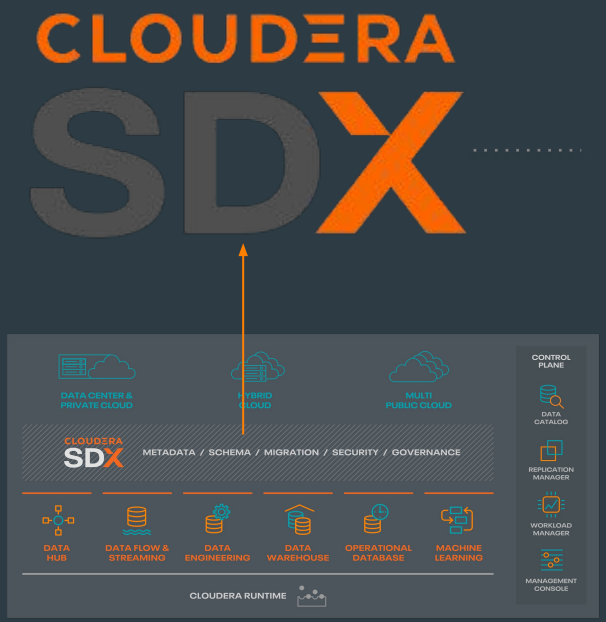
Hybrid & Multi-Cloud
to leverage existing investments and reduce risk

Secure & Governed
to simplify data protection, sharing and compliance

Open & Extensible
to support more use cases faster and at lower cost

CONSISTENT SECURITY AND GOVERNANCE

Built for multi-functional analytics anywhere



Data Catalog: a comprehensive catalog of all data sets, spanning on-premises, cloud object stores, structured, unstructured, and semi-structured

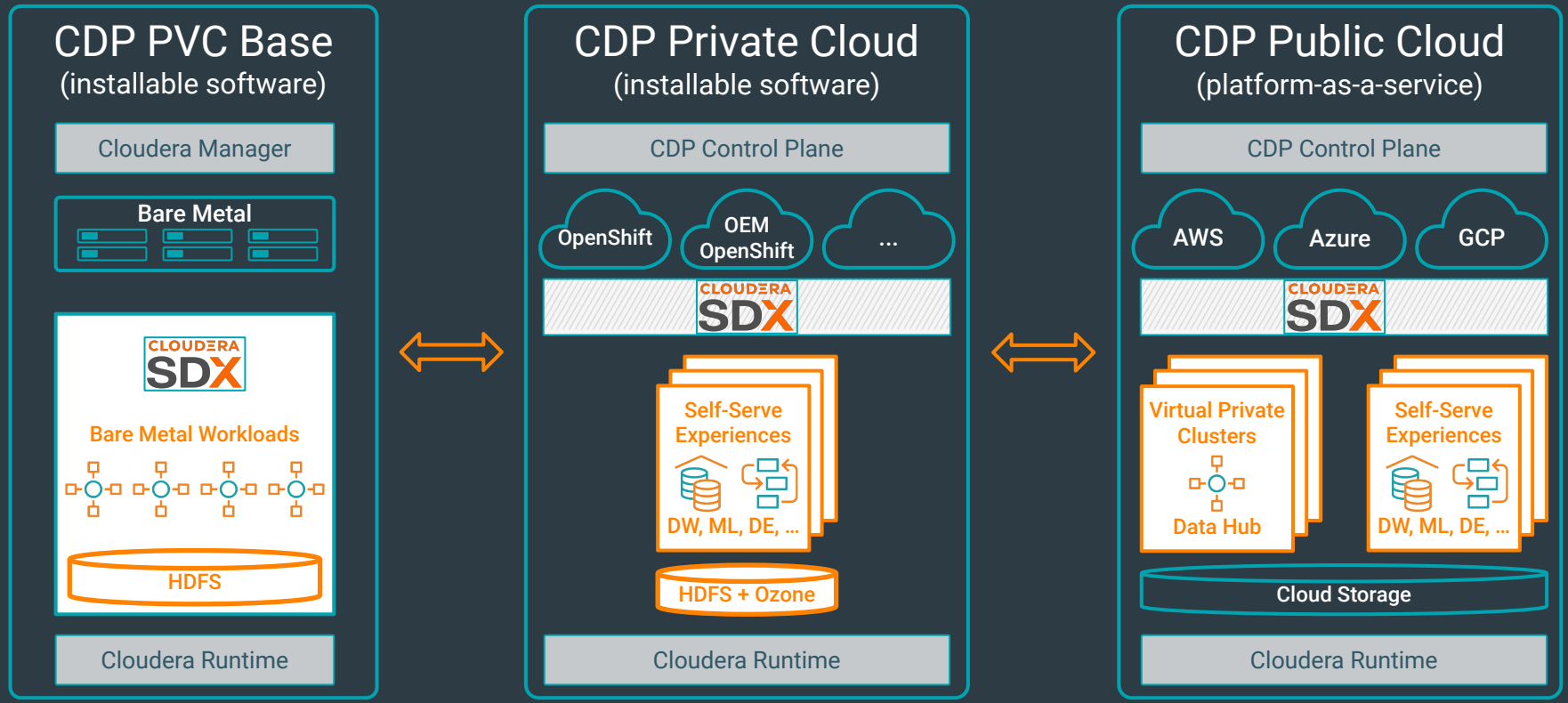
Schema: automatic capture and storage of any and all schema and metadata definitions as they are used and created by platform workloads

Security: role-based access control applied consistently across the platform. Includes full stack encryption and key management

Governance: enterprise-grade auditing, lineage, and governance capabilities applied across the platform with rich extensibility for partner integrations

Replication: deliver data as well as data policies there where the enterprise needs to work, with complete consistency and security

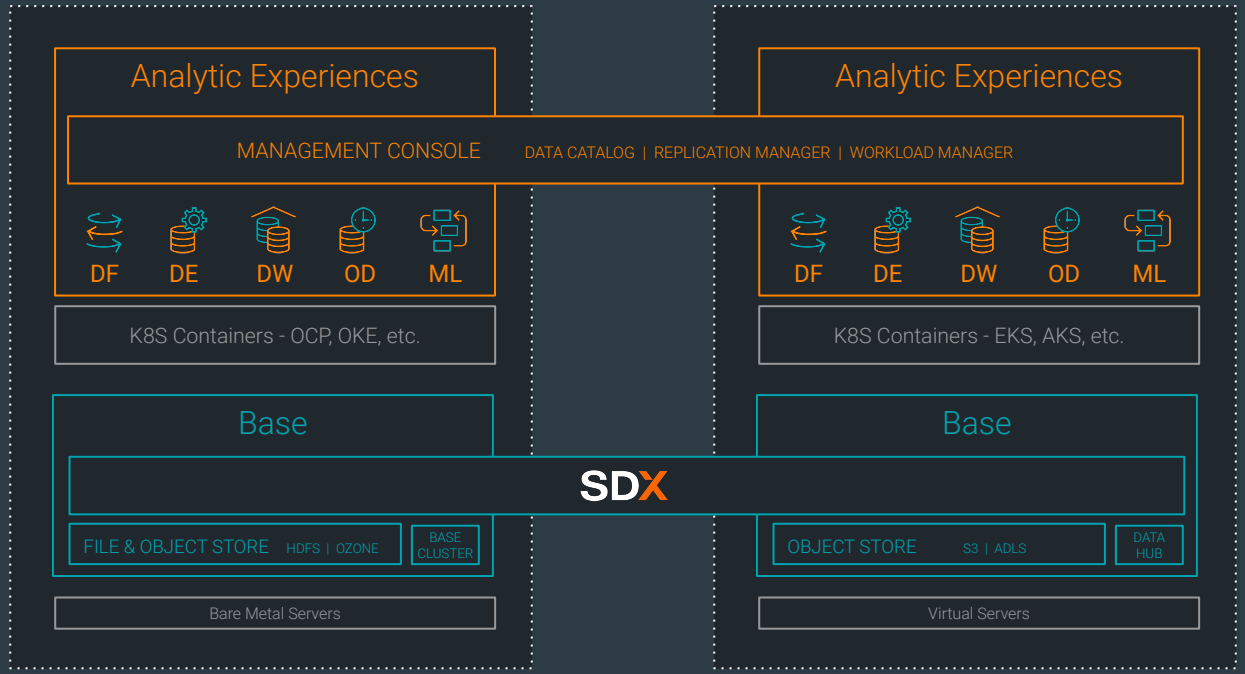
CDP - ONE PLATFORM, THREE FORM FACTORS



CDP HYBRID CLOUD

Consistent operations and analytics experiences across private and public clouds

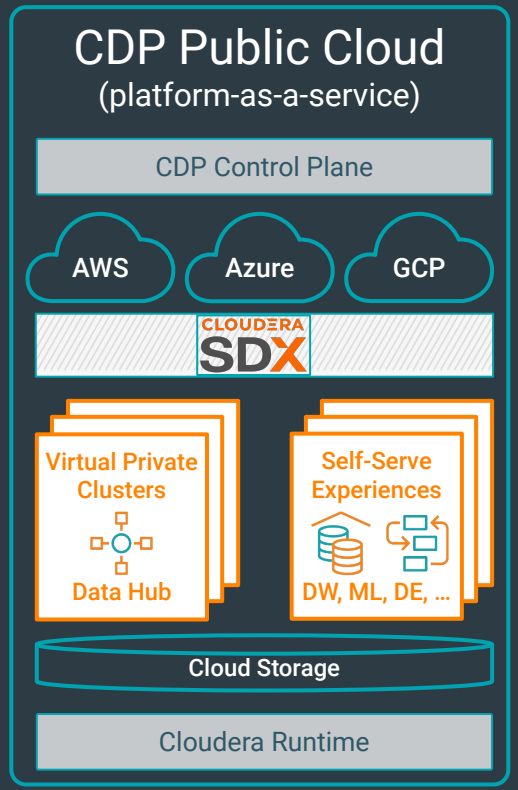
CDP Private Cloud



CDP Public Cloud

CLOUDERA DATA PLATFORM – PUBLIC CLOUD

- Available on AWS, Azure & GCP
- VM-based Data Lake and Data Hub clusters
- Containerized workloads:
  - Cloudera Data Warehouse (CDW)
  - Cloudera Machine Learning (CML)
  - Cloudera Data Engineering (CDE)
  - Cloudera Operational DB (COD)
  - Cloudera Data Flow
- Unlike other public cloud services, your data will always remain under your control in your VPC
- Control cloud costs by automatically spinning up workloads when needed and suspending their operation when complete



CDP PUBLIC CLOUD | UNIQUE CAPABILITIES

Self-Service Analytics

- Data warehouse
- Machine learning
- Data hub
- Flow management
- Shared data experience

Intelligent Migration

- Improve cluster utilization with highly variable jobs
- Deliver optimal capacity to meet workload SLAs
- Improve cost efficiency by freeing on-prem resources for more predictable workloads

Adaptive scaling

- Adjust capacity up or down to optimize workload performance automatically
- Eliminate the need to size workload requirements that can't be reliably predicted
- Speed up deployment while effectively managing costs

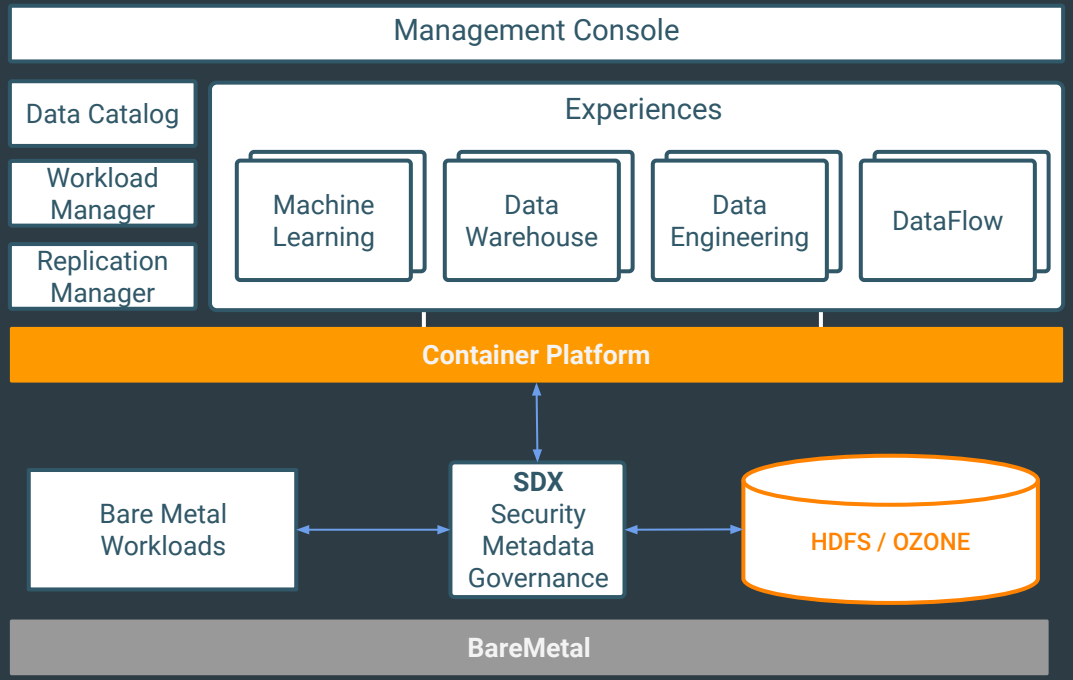
Burst to Cloud

- Easily and quickly move workloads, data, metadata, policies, etc.
- Provide the “right” amount of cloud capacity to meet SLAs
- Isolate “noisy neighbors”

CLOUDERA DATA PLATFORM - PRIVATE CLOUD

A new product offering, CDP Private Cloud provides the ability to:

- Extend compute capacity from today's VM/Bare-metal based CM/CDH deployments onto Kubernetes infrastructure
- Leverage Cloudera workloads (ML, Spark, Impala, Hive etc.) that they already leverage in CDP Public Cloud (AWS and Azure) on-premises.



WHY CDP PRIVATE CLOUD?

1. Workload Isolation

No Noisy Neighbours
Dedicated compute per tenant

Independent Upgrades
Upgrade when needed

Modern Standards
Container-based multi-tenancy

2. Simplified Onboarding

Push-button Provisioning
Up and running in seconds

Redesigned User Interfaces
Use-case optimised workflows

3. Better Infrastructure Utilisation

Auto-scale, Auto-suspend
Use what you need, when you need it

Shared Kubernetes
All experiences on a single platform

Quota Management
Set mins and max per tenants

DATA HUB CLUSTERS AND DATA SERVICES

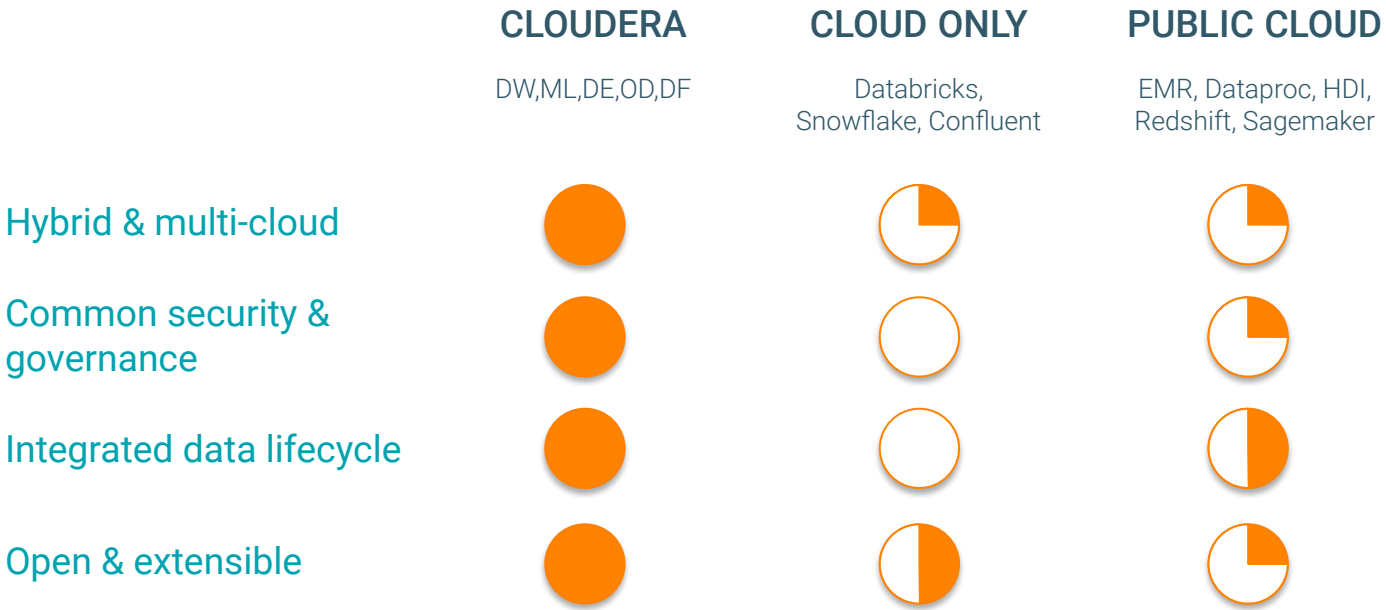
What are the consumption options?

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

An **Experience** is a container-based compute environment for specific purposes: ML, DW, DE, OD, DF

COMPETITIVE LANDSCAPE

Cloudera stands out as a hybrid and multi cloud platform for data management and analytics with enterprise security and governance powered by open-source innovation



CASE STUDY: FINANCIAL SERVICES CUSTOMER

Customer saves $10M

Before CDP

Amazon EMR

$300K **$12-16M**

After CDP

Amazon EC2 Amazon S3

$850K **$3M**

Moving Customers to CDP - What's in it for you?

It's not just an upgrade - it's about re-architecting the data infrastructure to be truly Hybrid / Cloud-Native

Great opportunity for both high-value consulting as well as project delivery services

Delivers a cloud-native platform for the client that can handle many varied data workloads

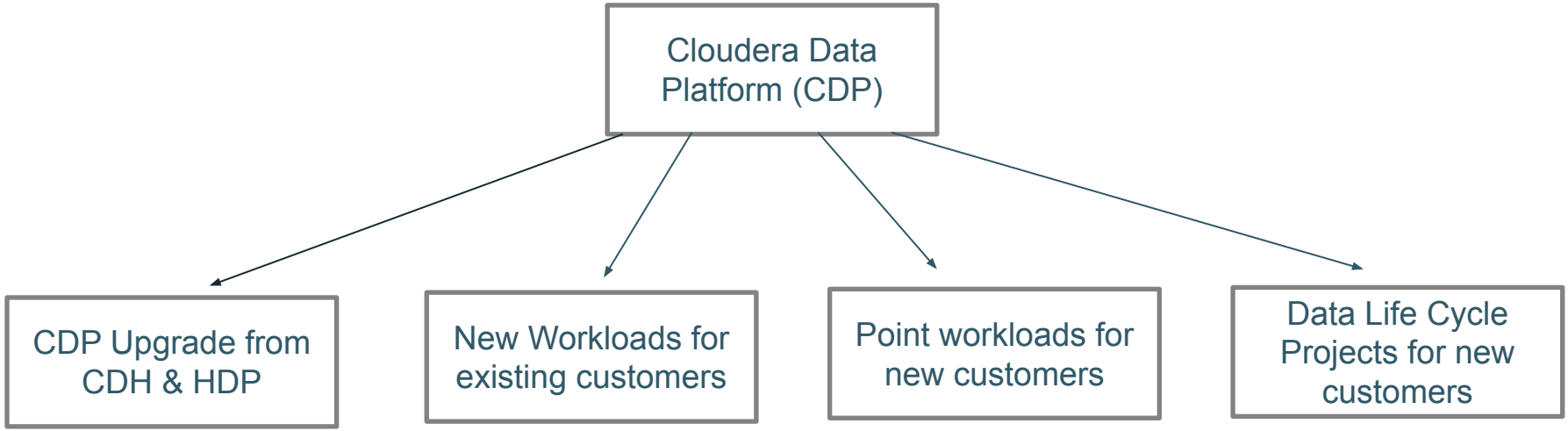
Deeper strategic client engagements with longer, multiple cycles

Deliver both cloud strategy and data strategy without compromise

Lower-risk integration of multiple data workloads in a single platform

Take advantage of AWS/Azure funding

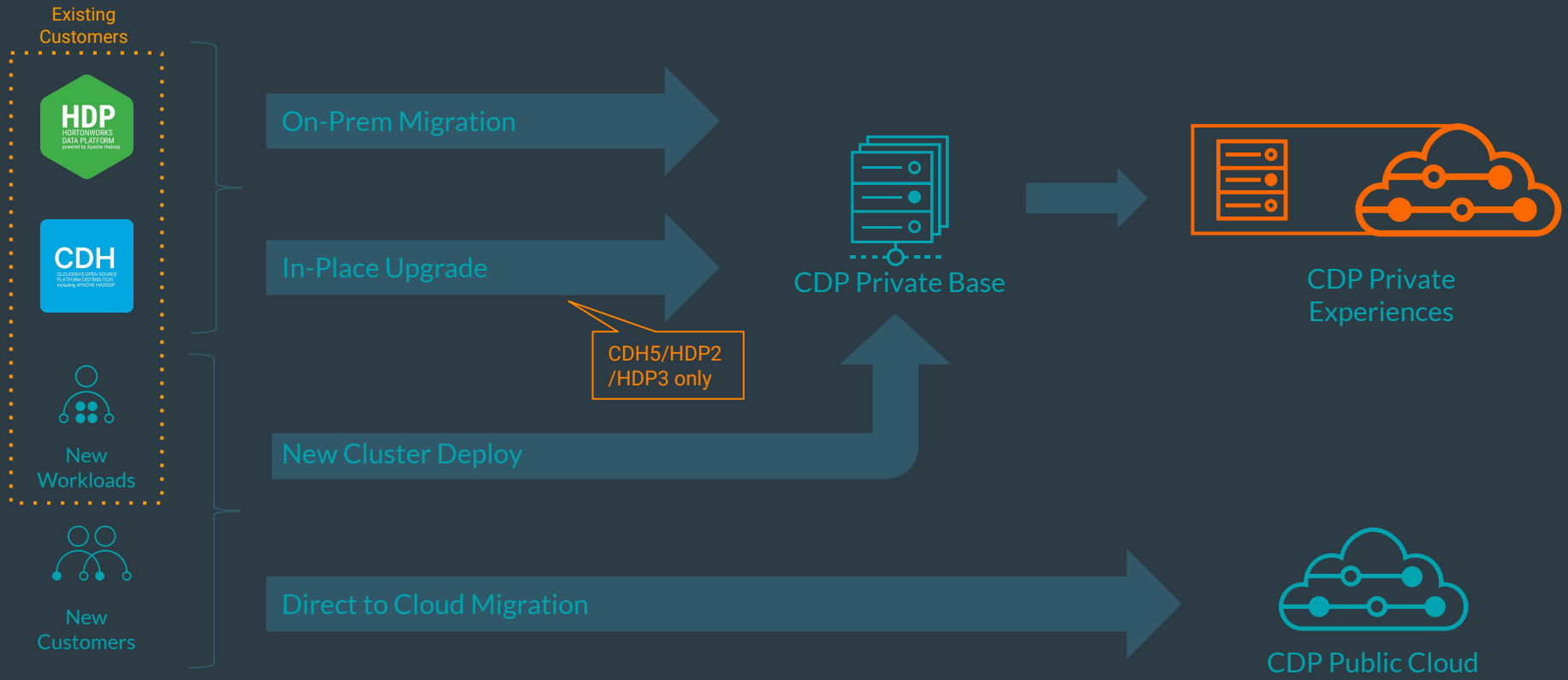
Opportunities ahead of us



Red Hat → 3-way partnership – You, Cloudera and Redhat – winning Private Data Cloud Business

aws Azure Google Cloud → 3-way partnership – You, Cloudera and Cloud Infrastructure Vendors (AWS, Azure &GCP) – Winning Public Data Cloud business

THE PATH



MIGRATION PROCESS OVERVIEW

Scoping and Planning

Workshops for identification of risk areas, migration strategy (in place, side-car, target version).
Planning for next steps

Dev CDP Cluster Installation

Perform installation of new cluster or upgrade of existing dev environment for dev and test purposes.

App Dev and Test

Tenant(s) work to test against newer versions and perform code modifications if required.

UAT / Pre-Prod Migration

Development and execution of a real-life upgrade/migration.

Used to rehearse and finalise cut-over activities and backup/restore strategy

Prod migration and go-live, DevOps integration readiness

Live Migration

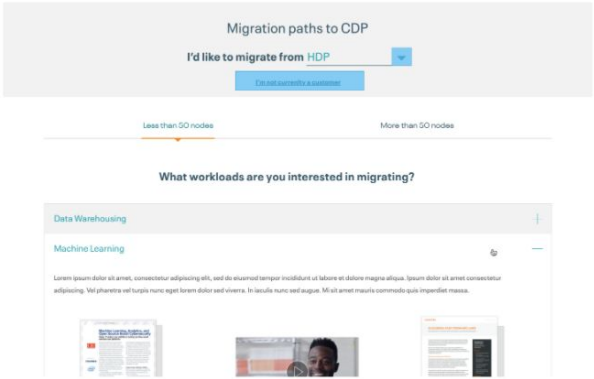
Installation of production cluster

Where doing side-car migrations, cut over of tenants or workloads on an individual basis.

UPGRADE ADVISOR & MYCLOUDERA UPGRADE ADVISOR

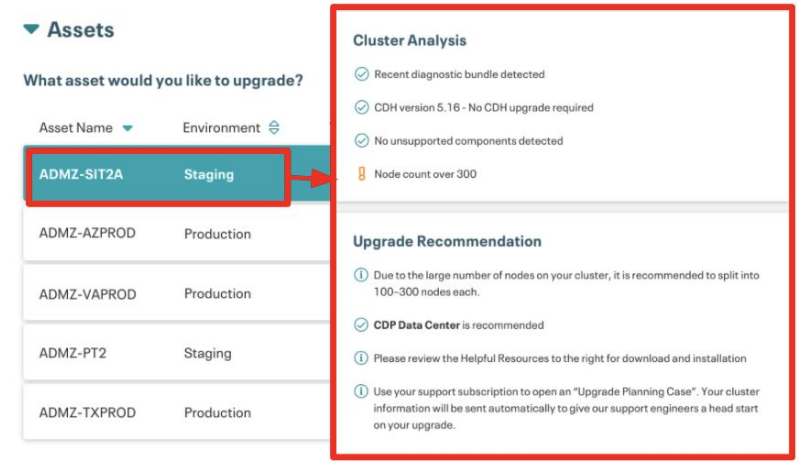
Early mockups

CDP Journey Advisor



- 3 questions
- Provides targeted collateral (1-5 pages)
  - What's New
  - What's Changed

MyCloudera.com CDP Journey Advisor

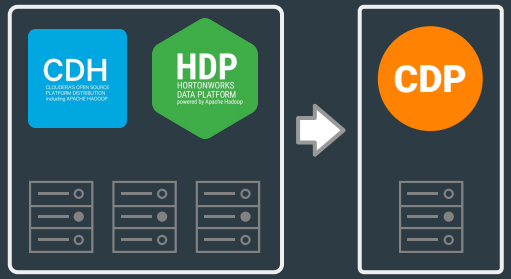


- Provides cluster specific guidance
- Warns for deprecated components
- Provides upgrade prerequisites
- Provides semi-customized PS & Training proposals
- Provides account team contact info

Migration to CDP Private Cloud

PATHS TO CDP PRIVATE CLOUD BASE

Migrate to Private Base



Build a new CDP Private Cloud Base cluster on-premises; copy data and metadata from existing classic cluster; and migrate existing workloads.

Use this when:

- **Customer has new hardware/spare capacity**
- **Can afford higher overall transition time**
- **Requires more granular cutover/failback capabilities**

Upgrade to Private Base

Upgrade from classic cluster to CDP Private Cloud Base in-place on the same hardware infrastructure.

Use this when:

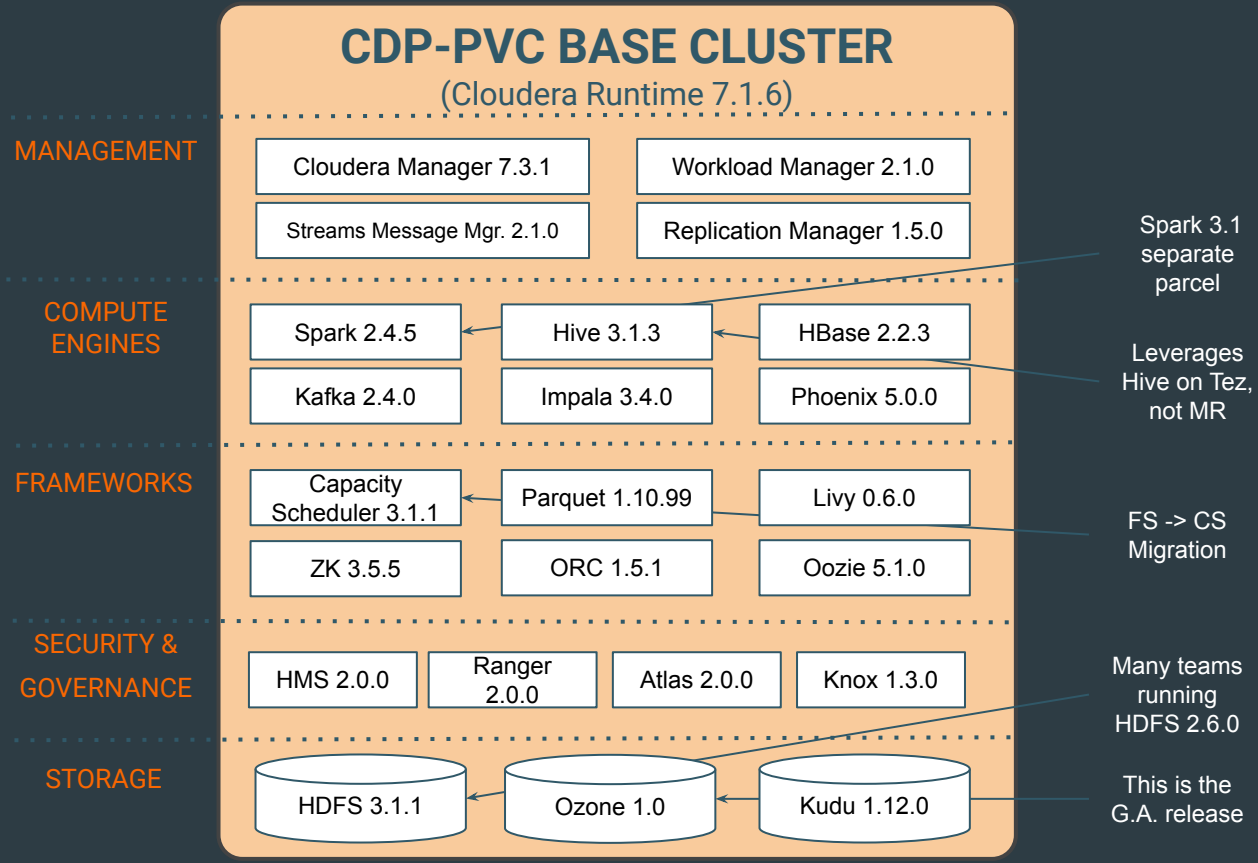
- **Customer has limited spare capacity**
- **Has fewer tenants / integrations**
- **Can afford a single-cutover with limited rollback options**

CDP RUNTIME 7.1

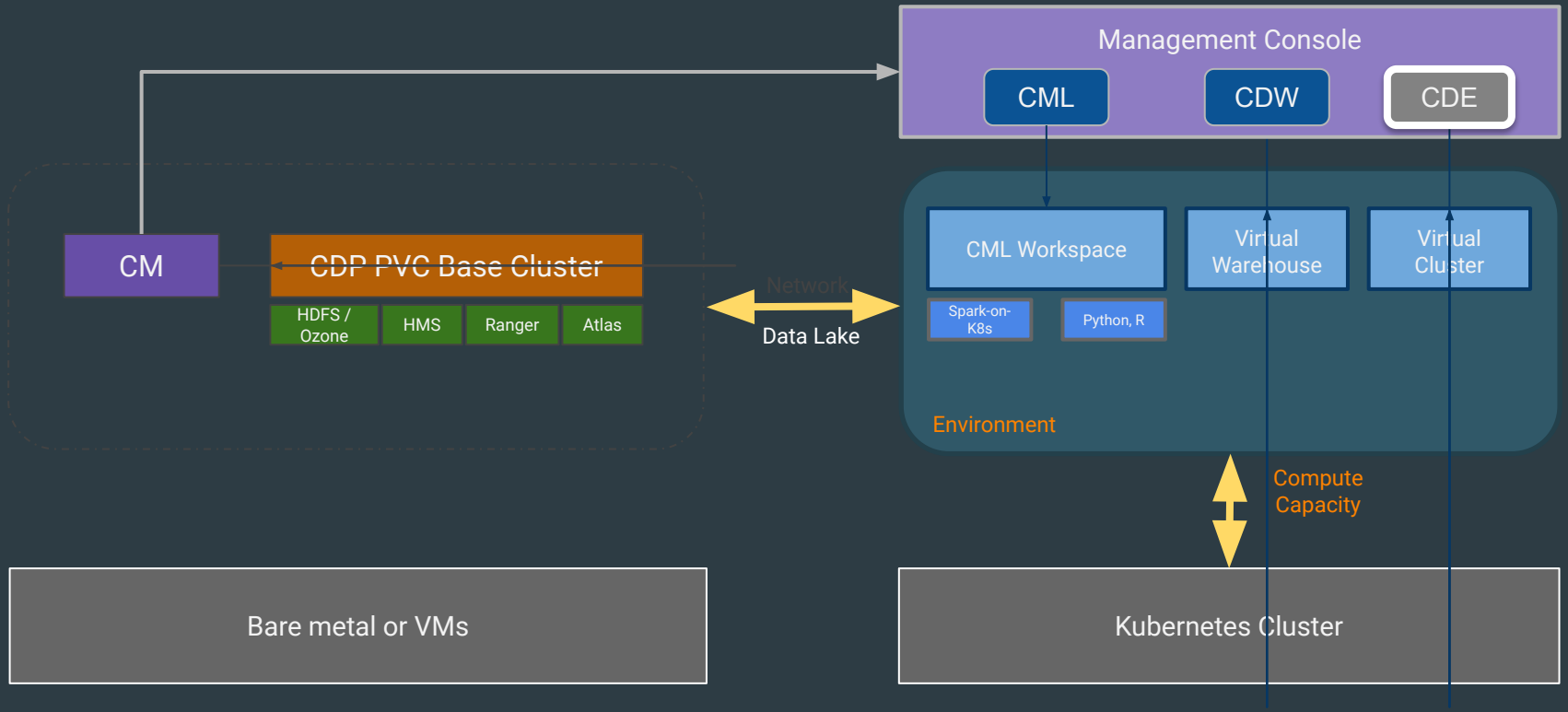
A CDP Private Cloud Base cluster (powered by Cloudera Runtime), can serve as a traditional "data lake" (storage & compute) cluster, or as a "base storage cluster" (storage only) serving compute workloads running on Kubernetes.

This image shows the component versions in Cloudera Runtime 7.1.6.

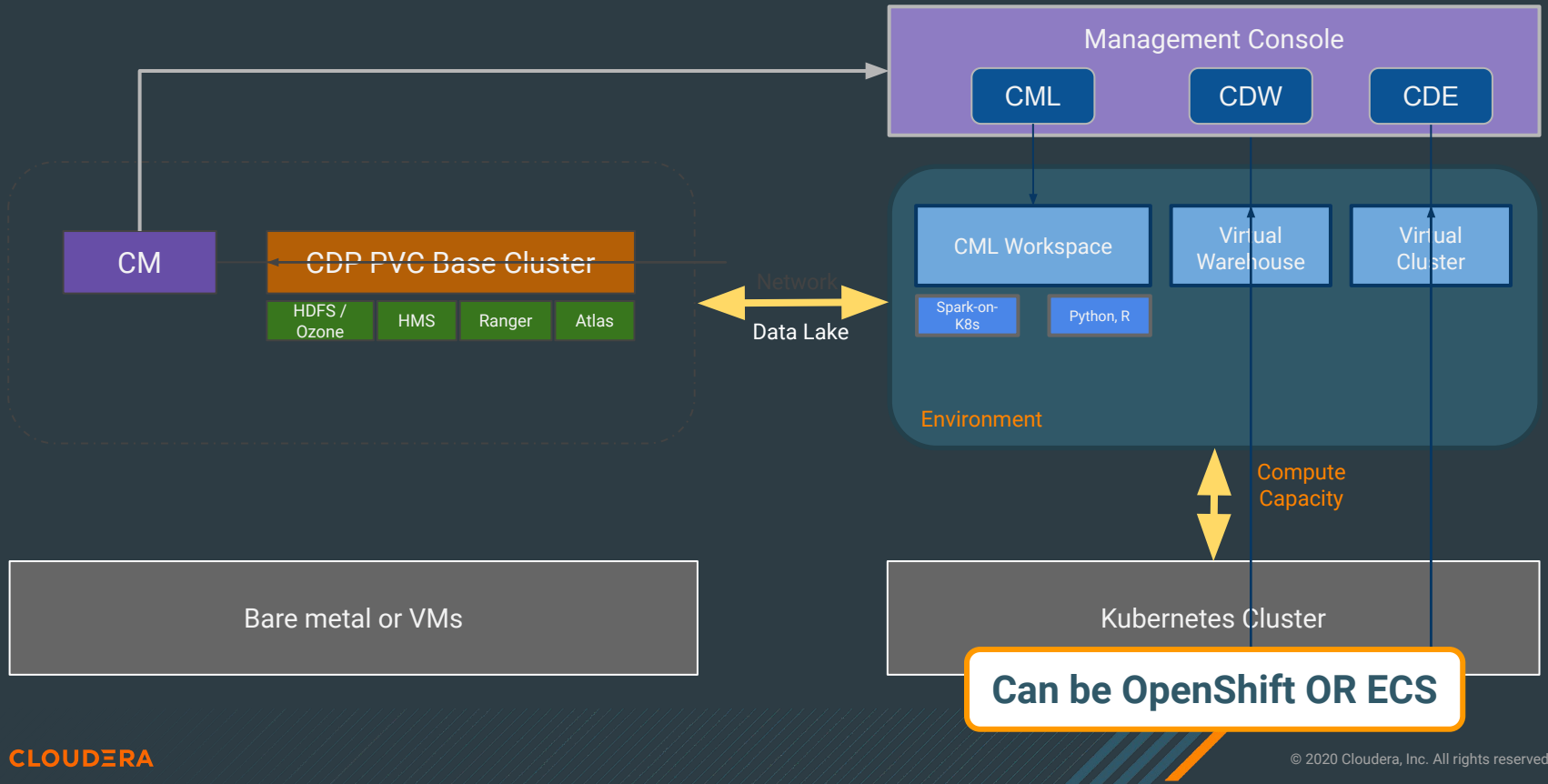
Click HERE for the complete list of supported components.



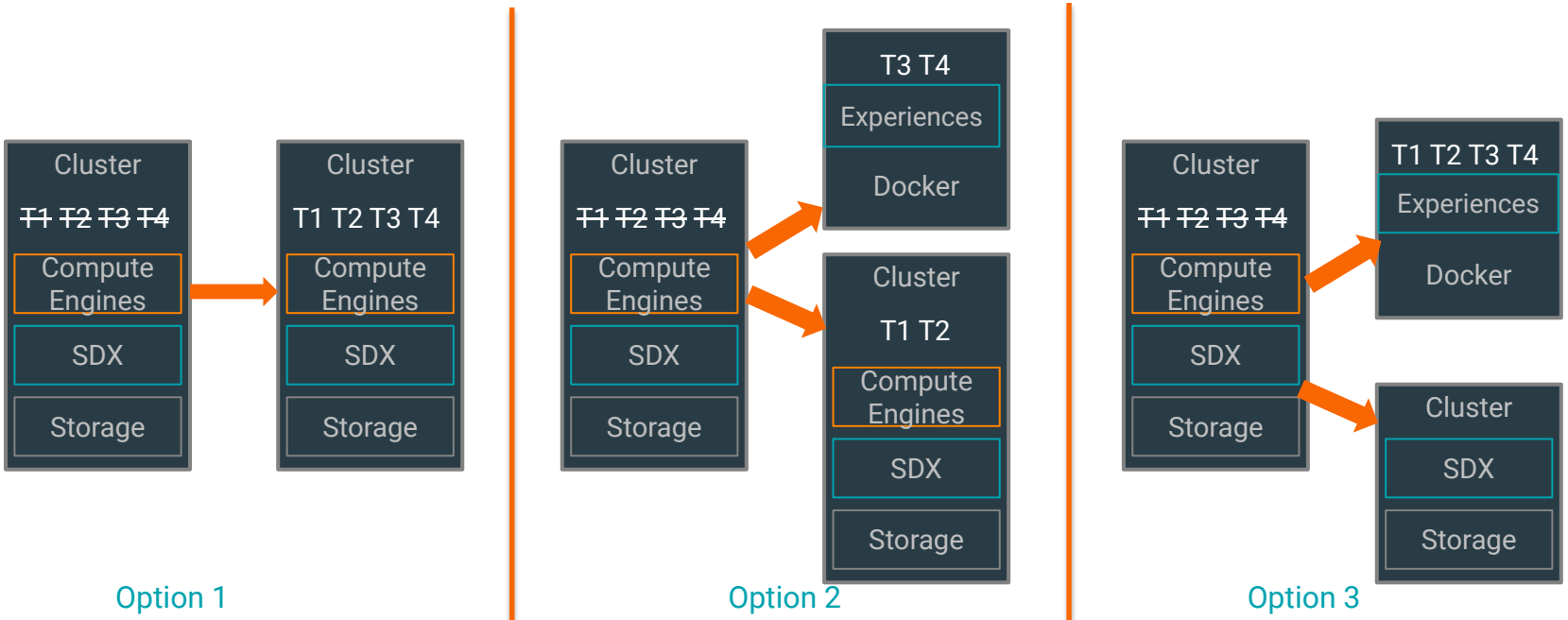
CDP PRIVATE CLOUD DATA SERVICES ARCHITECTURE



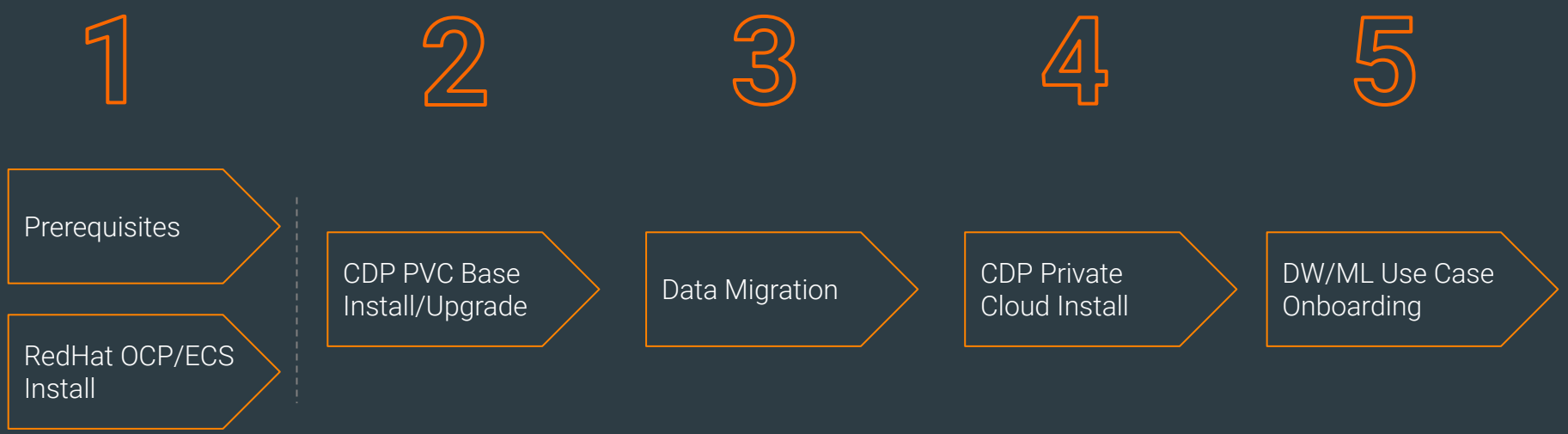
CDP PRIVATE CLOUD 1.3 DATA SERVICES ARCHITECTURE



WORKLOAD MANAGEMENT IS MORE FLEXIBLE AFTER UPGRADING TO CDP



PATH TO CDP PRIVATE CLOUD BASE AND EXPERIENCES



Migration to CDP Public Cloud

CDP Public Cloud – Key Benefits

There are a number of common benefits that CDP delivers

Simplify Data Analytics

You Own Your Data

First Class Security

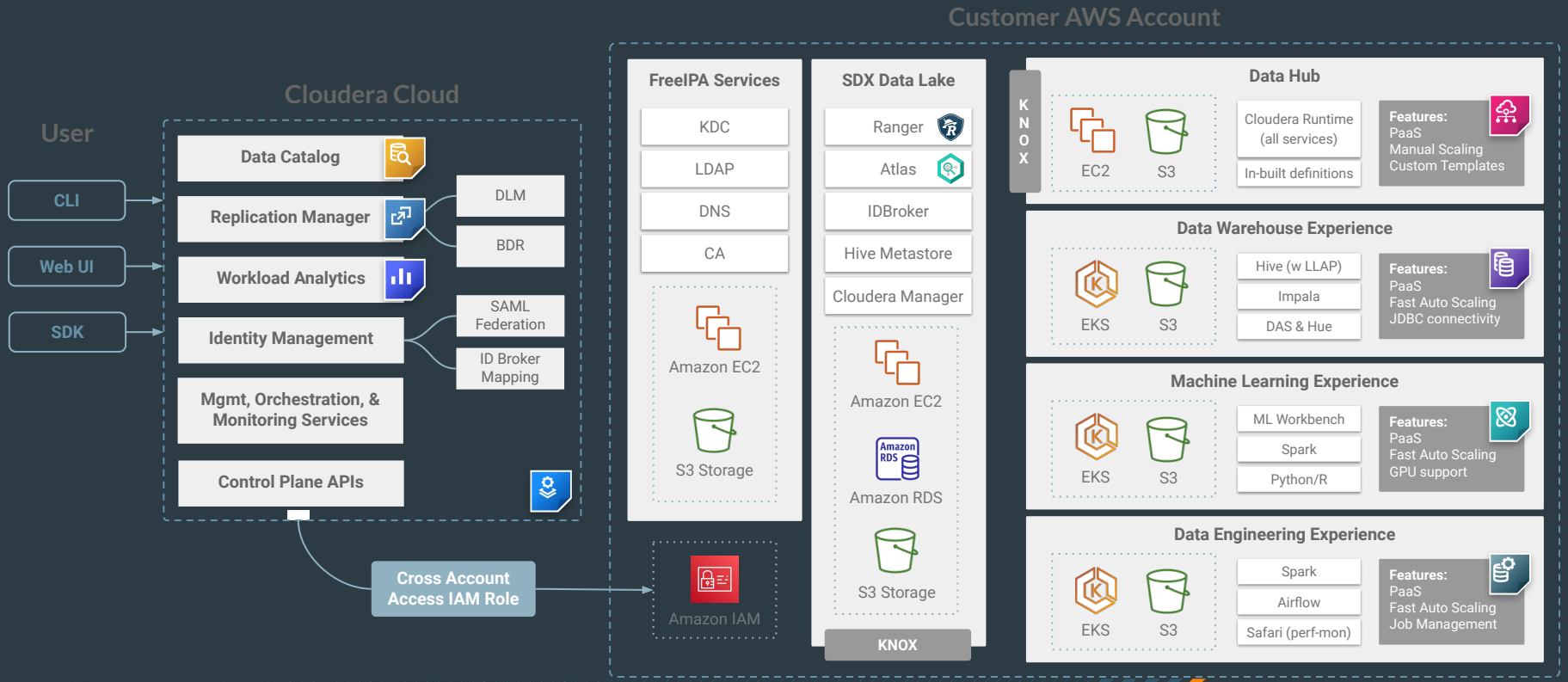
Hybrid Flexibility

Common Skill Set

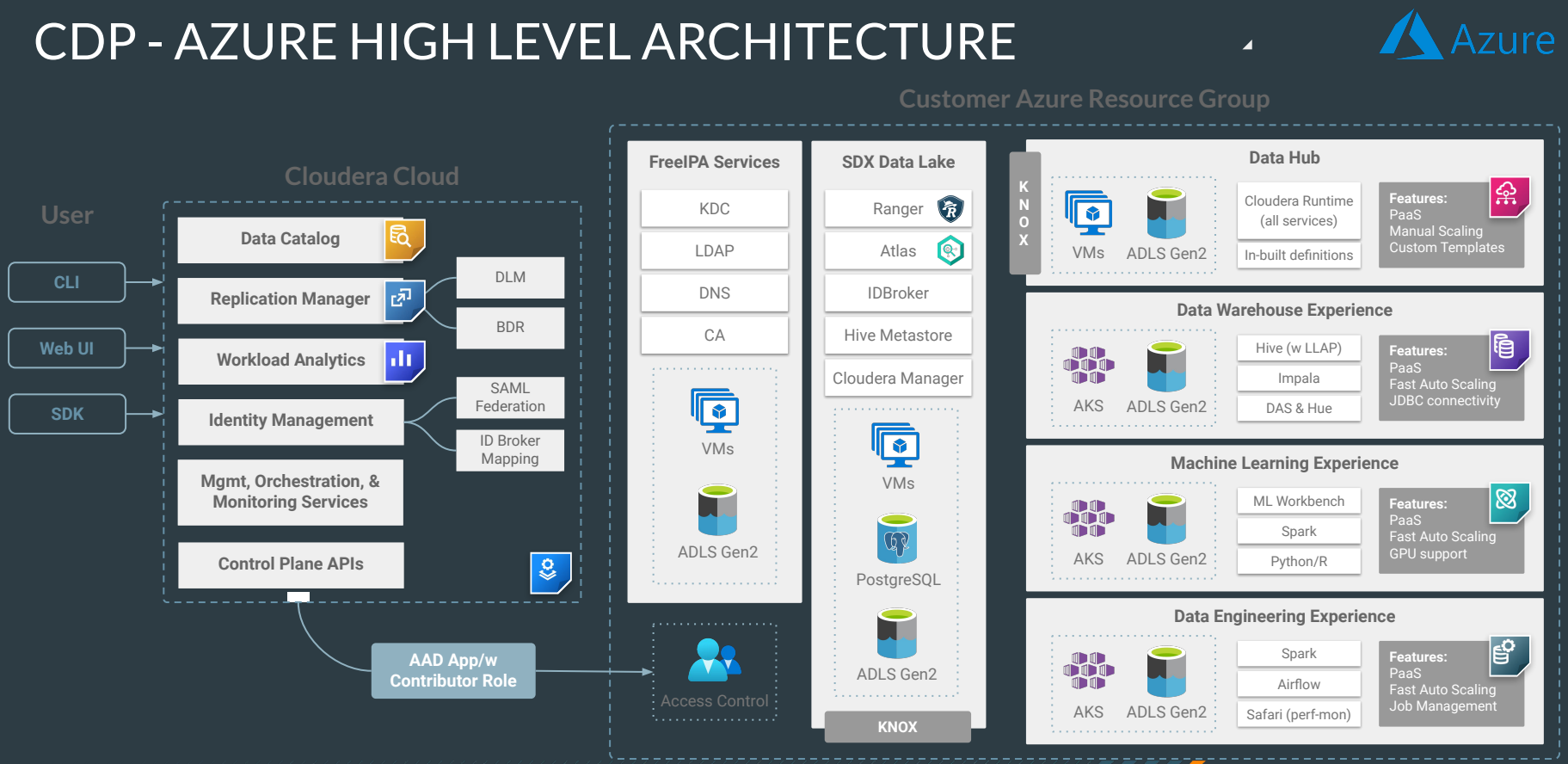
Data Lifecycle

Easy and Portable

CDP - AWS HIGH LEVEL ARCHITECTURE



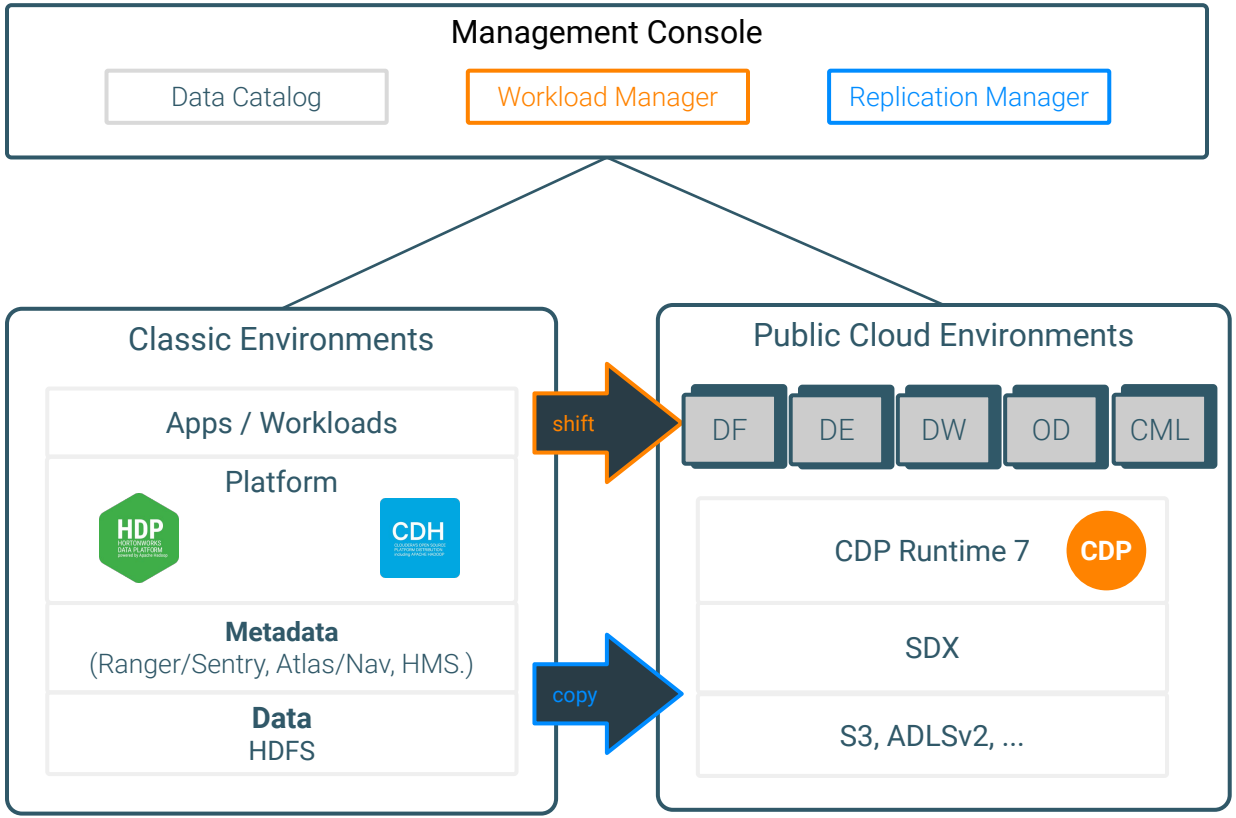
CDP - AZURE HIGH LEVEL ARCHITECTURE



MIGRATE TO PUBLIC CLOUD

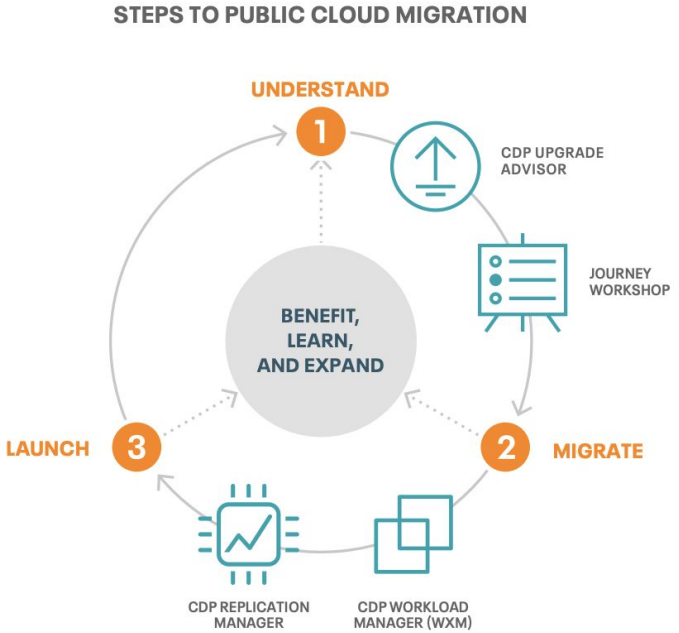
Process

- Set up public cloud environments
- Register classic cluster(s)
- Identify candidate workloads
- Migrate workload data and metadata using Replication Manager (“Burst to Cloud”)
- Test and promote to production



CDP Public Cloud Migration

- **Step 1—Evaluate Clusters and Workloads:** CDP Upgrade Advisor is a self-service tool that evaluates your legacy assets to recommend upgrade paths, highlight risks, and explain why. Armed with that knowledge, the Journey Workshops provide a constructive expert forum to carefully consider optimal placement for both existing and new workloads.
- **Step 2—Enable Migration Activities:** CDP Workload Manager and CDP Replication Manager are tools that automate how you replicate or migrate your data and workloads to the public cloud.
- **Step 3—Launch Production Workloads:** Seamlessly extend to production, leveraging the unified data experience of CDP. This strategic and progressive plan lets you benefit and learn from small wins and expand into more sophisticated and value driven use cases.



Using WXM to migrate to CDP Public Cloud

When to use WXM

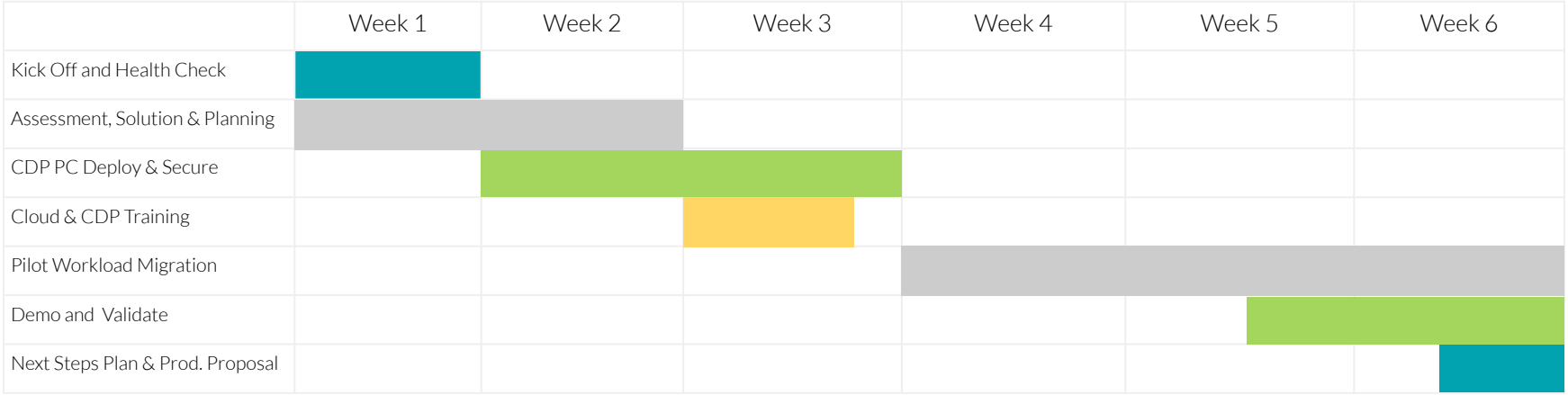
- Explore cluster & workload health before migration
- Identify right workloads to migrate
- Optimize workloads before migrating

Value proposition

- Generate ‘cloud friendliness’ Score
- Auto-generated sizing / capacity plan for target environment
- Automated Replication plan
- Mitigate risk of run-away costs in the cloud due to suboptimal workloads

PLAN: HIGH LEVEL SCHEDULE FOR MIGRATION PILOT

Scope of services include setting up CDP Data Center with Security and deploying workloads inscope to develop a production deployment plan.



Assumptions:

- Pilot scope assumes 1-2 customer workloads
- Pilot scope will be finalized during the pre-sales phase of the sales cycle. Customer will provide the data sets as extracts for pilot
- Customer will provide all necessary resources for supporting anything outside of Cloudera ecosystem

Tools to expedite Migration

Automation Toolkit

Cluster Deployment

- Ansible Automation deploys and fully secures a Private Cloud Base cluster in less than an hour
- Improves repeatability and consistency when deployment multiple environments.

Sentry to Ranger Migration

- Automatically migrate policies with Cloudera Manager and Replication Manager
- Leverage as part of the upgrade process, data migration, backup/disaster-recovery scenarios

Cloudera Navigator to Apache Atlas

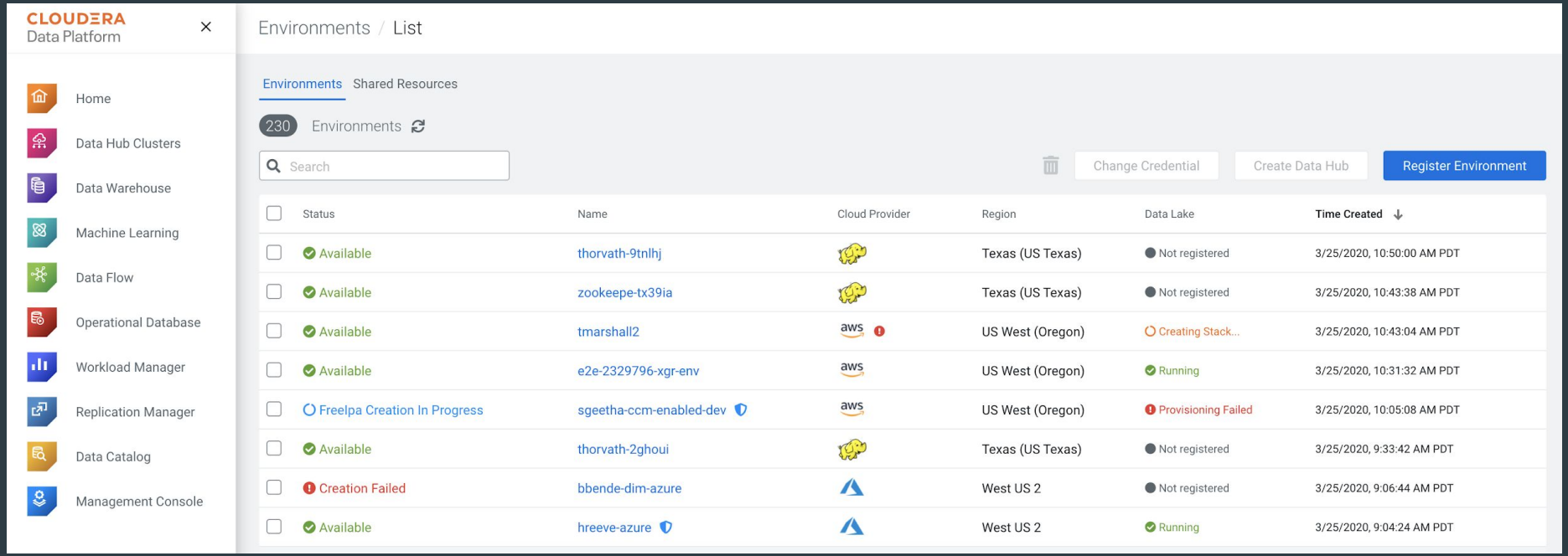
- Upgrading tooling to migrate data lineage and catalog data

YARN Fair Scheduler to Capacity Scheduler

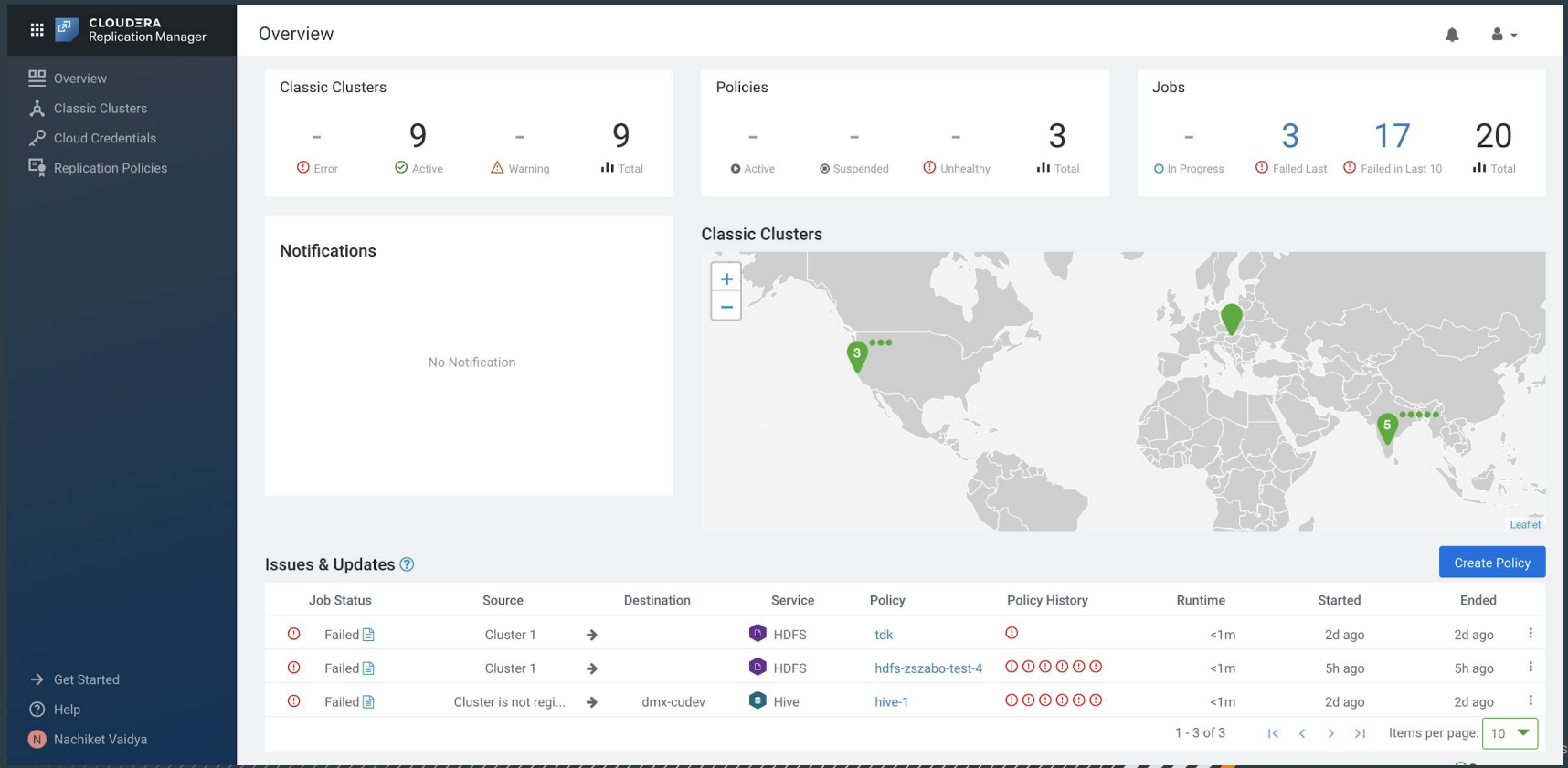
- Built in mechanism to migrate configurations to equivalent settings

SINGLE PANE OF GLASS

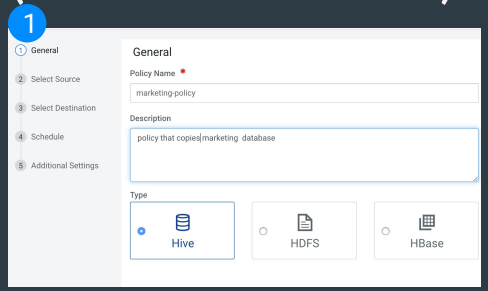
All registered environments on-prem & cloud

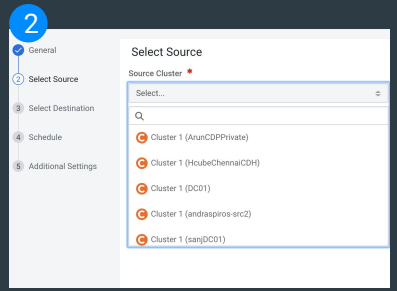


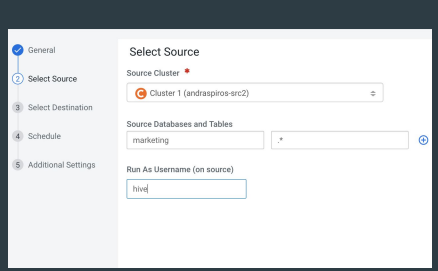
SINGLE PANE OF GLASS FOR ALL REPLICATION POLICIES

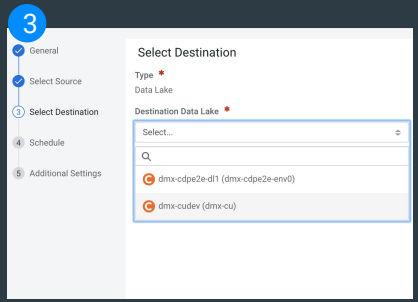


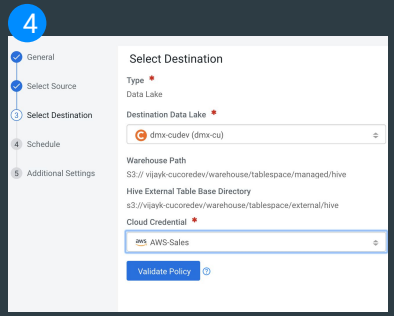
WIZARD BASED POLICY CREATION FOR DATA WAREHOUSE (HIVE & IMPALA)

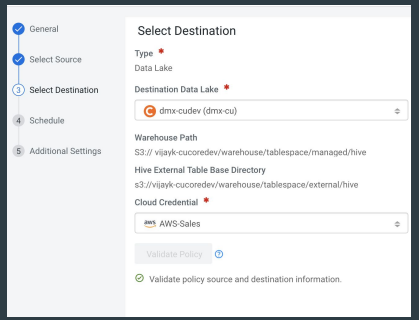


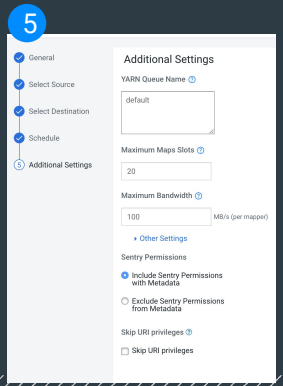


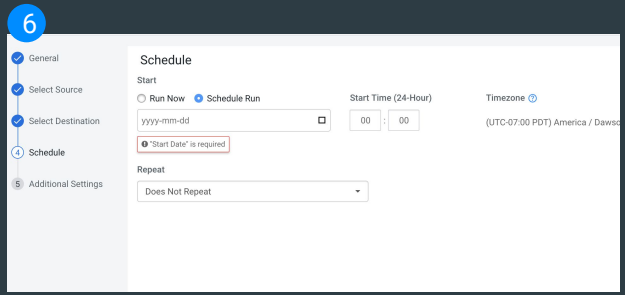






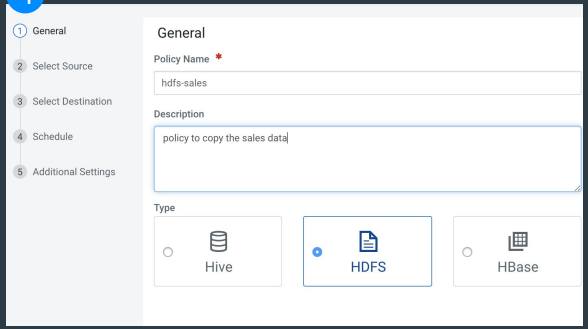




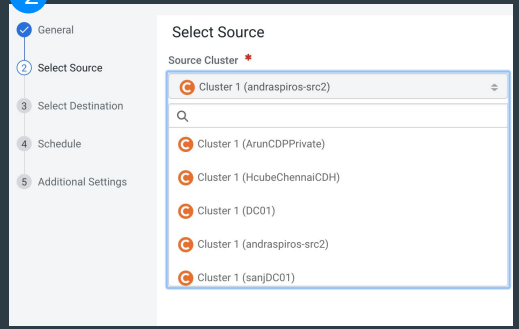


WIZARD BASED POLICY CREATION FOR DATA ENGINEERING & DATA SCIENCE (HDFS)

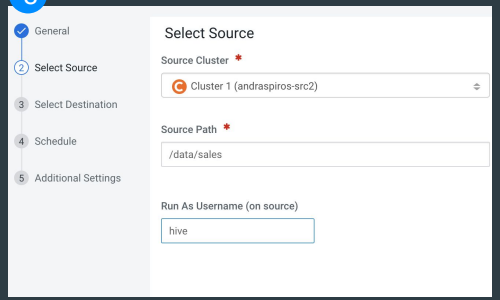
1



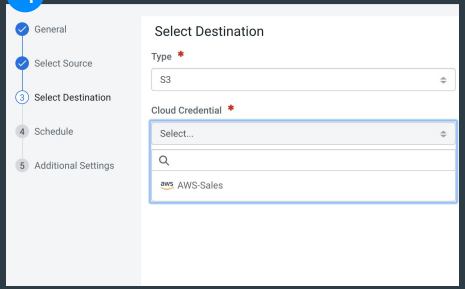
2



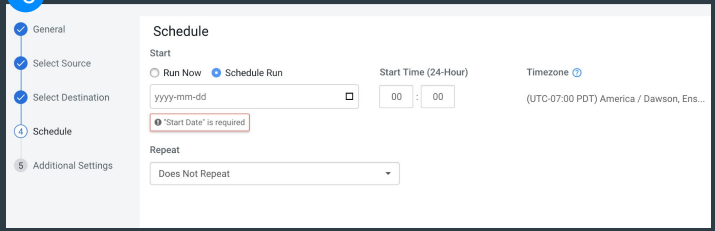
3



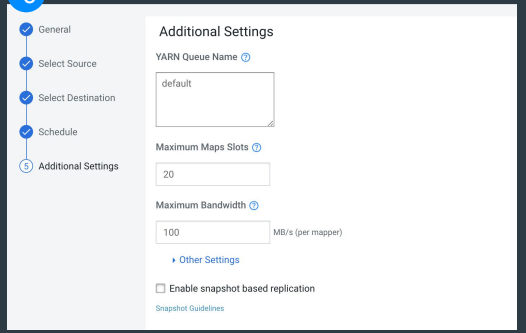
4



5

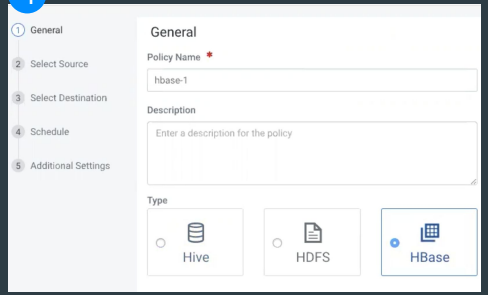


6

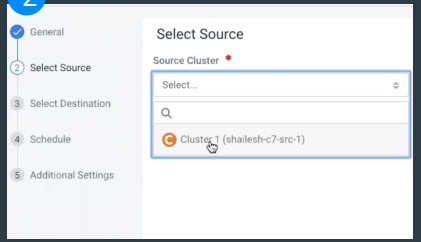


WIZARD BASED POLICY CREATION FOR OPERATIONAL DATABASE (HBASE/PHOENIX)

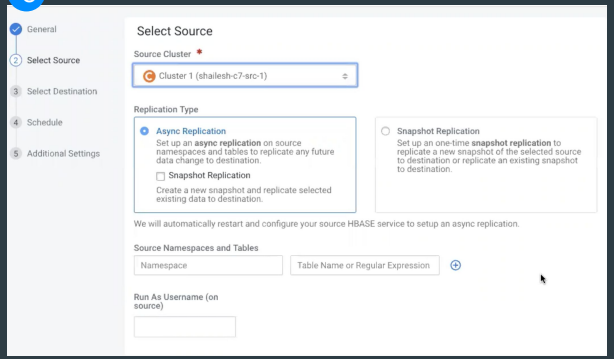
1



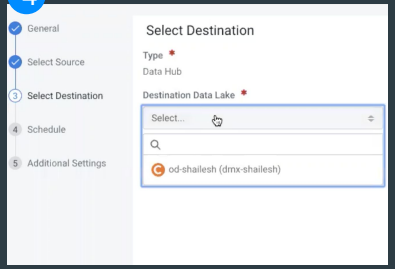
2



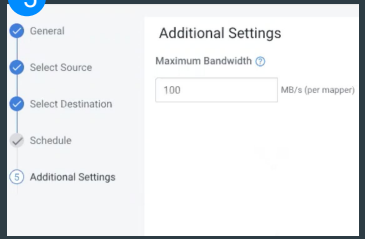
3



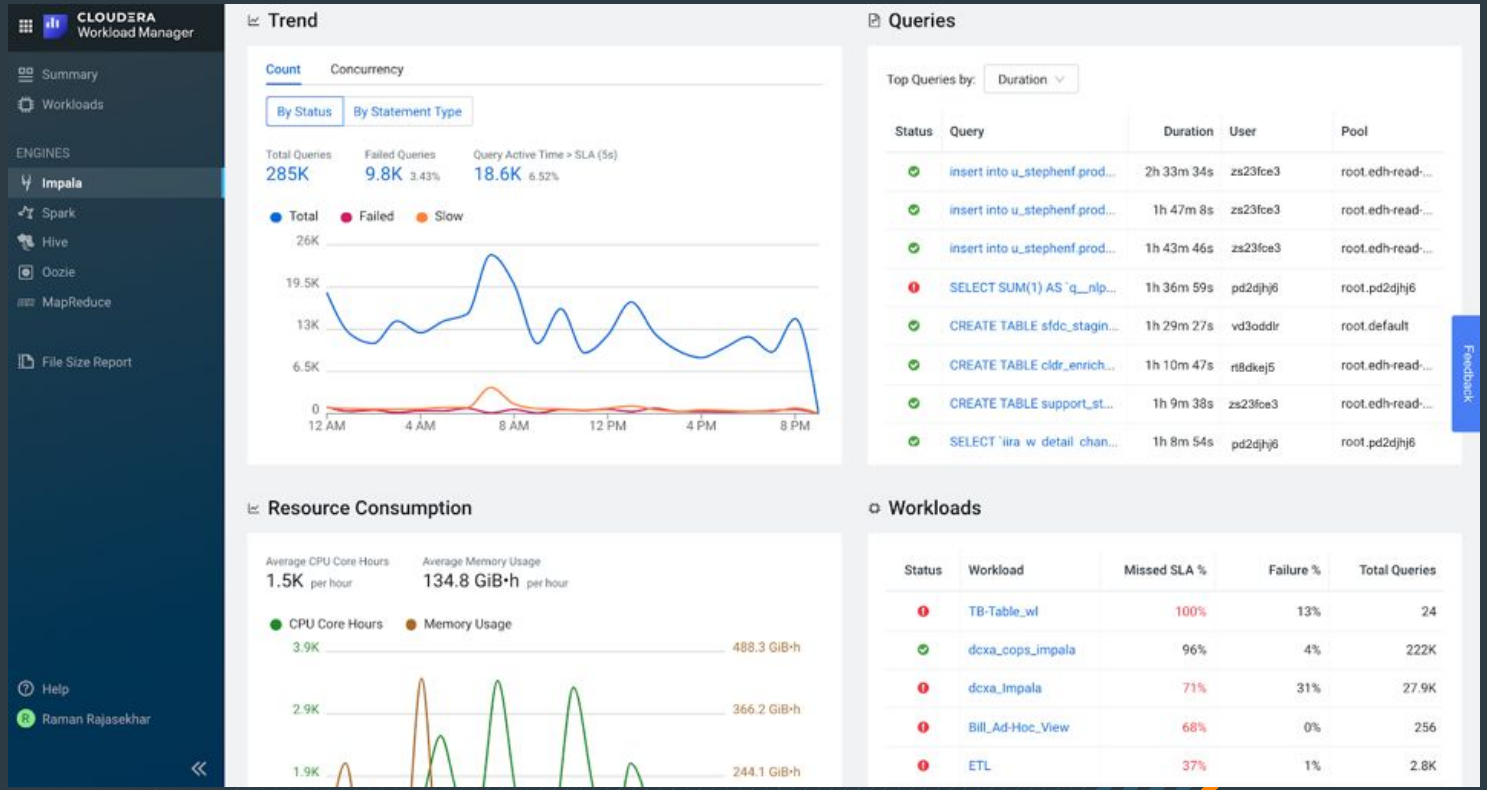
4



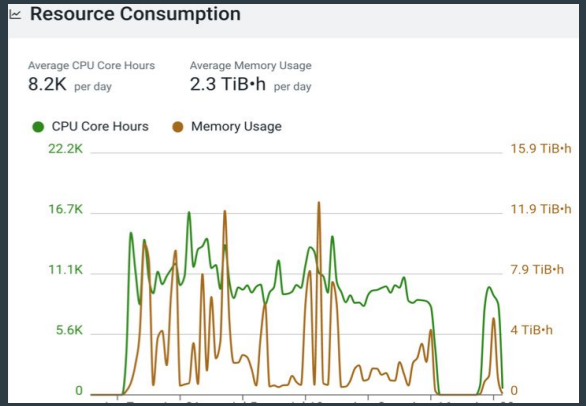
5



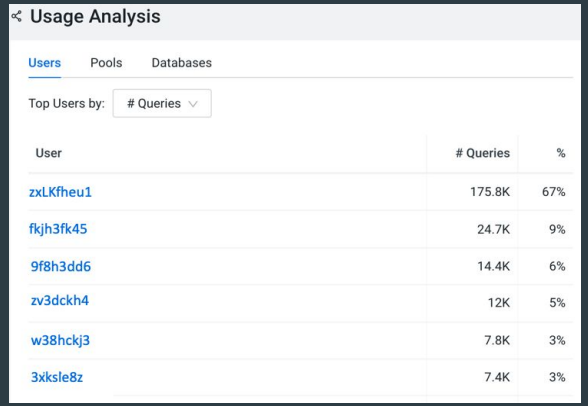
ACCELERATE UPGRADES WITH WORKLOAD MANAGER



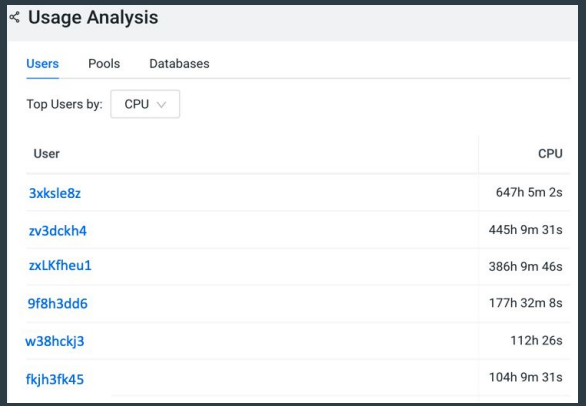
01 - Analyze Resource Consumption



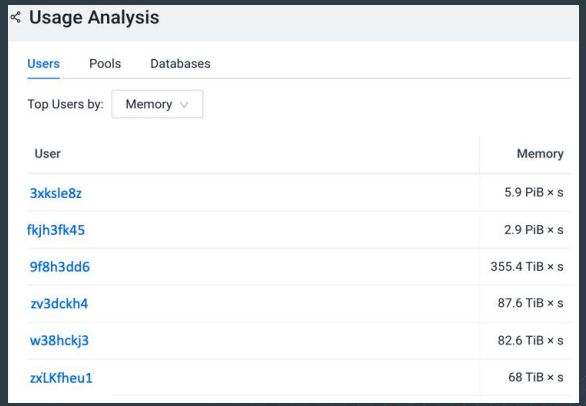
02 | 1 - Identify Rogue Users



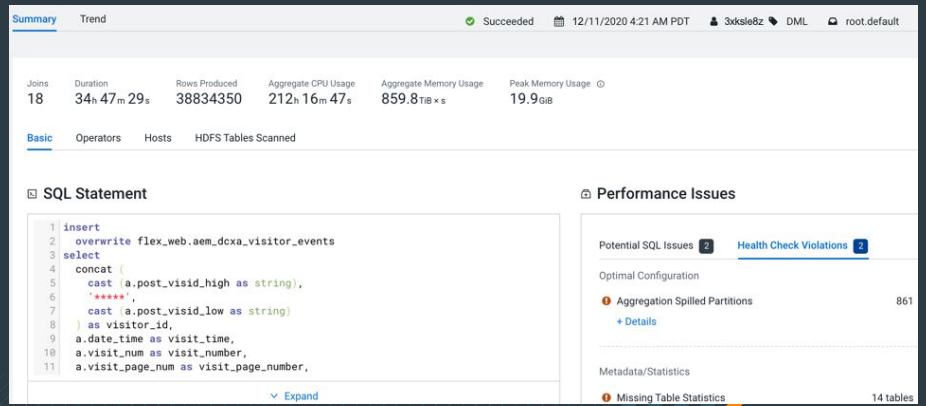
02 | 2 - Identify Rogue Users



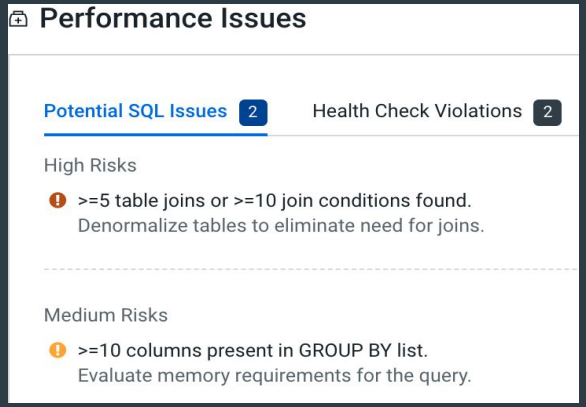
02 | 3 - Identify Rogue Users



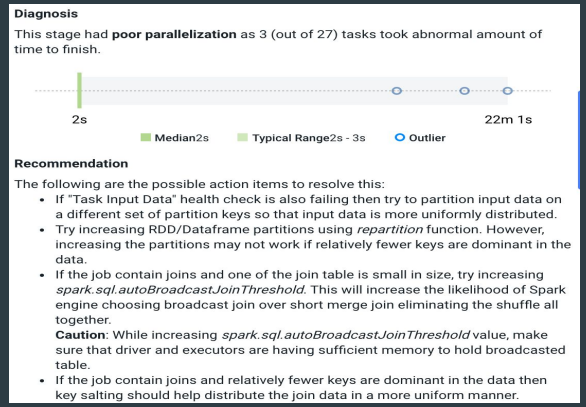
04 - Analyze Rogue User's Queries / Jobs



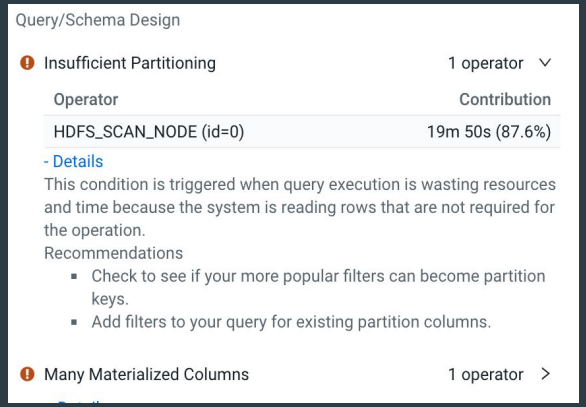
05 | 1 - Plan Optimization Pre/Post Upgrade



05 | 2 - Plan Optimization Pre/Post Upgrade



05 | 3 - Plan Optimization Pre/Post Upgrade



06 | 1 - Performance Benchmark Pre v/s Post Upgrade

Start Time	Status	Duration	User	Health Issue
12/07/2020 5:10 PM PST	Succeed...	6m 20s	tserver	Corrupt Table Statis
12/07/2020 1:03 PM PST	Succeed...	7m 57s	tserver	Corrupt Table Statis
12/07/2020 9:03 AM PST	Succeed...	3m 37s	tserver	Corrupt Table Statis
12/07/2020 5:02 AM PST	Succeed...	8m 14s	tserver	Corrupt Table Statis
12/07/2020 1:11 AM PST	Succeed...	3m 7s	tserver	Corrupt Table Statis

06 | 2 - Performance Benchmark Pre v/s Post Upgrade

Metric	Baseline	Current Job	
Number of Task attempts	134	1K	+870
Number of Tasks	134	995	+861
Output bytes	728.5 MiB	0 B	-728.5 MiB
Output records	396.6K	0	-396.6K
Shuffle Read bytes	265.5 B	38.5 MiB	+38.5 MiB
Shuffle Read records	4.5	796K	+796K
Shuffle Write bytes	265.5 B	52.6 MiB	+52.6 MiB
Shuffle Write records	4.5	1.2M	+1.2M
Succeeded Task attempts	134	992	+858

Additional Details on Migration

VERSION CHANGES (1 OF 2)

Component	CDH 5.16	CDH 6.3	CDP 7.1
Apache Hadoop	2.6.0	3.0.0	3.1
Apache HBase	1.2.0	2.1.4	2.2
Apache Hive	1.1.0	2.1.1	3.1
Hue	3.9.0	4.3.0	4.3
Apache Impala	2.12.0	3.2.0	3.4
Apache Kudu	1.7.0	1.10.0	1.12
Apache Solr	4.10.3	7.4.0	8.4
Apache Oozie	4.1.0	5.1.0	5.1
Apache Sentry	1.5.1	2.1.0	Replaced by Ranger2.0
Navigator	2.15	6.3	Replaced by Atlas2.0

VERSION CHANGES (2 of 2)

Component	CDH 5.16	CDH 6.3	CDP 7.1
Apache Spark	1.6.0	2.4.0	2.4.5
Apache Kafka	1.0.1	2.3.0	2.3.0
Apache Sqoop	1.4.6	1.4.7	1.4.7
Apache ZooKeeper	3.4.5	3.4.5	3.5.5
Apache Phoenix	4.7.0	5.0.0	5.0
Apache Parquet	1.5.0	1.9.0	1.10
Apache Avro	1.7.6	1.8.2	1.8.2
Apache Flume	1.6.0	1.9.0	Replaced by NiFi
Apache Knox	-	-	1.3
Apache TEZ	-	-	0.9
Apache Zeppelin	-	-	0.8

COMPONENT MIGRATION RECOMMENDATIONS

	Approach	Tooling
Flume	Port to CFM (Apache NiFi)	No
Storm	Port to CSA (Apache Flink)	No
Sqoop2	Port to Sqoop 1 (or Apache NiFi)	No
Pig	Port to Apache Spark	No
Crunch	Port to Apache Spark	No
YARN Fair Scheduler	Configure Capacity Scheduler in CM UI	Yes
Sentry	Migrate to Ranger	Yes
Navigator	Migrate to Atlas	Yes
Ambari	Migrate to Cloudera Manager	Yes

ON-PREM UPGRADE PREPARATION

Component	Approach
Spark	Migrate all jobs to Spark 2
JDK	Install JDK 1.8 on all nodes
OS	Upgrade all nodes to RHEL 7.6+
HDP	Upgrade to 2.6.5
CDH	Upgrade to 5.13 - 5.16 if older than 5.13
Pig/Crunch/Sqoop/Flume	Migrate to replacement (Spark, Flink, NiFi)
DB	Assess if DB upgrade required
3rd Party	Validate support/certification on CDP-DC

SUMMARY: MIGRATING DATA WAREHOUSING & OP DB to CDP

	Component	Approach summary
Warehousing	Hive / Impala	• WXM (workload centric approach) or Replication manager
	Druid	• Manual migration to Public Cloud (only)
	SOLR	• Backup your data and schema etc. • Use the scanner to figure out what in your schema needs to change. • Use Replication Manager to copy your data to CDP cluster. • Redesign your schema for Solr 8 and reindex using MRindexer
	HUE	• Export & import the Hue database
	Kudu	• Kudu backup tool to backup & restore to new cluster. Manually migrate configs
	Hive LLAP	• Public Cloud and Private Cloud
OP DB	HBase Phoenix	• Replication manager to migrate data to CDP Public Cloud or HBase native replication to migrate data to CDP D/C • Migrate configurations manually (& retune) • Manually migrate ACLs to Ranger or continue to use native HBase ACLs
	Accumulo	• n/a – Will be supported in CDP DC 7.2

SUMMARY: MIGRATING ML, DE & DF TO CDP

	Component	Approach summary
ML / DE	CDSW	Migrate to CDSW side-car or CML Public Cloud/Private Cloud. Project piecemeal , or backup/restore coming soon.
	Zeppelin	Manually copy notebook files to CDSW as Zeppelin 3rd party editor or CDP DC Zeppelin.
	Spark	Replicate HDFS/Hive. Will need to use HiveWarehouseConnector when working with Hive 3 managed tables (ACID)
	Livy	n/a
	Oozie	Copy associated jars, files, spark code, including the Oozie WF, Coord, Bundles from oozie workspace to new cluster.
DF	Kafka	If Schema Registry is being used, back up schemas. In CDP-DC, set up Kafka, SR, SRM. Import Schemas in new Schema Registry. Set up SRM on target CDP-DC cluster to replicate data from source cluster. Migrate existing Sentry policies to ranger with policy migration tool. Migrate Ranger policies to Ranger in CDP-DC
	Nifi	Upgrade CFM/HDF to latest 3.5/1.1. Migrate other workloads. Update flows on CFM/HDF to use appropriate processors for new component versions in CDP-DC. Install CFM on CDP-DC. Move flows. Migrate state if needed. Migrate NiFi Registry. Migrate auth policies either from Ranger/built-in NiFi

SUMMARY: MIGRATING SDX TO CDP

Component	Approach summary
Sentry	• Replication Manager will convert policies and migrate policies to Ranger as part of Hive/Impala replication (for CDP-PC) • Kafka/Solr permissions need to be manually converted into Ranger policies • HDFS ACLs which are automatically set-up by Sentry will need to be manually converted into Ranger policies
Navigator	• Navigator Managed metadata tags & any manually ported to Atlas Business Metadata Tags • Audit data kept in Read Only mode in CM until expired
Ranger	• Ranger Policy Import/Export feature can be used to migrate existing Ranger policy into CDP-PC and/or CDP-DC
Ranger KMS	• Use distcp to copy data into Cloud Native encrypted storage (CDP-PC) or into another HDFS encryption zone (CDP-DC) • Data re-encryption will take place during copy
Atlas	• Managed In place upgrade to CDP-DC from CDH (as of 7.1) / HDP(roadmap) • CDP Public cloud will have Atlas wired up to all workloads, and replication manager and porting of jobs will recreate lineage. ? • Atlas Import / Export can be used to port legacy atlas data to new deploy.
Key Trustee Server Key Trustee KMS Key HSM HSM KMS	• Use BDR/Replication Manager to migrate encrypted data to Public cloud or to CDP-DC
NavEncrypt	• Migrate data from encrypted volumes to Cloud native encrypted storage (for CDP-PC) or to another NavEncrypt encrypted volume (in CDP-DC) • Data re-encryption will take place during migration
Knox	• n/a – No migration required

SUMMARY: MIGRATING PLATFORM COMPONENTS TO CDP

Component	Approach summary
HDFS	Replication Manager to replicate the data
Parquet / ORC / AVRO	n/a
YARN	Migrate configurations, recreate queues
Cloudera Manager	Adapt to new cluster manager. Automate scripts or other api interfaces.
BDR	BDR renamed to Replication Manager in CM 7.0 (CDP DC only). Policies need to be manually migrated
Ambari	n/a. Learn how to use Cloudera Manager
DLM	PS to migrate replication policies to Replication Manager
ZooKeeper	n/a. There is nothing to migrate

PRECAUTIONS FOR IN-PLACE UPGRADE

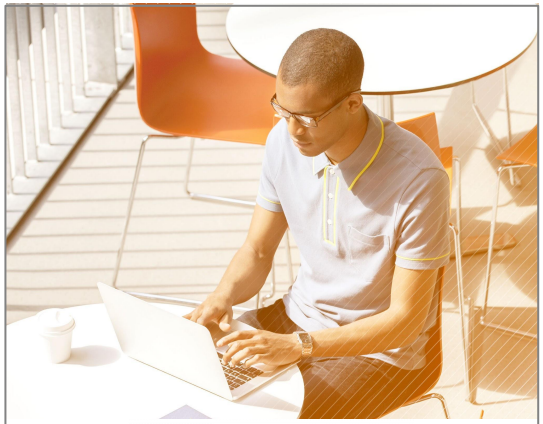
- Testing before the upgrade is crucial. Almost all the problems encountered after the upgrade are caused by the early neglect of system testing or application testing
- The test environment must be configured with similar parameters as the production environment, even if it may take a lot of time to perform consistent matching
- Develop a detailed upgrade plan. It is best to accurately record the command that must be executed during the upgrade process, and it is not recommended to temporarily modify the steps during the upgrade.
- The rollback operation is also very important. Even if we never plan to use rollback, testing the rollback steps makes us more at ease
- Although data migration is not necessary for in-place upgrade, strongly recommend customers make data backups in advance
- Major version upgrades usually have service interruption time, generally no more than 8 hours (scheduled at night/weekend time). Be sure to coordinate the downtime window of each business department on the platform. If you have special requirements, you can contact Cloudera to provide additional solutions

OVERALL IDEA OF MIGRATION UPGRADE

- Build a new cluster of CDH/CDP
- replication all the data from old cluster to new cluster
- Since the amount of data in the old cluster is relatively large during the full copy process, it will take a long time, and new data will be written to the old cluster, so incremental copying is required
  - The amount of data for incremental copy is relatively small and the time required is relatively short. You can iterate multiple incremental copies until the final incremental copy time reaches an acceptable time range
  - Choose an appropriate time point, stop the application of the old cluster, ensure that no new data is written, and perform incremental copying. (If there is an application data supplementary recording mechanism, you can also use the data supplementary recording function to complete the data)
- After the incremental copy is completed, theoretically the new and old cluster data is consistent, and the data consistency check starts
- Finally, switch the cluster

Customer Story

CLOUDERA POWERING DATA-DRIVEN CUSTOMERS



LinkedIn

At-scale migration from on-premises to public cloud for increased flexibility and scale

REEF

Enabling new growth opportunities via CDP Public Cloud and streaming analytics

ExxonMobil

Increasing ROI via a hybrid cloud for massively scalable, real-time data ingestion

CUSTOMER UPGRADE: REGIONAL US BANK

CDH 5.14.2 to CDP PVC BASE 7.1.2

10 Weeks

3 Clusters

1PB Data

CHALLENGE

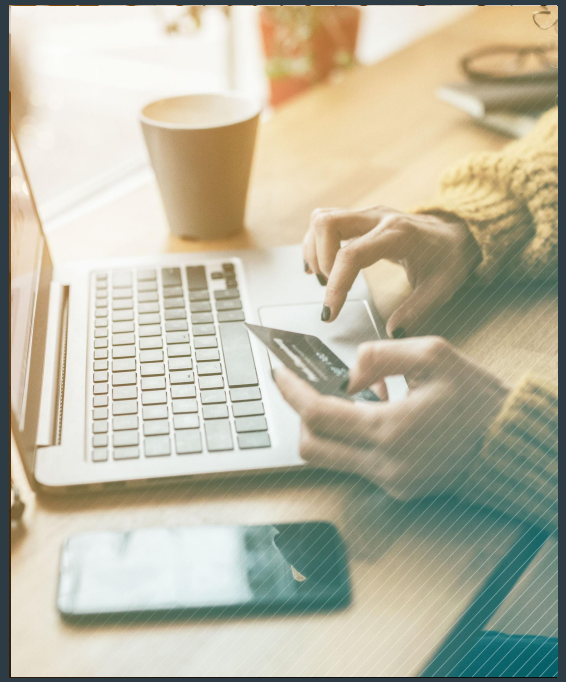
Need to modernize architecture to ingest real-time data using new CDP streaming capabilities while leveraging existing infrastructure and minimizing downtime

SOLUTION

In-place upgrade to minimize downtime and risk for test, dev and production clusters. Cloudera PS supported upgrade planning, process and implementation

OUTCOMES

Successful upgrade to CDP PVC Base, enable use cases to leverage new capabilities: Apache Ranger, Atlas, Kafka, Hive 3



Cloudera PS & Training

SMARTSERVICES

Cloudera Professional Services to Enable the Path to CDP PVC Base

SMARTARCHITECTURE

Begin defining a **hybrid strategy** and architecture to modernize your analytics with CDP

Develop an architecture and roadmap to a tailored Enterprise Data Cloud strategy

SMARTDEPLOY

Quickly **install and secure your first cluster** on CDP Private Cloud based on best practices

Plan use case implementation with roadmap

SMARTUPGRADE

Enable CDP Private Cloud with either **in-place upgrade or migration** from existing CDH or HDP deployments

Sizing calculators assess time of engagement based on risk and complexity of environment

SMARTPRIVATE

Begin your analytics journey with secure platform deployment on **PVC Experiences and pilot workload on CDW or CML**

Includes OnDemand training license

Fixed Fee Offering

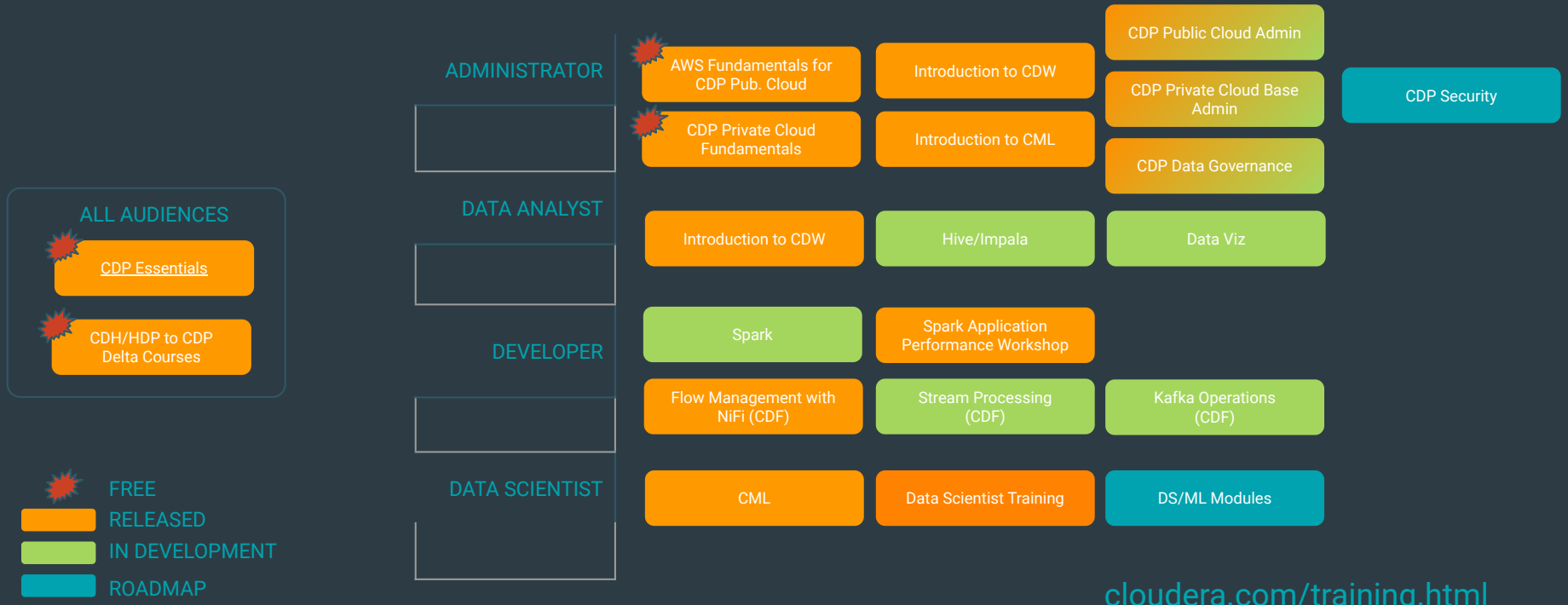
TRAINSMART

Now Available! Administrator Training: CDP Private Cloud Base

CLOUDERA

CLOUDERA EDUCATIONAL SERVICES

Supporting Customers' Move to the Enterprise Data Cloud



cloudera.com/training.html

Key Resources

Key Bookmarks

CDP Upgrade/Migration
https://docs.cloudera.com/cdp-private-cloud-upgrade/latest/upgrade/topics/cdpdc-cdp-upgrade-migrations-paths.html

Blog https://blog.cloudera.com/upgrade-journey-the-path-from-cdh-to-cdp-private-cloud/

Migration Blog https://blog.cloudera.com/migrate-to-cdp-private-cloud-base-a-step-by-step-guide/

Upgrade Advisor https://my.cloudera.com/

Upgrade Advisor (External) https://www.cloudera.com/products/cdp-migration/migration-from-cdh.html?tab=0

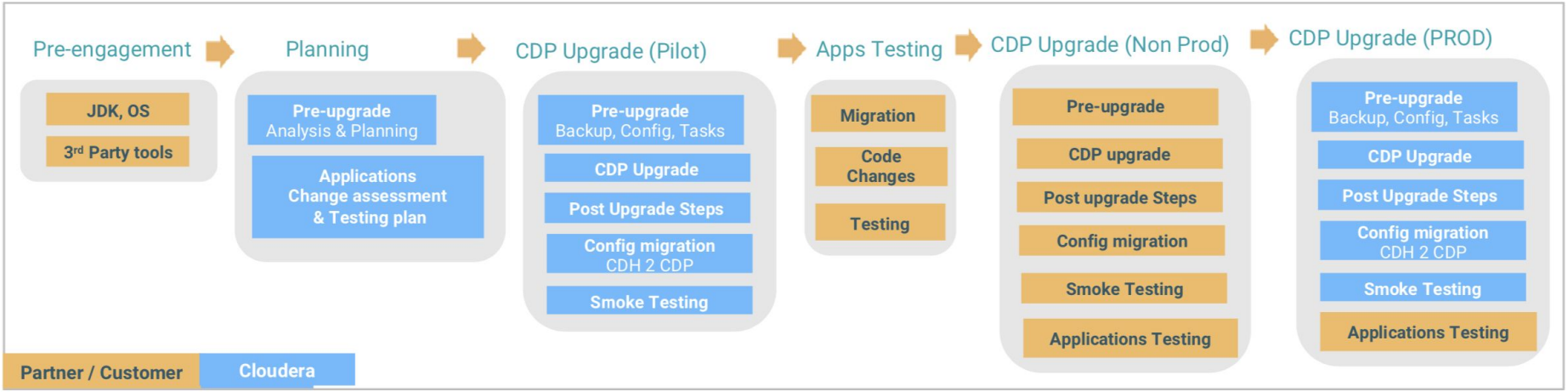
Cloudera PS Packages https://www.cloudera.com/about/services-and-support/professional-services.html

Cloudera Trainings https://www.cloudera.com/about/training.html#

THANK YOU

CLOUDERA

ACTIVITIES SUMMARY



Analysis & Planning

- Review CDH cluster deployment
- Review App arch design
- Review data pipeline
- Review known issues
- Review Sentry policies
- Assess 3rd party tools compatibility risks
- Applications changes assessment
- Validate pre-requisites for upgrade
- Create detailed plan

Cluster Upgrade (Pilot)

- Pre-upgrade prep & configuration
- Upgrade to CDP
- Post upgrade configuration
- Post upgrade platform configuration migration
- Smoke testing