

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

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



Secure & Governed



Open

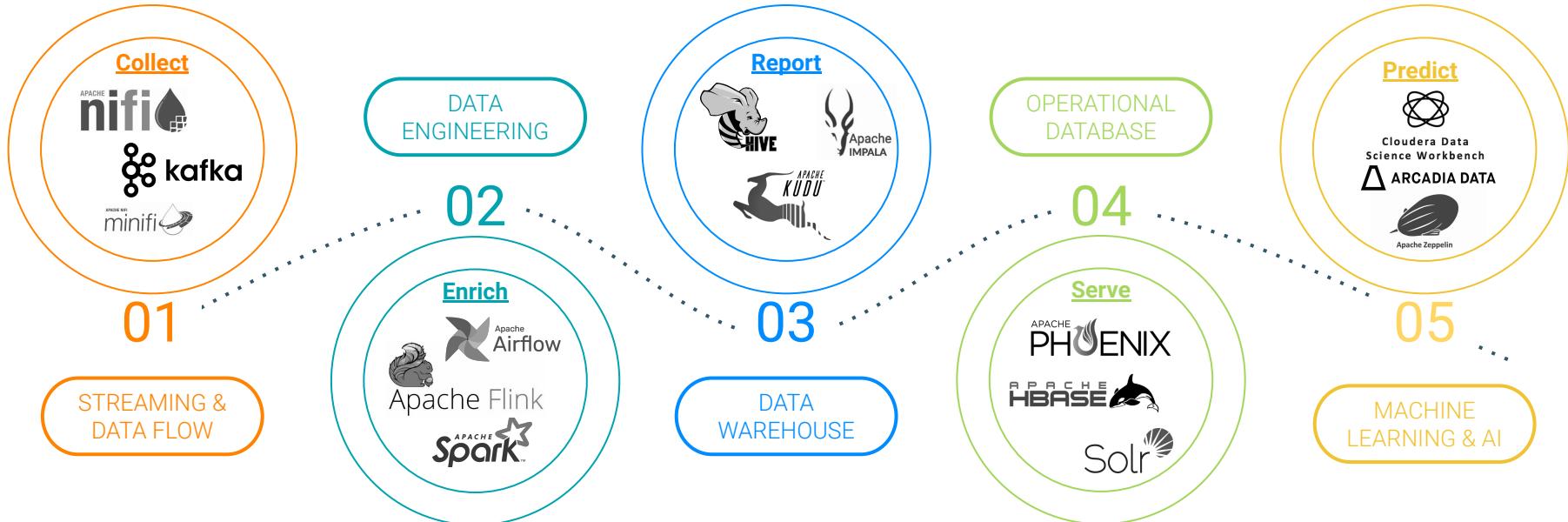
# CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter



SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

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

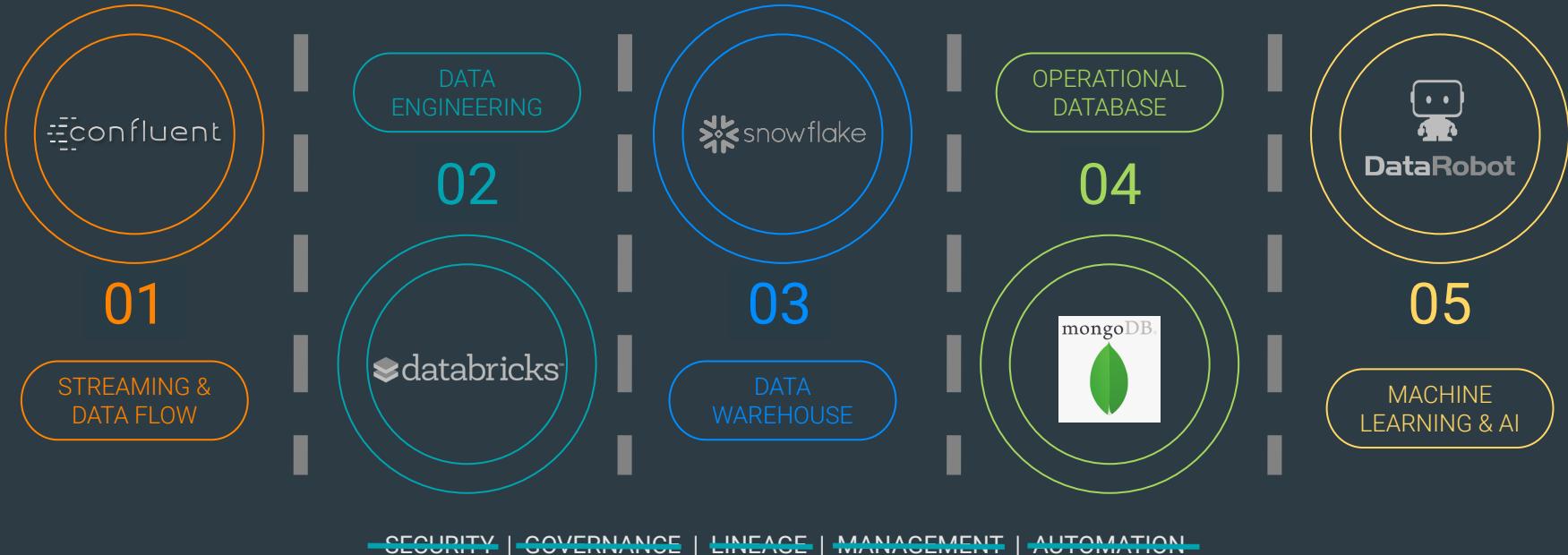


SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION



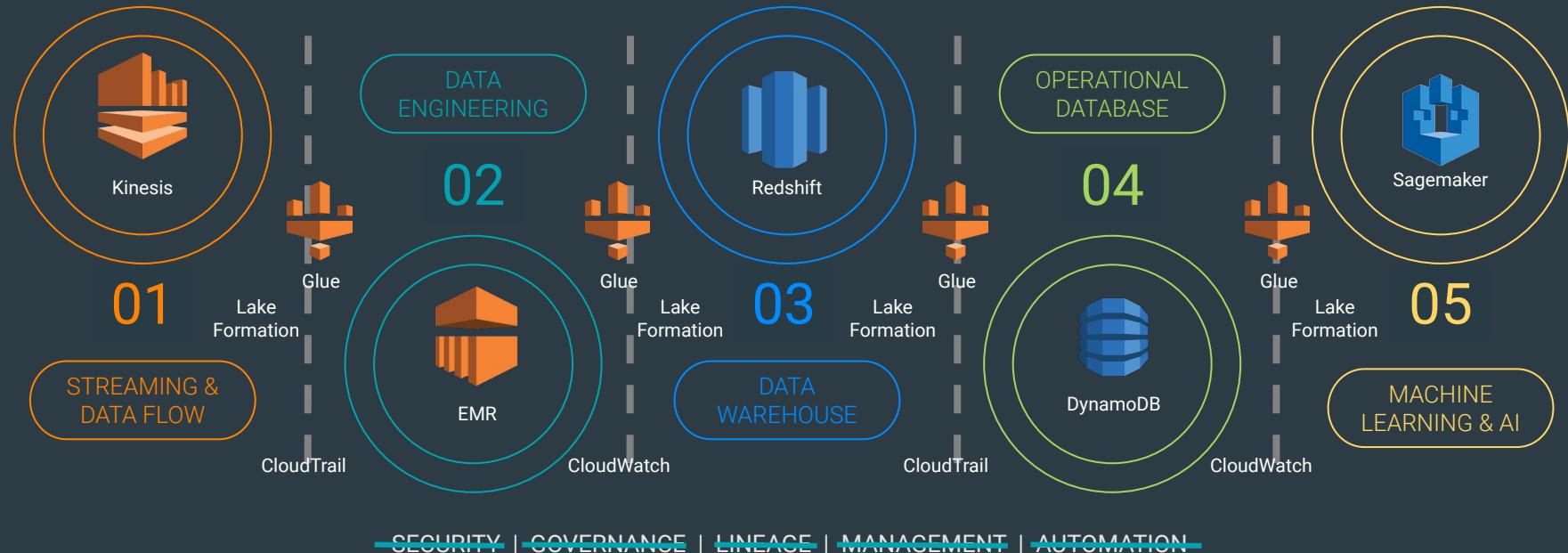
# POINT SOLUTIONS HAVE AN INTEGRATION TAX

Security & governance is an afterthought



# POINT SOLUTIONS FROM PUBLIC CLOUD PROVIDERS

## Building blocks

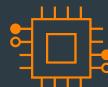


# HOW DO CUSTOMERS USE CLOUDERA?

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

KEY  
CUSTOMERS

- Barclays
- Citi
- Santander UK

- Customer analytics
- Threat detection
- Predictive support

- Cisco
- Intel
- Reef Technology

- Churn analysis
- Customer care
- Network optimization

- Globe Telecom
- Deutsche Telecom
- Robi Axiata

- Patient care (IoT)
- Genomics research
- Regulatory compliance

- GlaxoSmithKline
- Clearsense
- Cerner

- Predictive maintenance (IoT)
- Supply chain optimization
- Remote monitoring

- Navistar
- Micron
- Sikorsky

# INDUSTRY ANALYST RECOGNITION

## Enterprise Data Cloud

*Enterprise Data Platform*



January 2021

*Cloud Data Ecosystems*



January 2020

*Enterprise Intelligence Platforms*



December 2019

## Cloudera Data Platform (CDP)

*...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..."*



January 18, 2021



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

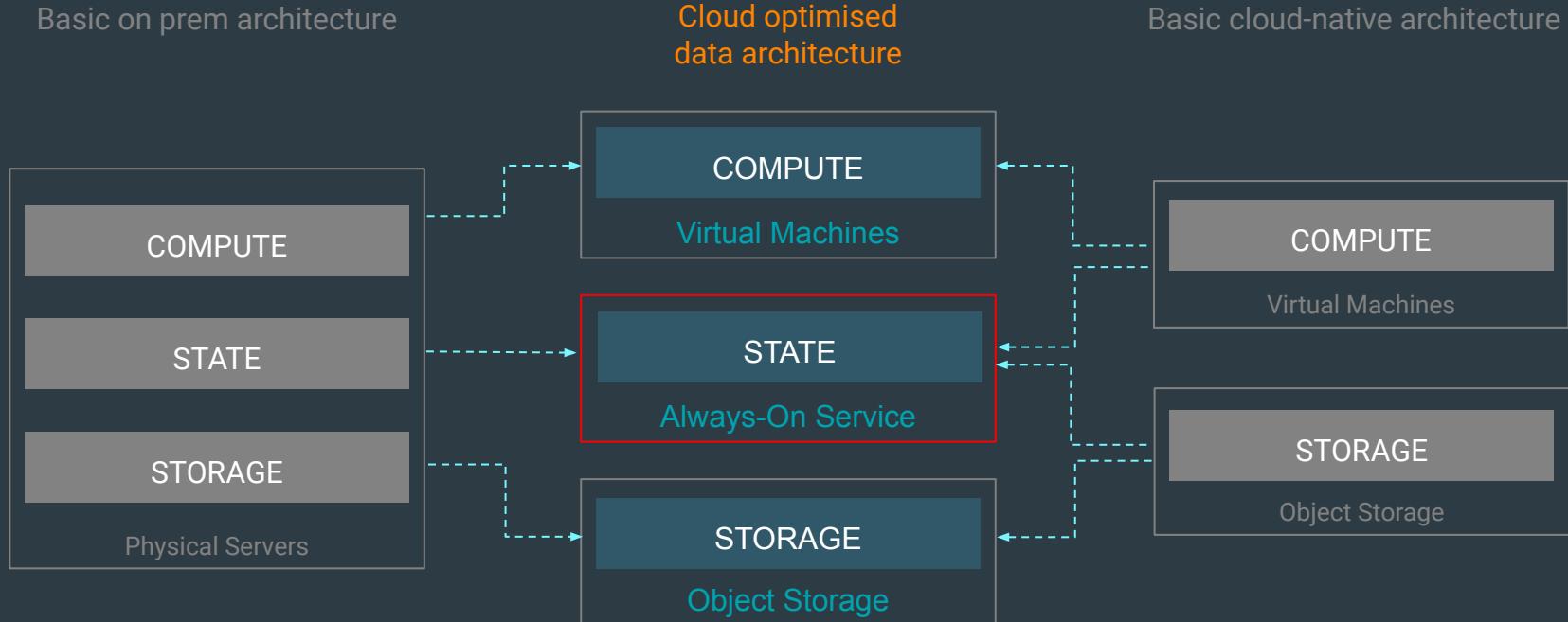
52%

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

PUBLIC CLOUDS  
compute & storage

DATACENTER  
compute & storage

SECURITY & GOVERNANCE

IOT, INGEST &  
STREAMING

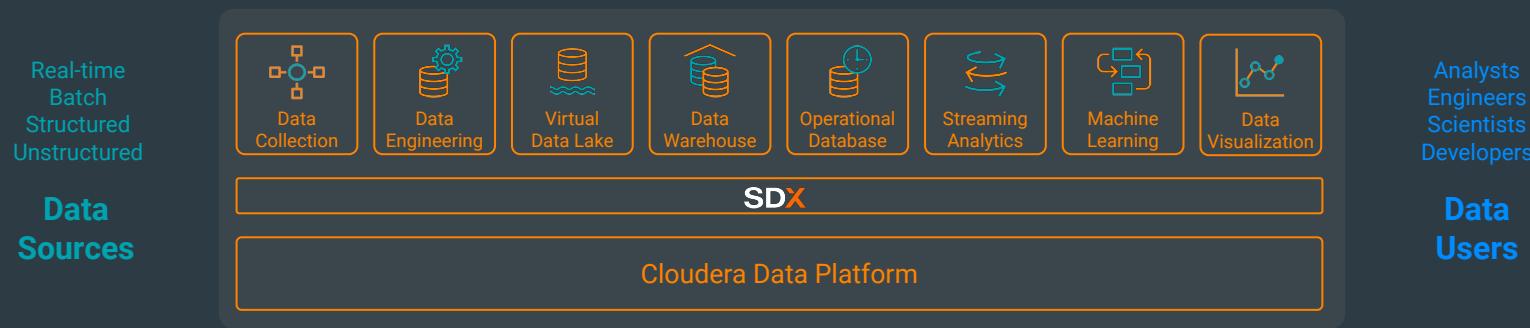
DATA  
WAREHOUSING

ML / AI  
DATA SCIENCE

---

# CLOUDERA DATA PLATFORM

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



**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

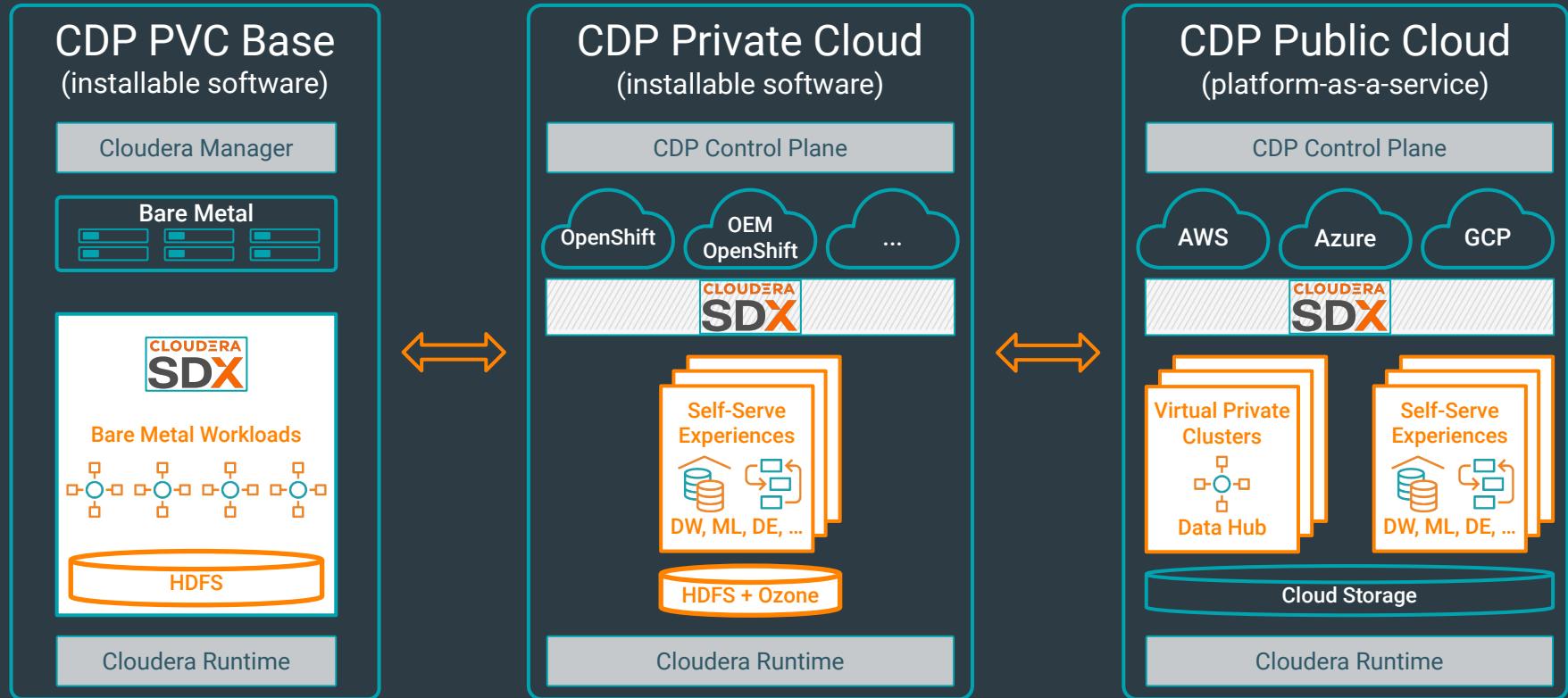
**Schema:** automatic capture and storage of any and all schema and metadata definitions as they are used and created by platform workloads

**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

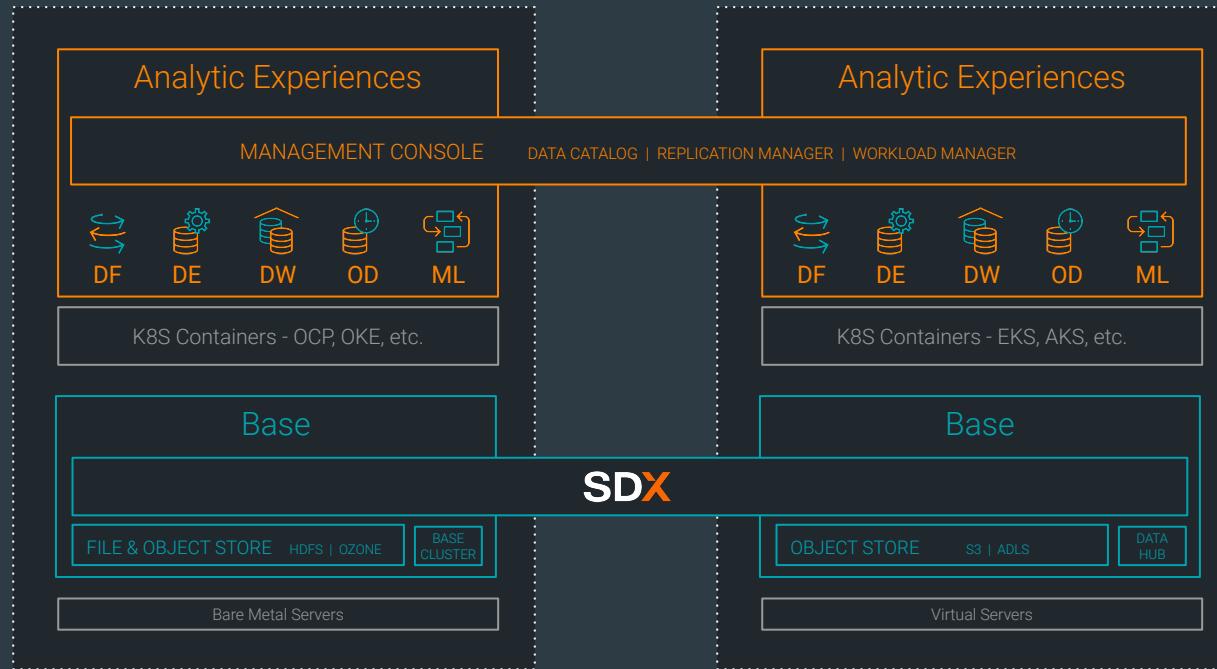
# CDP - ONE PLATFORM, THREE FORM FACTORS



# CDP HYBRID CLOUD

Consistent operations and analytics experiences across private and public clouds

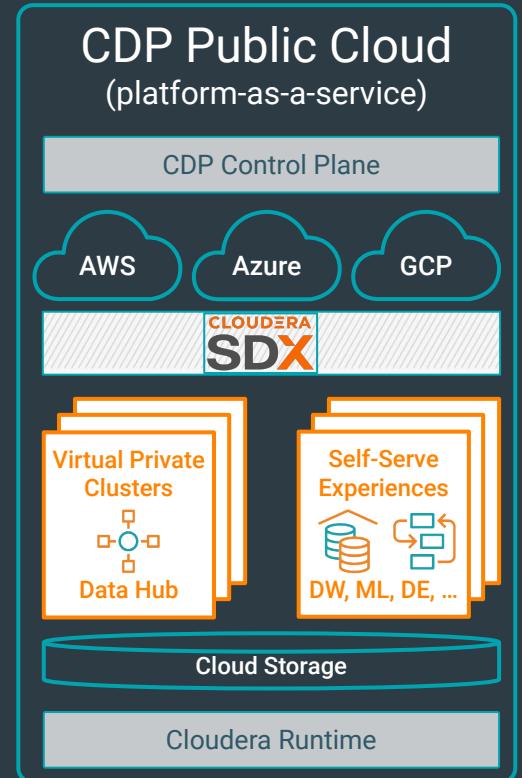
CDP  
Private  
Cloud



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



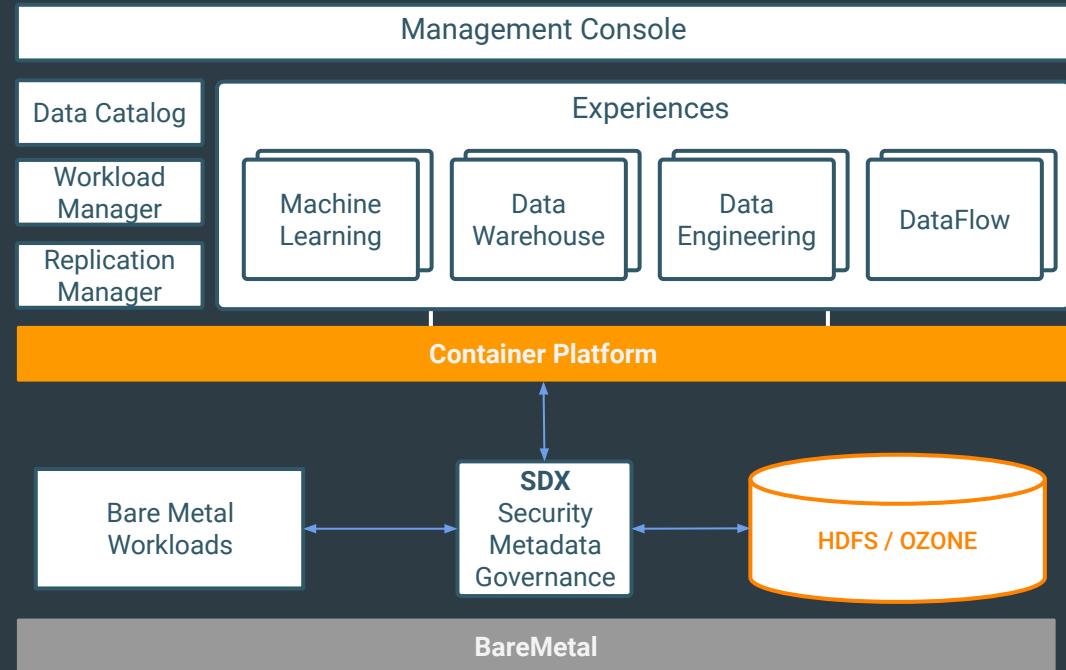
## Burst to Cloud

- 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?



Data Hub Clusters



DataFlow



Data Engineering



Data Warehouse



Operational Database



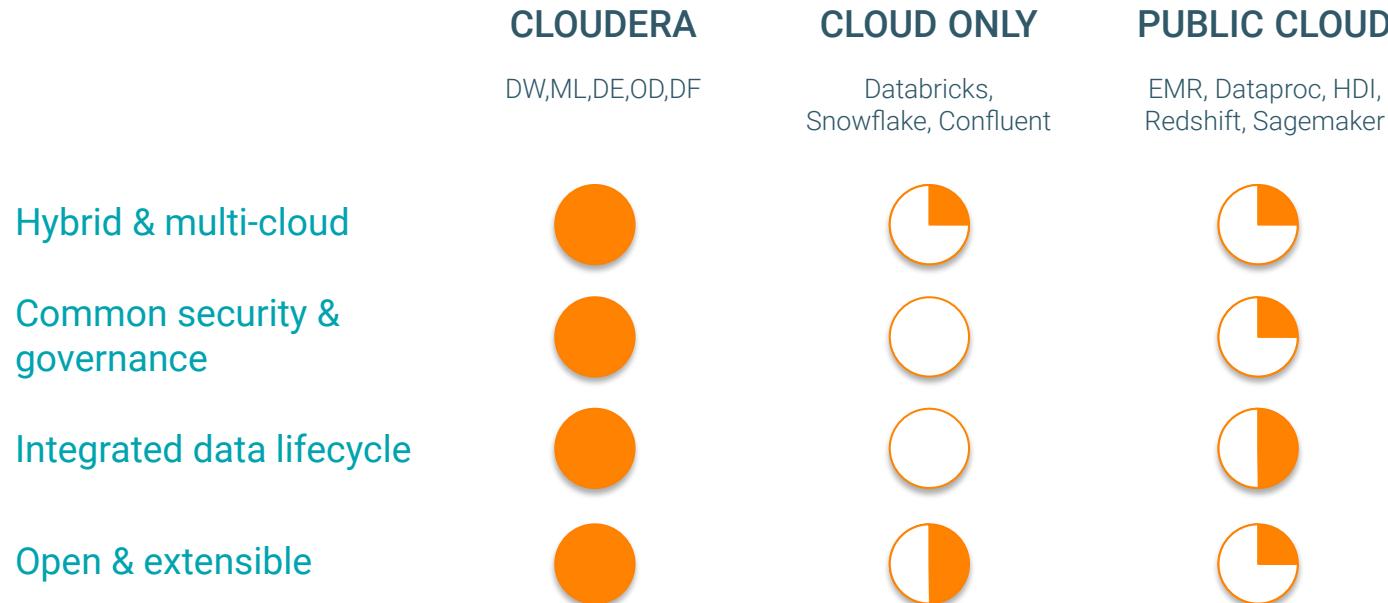
Machine Learning

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



\$300K

\$12-16M

## After CDP



\$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

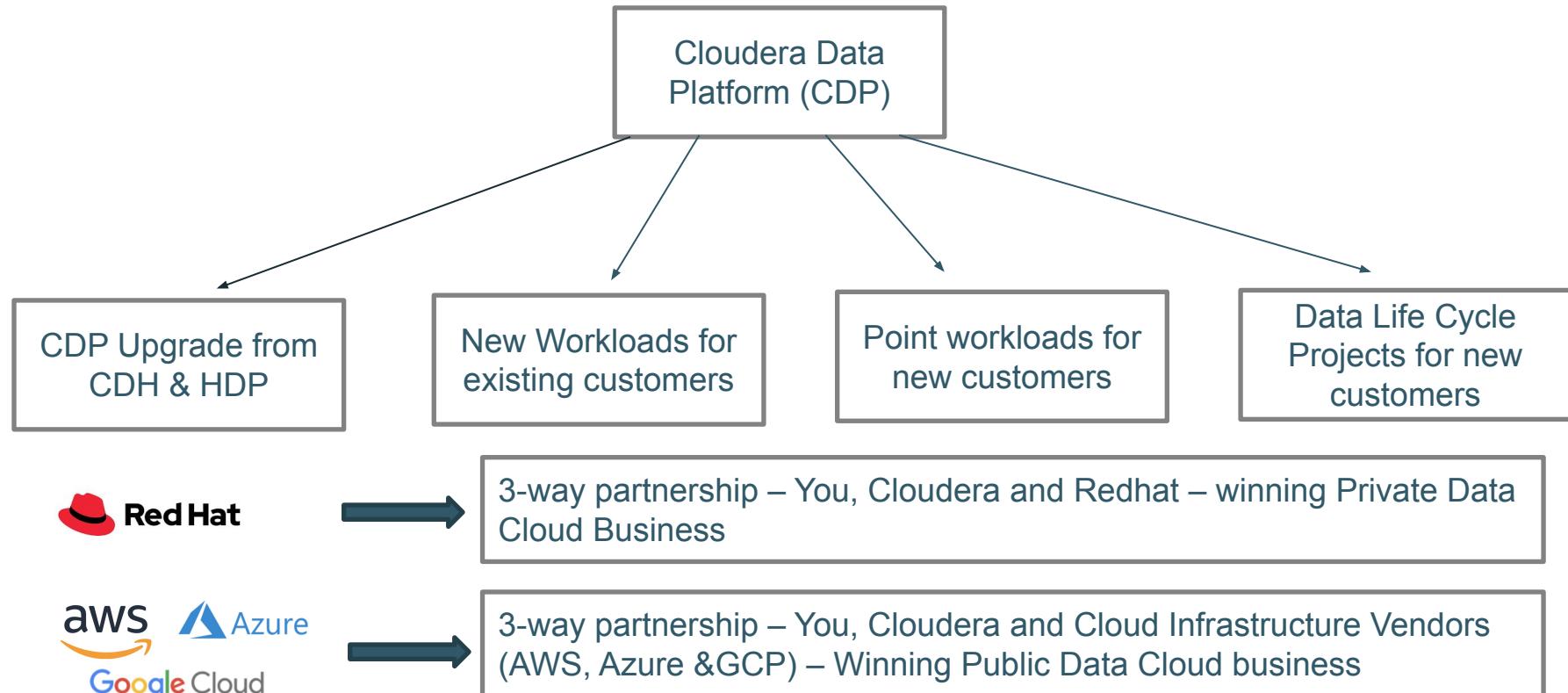
Deeper strategic client engagements with longer, multiple cycles

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



# THE PATH

Existing  
Customers



New  
Workloads



New  
Customers

On-Prem Migration

In-Place Upgrade

New Cluster Deploy

Direct to Cloud Migration

CDP Private Base

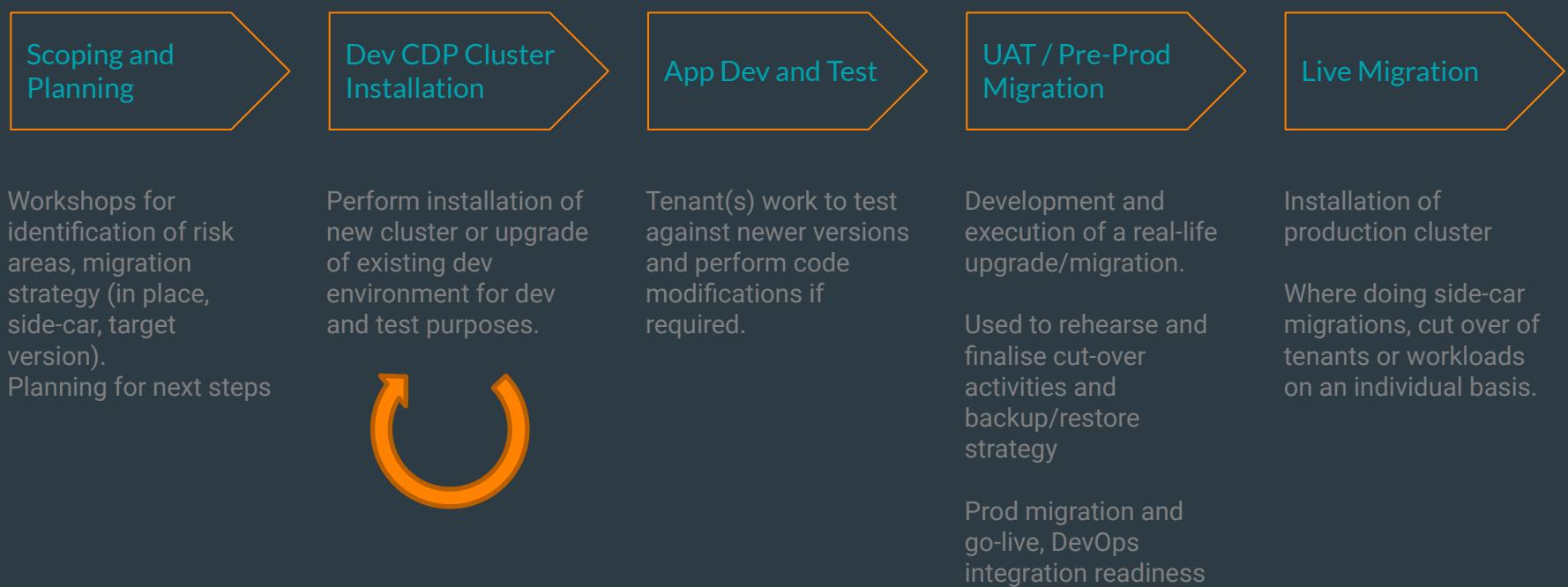


CDP Private  
Experiences



CDP Public Cloud

# MIGRATION PROCESS OVERVIEW



# UPGRADE ADVISOR & MYCLOUDERA UPGRADE ADVISOR

## Early mockups

### CDP Journey Advisor

The interface includes a dropdown for 'I'd like to migrate from HDP' (with options 'Less than 50 nodes' and 'More than 50 nodes') and a section for 'What workloads are you interested in migrating?' with categories 'Data Warehousing' and 'Machine Learning'. Below these are three small thumbnail images representing different workload types.

### MyCloudera.com CDP Journey Advisor

#### Assets

#### What asset would you like to upgrade?

Asset Name	Environment
ADMZ-SIT2A	Staging
ADMZ-AZPROD	Production
ADMZ-VAPROD	Production
ADMZ-PT2	Staging
ADMZ-TXPROD	Production

#### Cluster Analysis

- Recent diagnostic bundle detected
- CDH version 5.16 - No CDH upgrade required
- No unsupported components detected
- Node count over 300

#### Upgrade Recommendation

- (i) Due to the large number of nodes on your cluster, it is recommended to split into 100-300 nodes each.
- CDP Data Center is recommended
- (i) Please review the Helpful Resources to the right for download and installation
- (i) Use your support subscription to open an "Upgrade Planning Case". Your cluster information will be sent automatically to give our support engineers a head start on your upgrade.

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

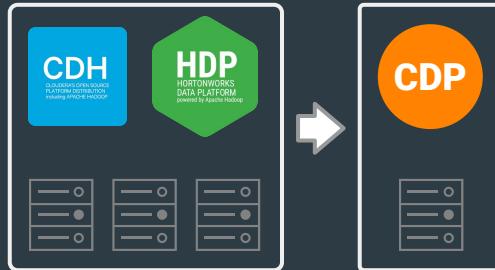
- Provides cluster specific guidance
- Warns for deprecated components
- Provides upgrade prerequisites
- Provides semi-customized PS & Training proposals
- Provides account team contact info

---

# 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:

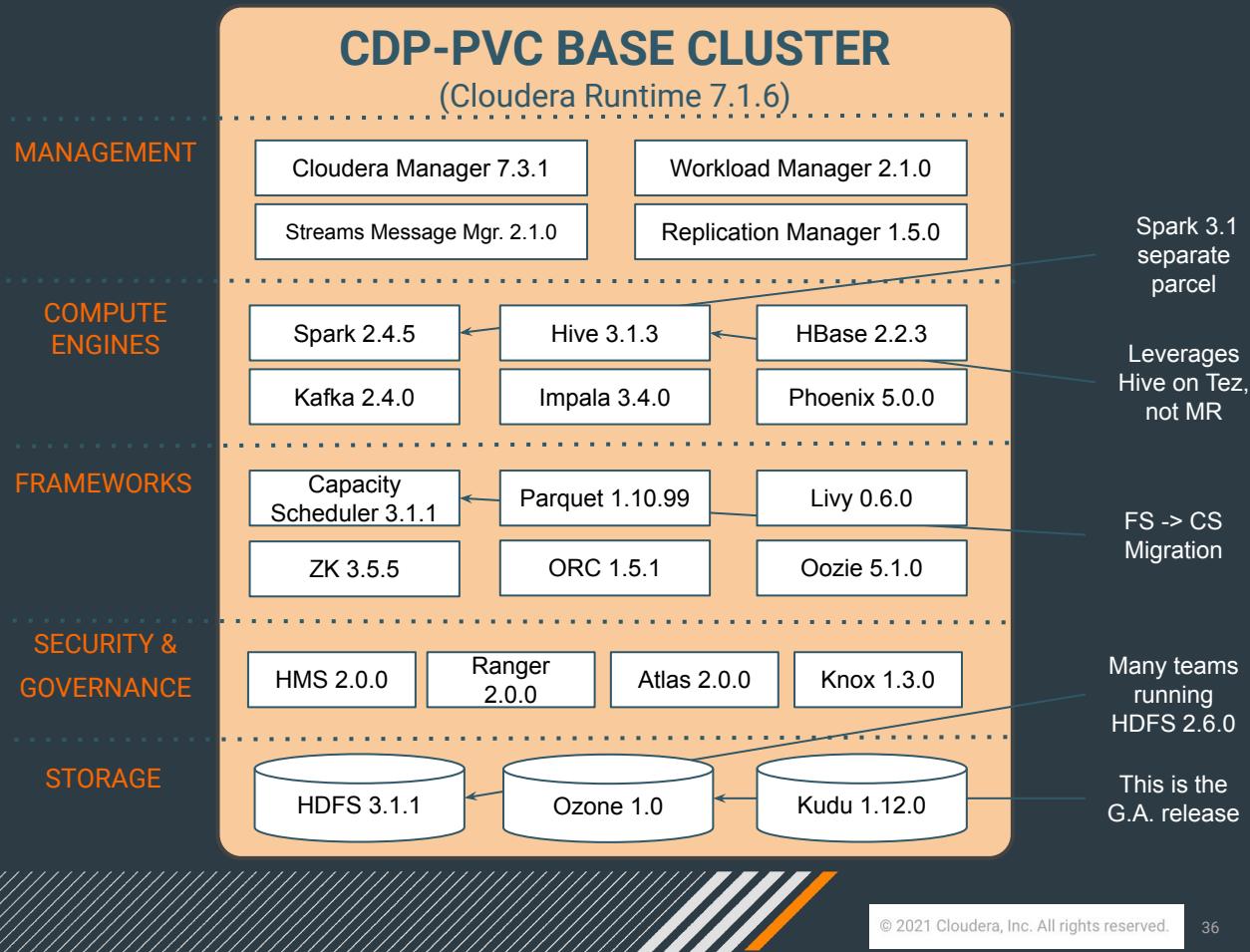
- Customer has limited spare capacity
- Has fewer tenants / integrations
- Can afford a single-cutover with limited rollback options

# CDP RUNTIME 7.1

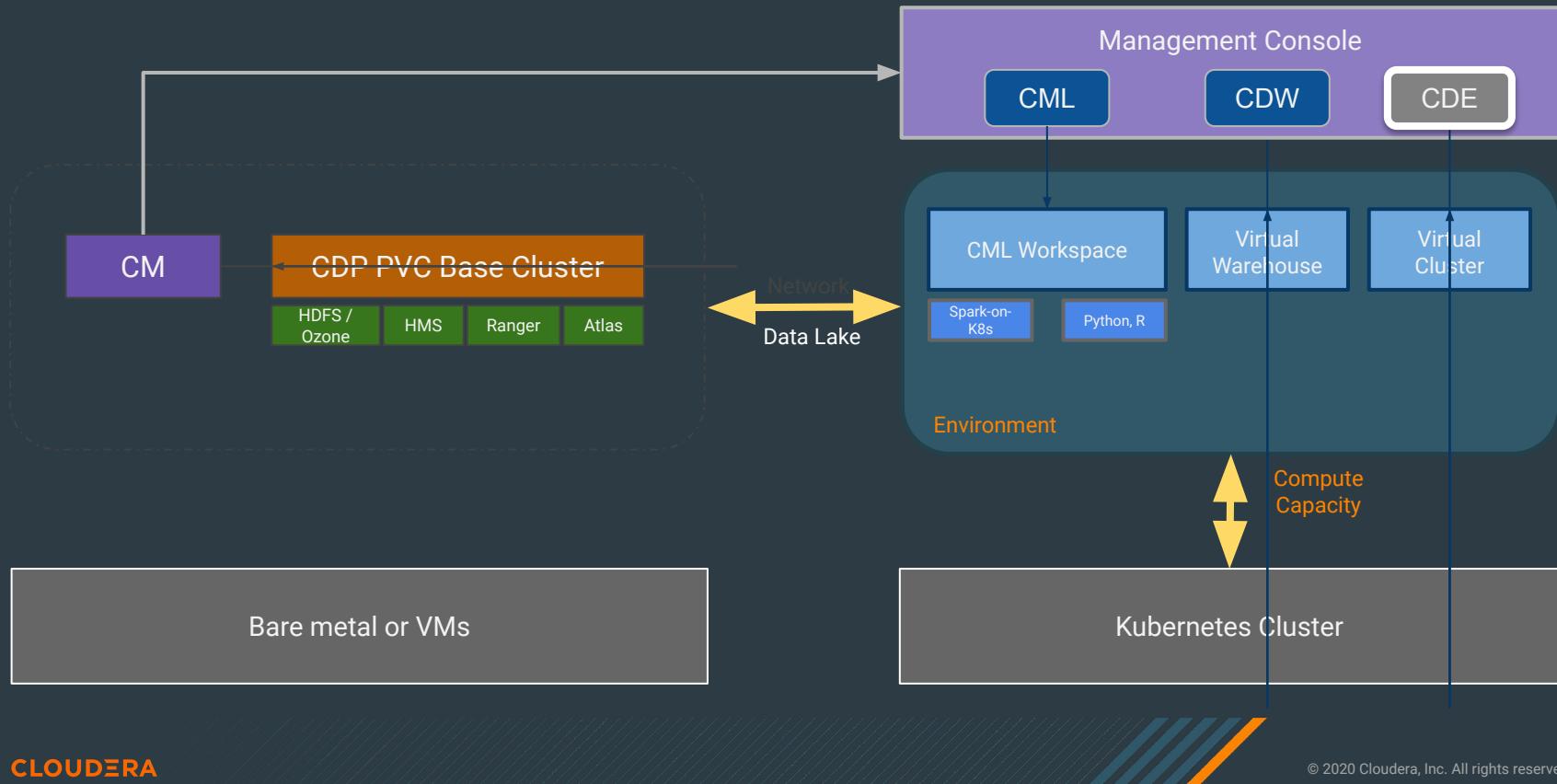
A CDP Private Cloud Base cluster (powered by Cloudera Runtime), can serve as a traditional “**data lake**” (storage & compute) cluster, or as a “**base storage cluster**” (storage only) serving compute workloads running on Kubernetes.

This image shows the component versions in Cloudera Runtime 7.1.6.

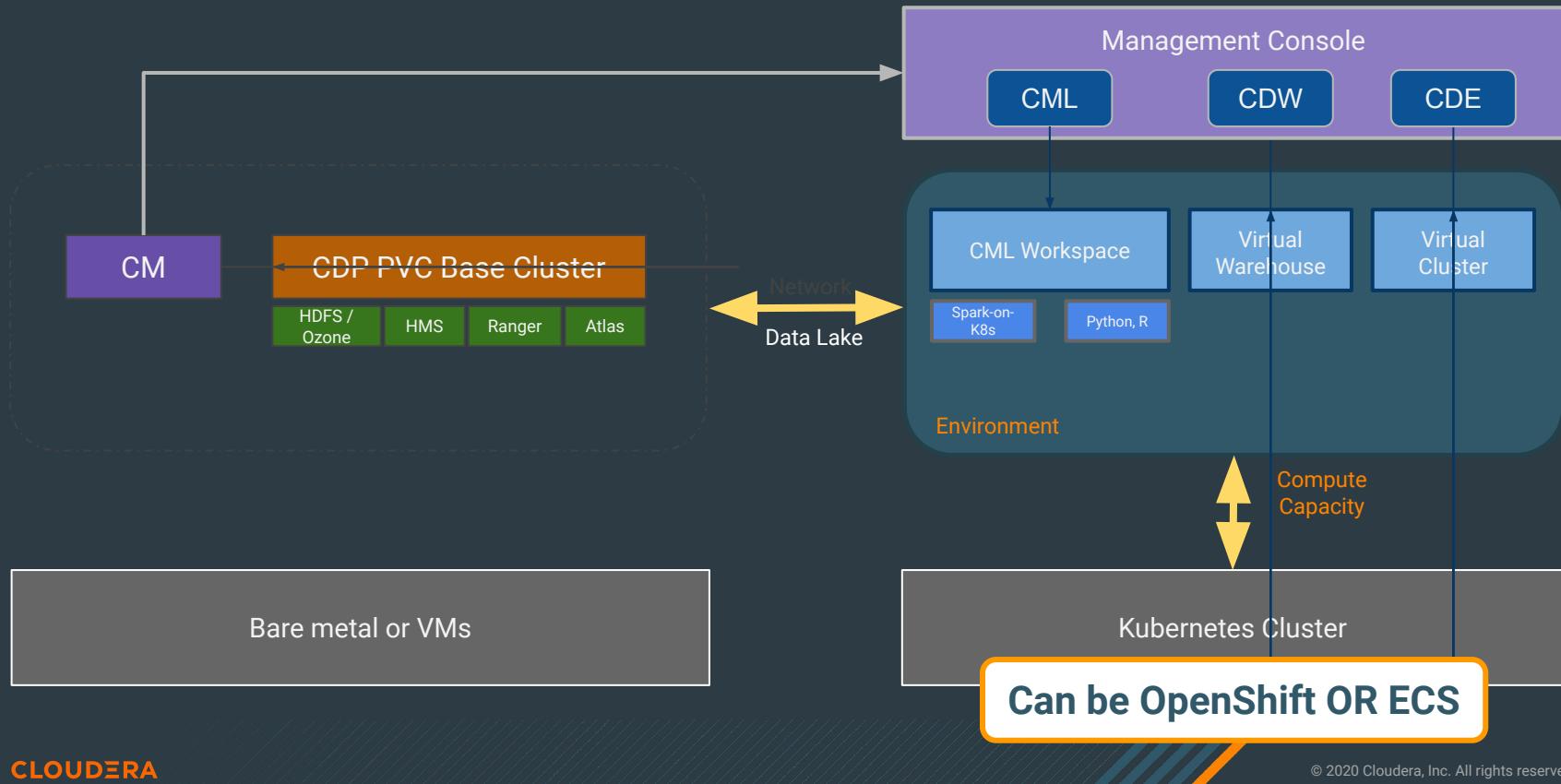
Click [HERE](#) for the complete list of supported components.



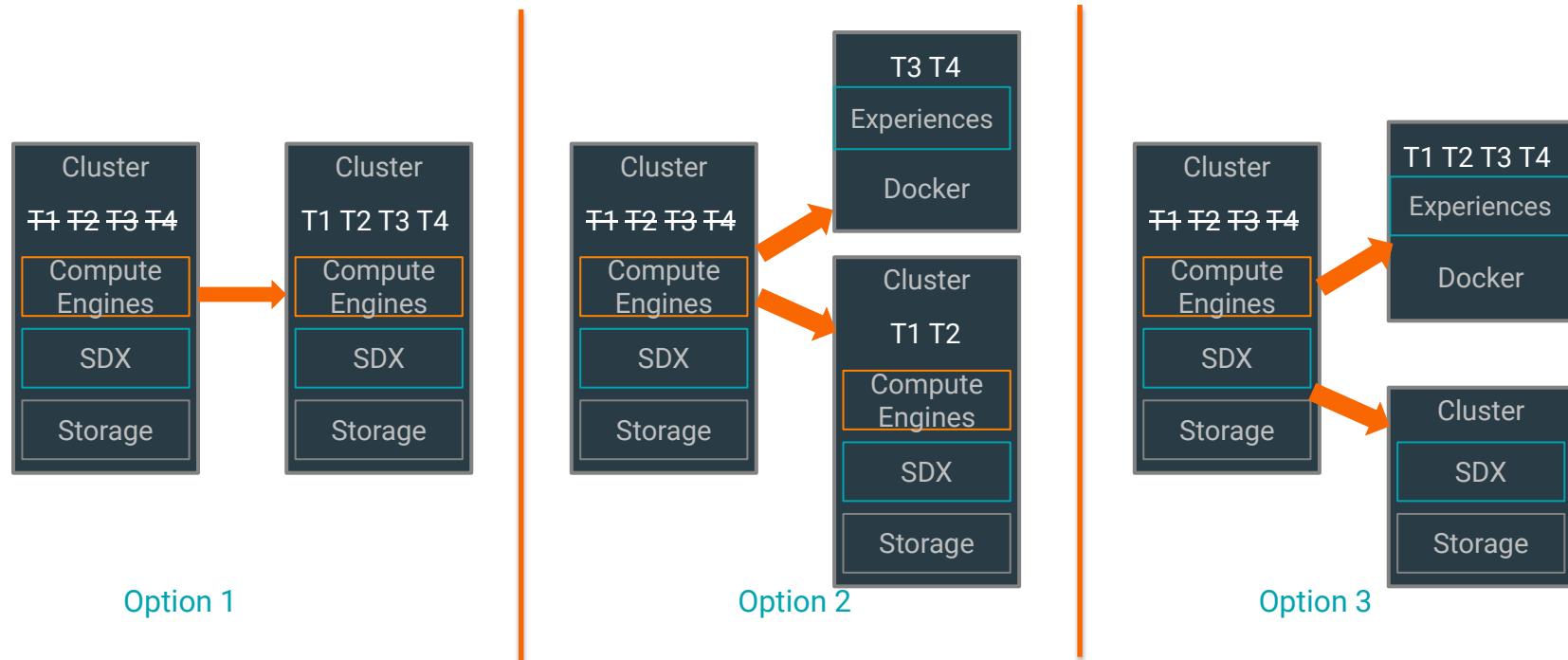
# CDP PRIVATE CLOUD DATA SERVICES ARCHITECTURE



# 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

1

Prerequisites

RedHat OCP/ECS  
Install

2

CDP PVC Base  
Install/Upgrade

3

Data Migration

4

CDP Private  
Cloud Install

5

DW/ML Use Case  
Onboarding

---

# 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

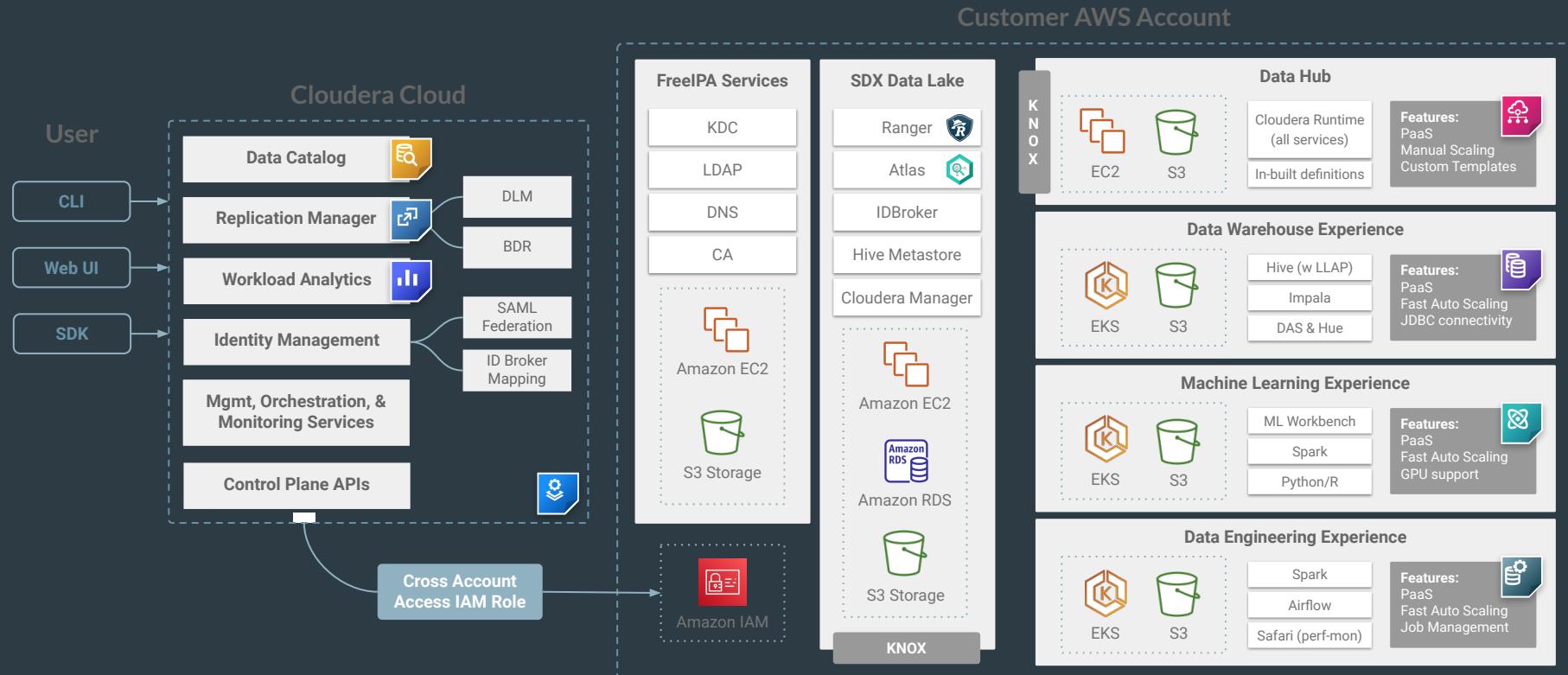


Data  
Lifecycle

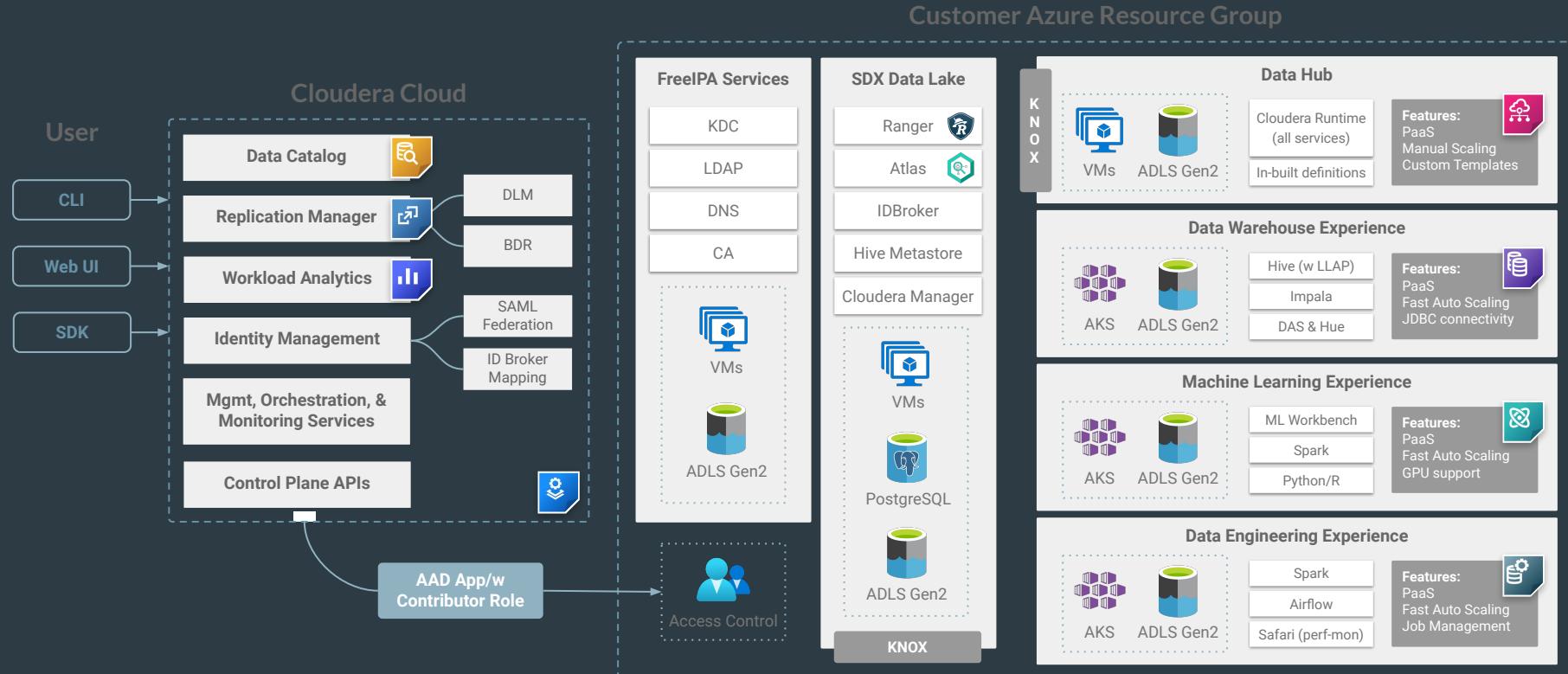


Easy and  
Portable

# CDP - AWS HIGH LEVEL ARCHITECTURE



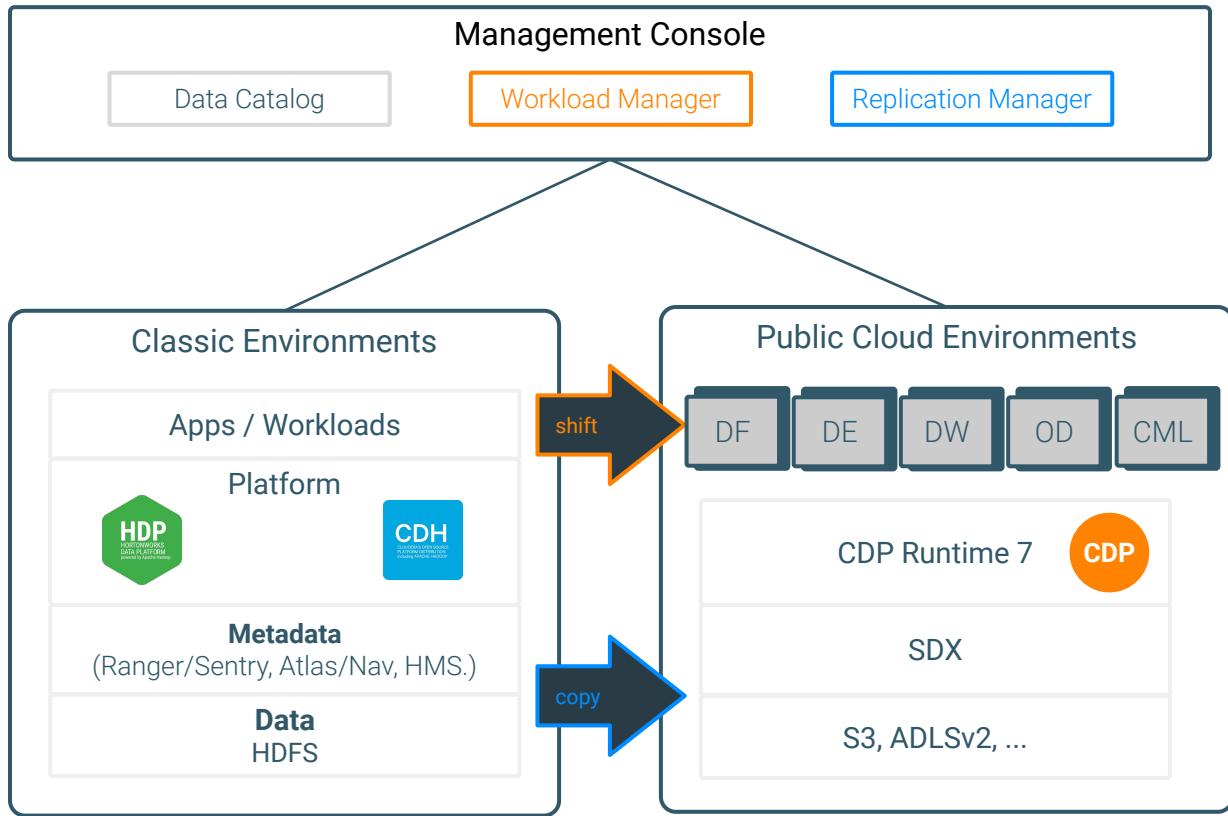
# CDP - AZURE HIGH LEVEL ARCHITECTURE



# MIGRATE TO PUBLIC CLOUD

## Process

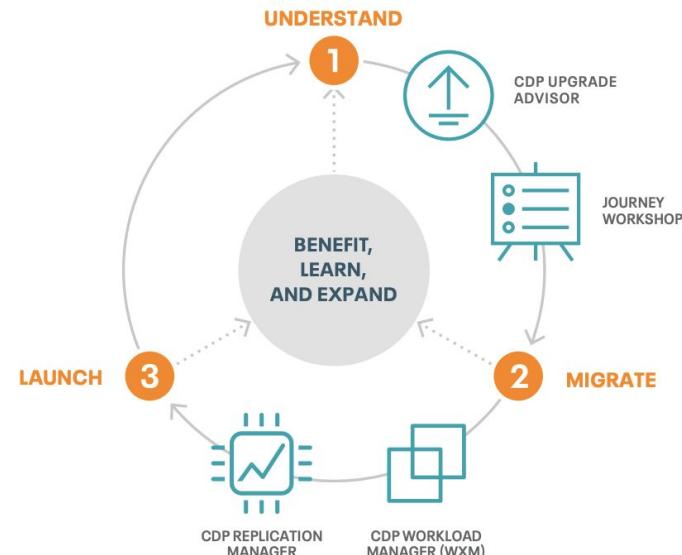
- Set up public cloud environments
- Register classic cluster(s)
- Identify candidate workloads
- Migrate workload data and metadata using Replication Manager ("Burst to Cloud")
- Test and promote to production



# 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.

STEPS TO PUBLIC CLOUD MIGRATION



## 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.

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Kick Off and Health Check	■					
Assessment, Solution & Planning		■				
CDP PC Deploy & Secure			■			
Cloud & CDP Training			■			
Pilot Workload Migration				■		
Demo and Validate					■	
Next Steps Plan & Prod. Proposal						■

## 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.



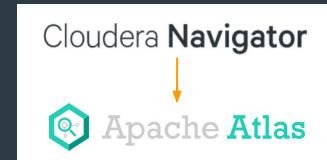
## Sentry to Ranger Migration

- 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

**CLOUDERA** Data Platform X

Environments / List

**Environments** Shared Resources

230 Environments

Search  Change Credential Create Data Hub Register Environment

<input type="checkbox"/> Status	Name	Cloud Provider	Region	Data Lake	Time Created
Available	thorvath-9tnlhj		Texas (US Texas)	Not registered	3/25/2020, 10:50:00 AM PDT
Available	zookepe-tx39ia		Texas (US Texas)	Not registered	3/25/2020, 10:43:38 AM PDT
Available	tmarshall2		US West (Oregon)	Creating Stack...	3/25/2020, 10:43:04 AM PDT
Available	e2e-2329796-xgr-env		US West (Oregon)	Running	3/25/2020, 10:31:32 AM PDT
Freelpa Creation In Progress	sgeetha-ccm-enabled-dev		US West (Oregon)	Provisioning Failed	3/25/2020, 10:05:08 AM PDT
Available	thorvath-2ghouli		Texas (US Texas)	Not registered	3/25/2020, 9:33:42 AM PDT
Creation Failed	bbende-dim-azure		West US 2	Not registered	3/25/2020, 9:06:44 AM PDT
Available	hreeve-azure		West US 2	Running	3/25/2020, 9:04:24 AM PDT

# SINGLE PANE OF GLASS FOR ALL REPLICATION POLICIES

CLOUDERA  
Replication Manager

- Overview
- Classic Clusters
- Cloud Credentials
- Replication Policies

Get Started

Help

N Nachiket Vaidya

## Overview

Classic Clusters

9 Error    9 Active    9 Warning    Total

Policies

- Active    - Suspended    - Unhealthy    3 Total

Jobs

- In Progress    3 Failed Last    17 Failed in Last 10    20 Total

Notifications

No Notification

Classic Clusters

Leaflet

Issues & Updates [?](#)

Job Status	Source	Destination	Service	Policy	Policy History	Runtime	Started	Ended	More
Failed	Cluster 1	→	HDFS	tdk	0	<1m	2d ago	2d ago	⋮
Failed	Cluster 1	→	HDFS	hdfs-zszabo-test-4	0 0 0 0 0 0	<1m	5h ago	5h ago	⋮
Failed	Cluster is not regi...	→	dmx-cudev	Hive	hive-1	<1m	2d ago	2d ago	⋮

Create Policy

1 - 3 of 3    < < > > Items per page: 10

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

1

General

Policy Name **\***  
marketing-policy

Description  
policy that copies marketing database

Type  
 Hive  HDFS  HBase

2

Select Source

Source Cluster **\***  
Select...  
Cluster 1 (ArundCDPPrivate) Cluster 1 (HcubeChennai(CDH)) Cluster 1 (DC01) Cluster 1 (andraspiros-erc2) Cluster 1 (sanJDC01)

3

Select Source

Source Cluster **\***  
Cluster 1 (andraspiros-erc2)

Source Databases and Tables  
marketing

Run As Username (on source)  
hive

4

Select Destination

Type **\***  
Data Lake

Destination Data Lake **\***  
dmx-cudev (dmx-cu)

Warehouse Path  
S3://vijayk-cucoreddev/warehouse/tablespace/managed/hive

Hive External Table Base Directory  
s3://vijayk-cucoreddev/warehouse/tablespace/external/hive

Cloud Credential **\***  
aws AWS-Sales

Validate Policy

5

Select Destination

Type **\***  
Data Lake

Destination Data Lake **\***  
dmx-cudev (dmx-cu)

Warehouse Path  
S3://vijayk-cucoreddev/warehouse/tablespace/managed/hive

Hive External Table Base Directory  
s3://vijayk-cucoreddev/warehouse/tablespace/external/hive

Cloud Credential **\***  
aws AWS-Sales

Validate Policy

6

Additional Settings

YARN Queue Name  
default

Maximum Map Slots  
20

Maximum Bandwidth  
100 MB/s (per mapper)

Other Settings

Sentry Permissions

Include Sentry Permissions with Metadata

Exclude Sentry Permissions from Metadata

Skip URI privileges

Skip URI privileges

Schedule

Start  
 Run Now  Schedule Run Start Time (24-Hour)  
yyyy-mm-dd 00 : 00

Timezone  
(UTC-07:00 PDT) America / Dawson Creek

Repeat  
Does Not Repeat

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

1

General

Policy Name \*

Select Source

Select Destination

Schedule

Additional Settings

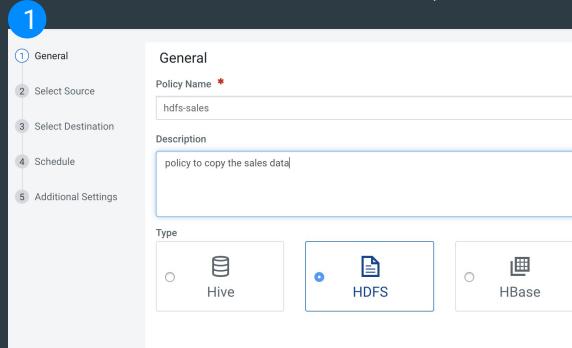
Description

Type

Hive

HDFS

HBase



2

General

Select Source

Select Destination

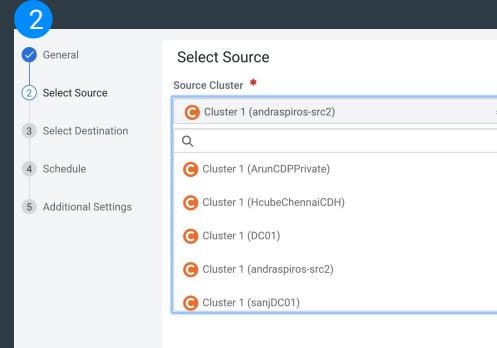
Schedule

Additional Settings

Select Source

Source Cluster \*

- Cluster 1 (andraspiros-src2)
- Cluster 1 (ArunCDPPrivate)
- Cluster 1 (HcubeChennaiCDH)
- Cluster 1 (DC01)
- Cluster 1 (andraspiros-src2)
- Cluster 1 (sanjDC01)



3

General

Select Source

Select Destination

Schedule

Additional Settings

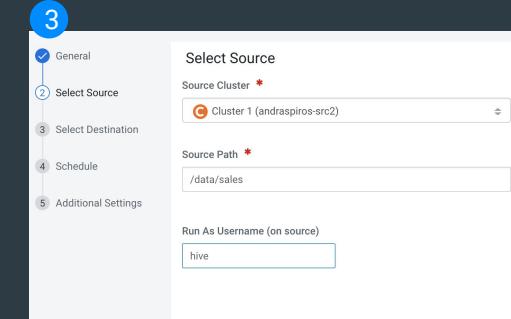
Select Source

Source Cluster \*

- Cluster 1 (andraspiros-src2)

Source Path \*

Run As Username (on source)



4

General

Select Source

Select Destination

Schedule

Additional Settings

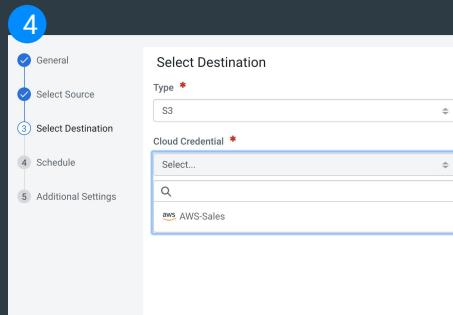
Select Destination

Type \*

- S3

Cloud Credential \*

Select...



5

General

Select Source

Select Destination

Schedule

Additional Settings

Schedule

Start

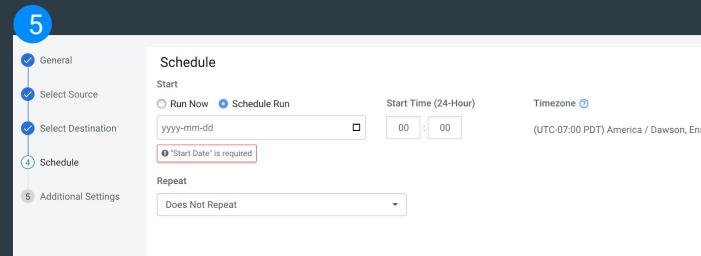
Run Now  Schedule Run

Start Time (24-Hour)  
yyyy-mm-dd   
00 : 00

Timezone

Repeat

Does Not Repeat



6

General

Select Source

Select Destination

Schedule

Additional Settings

YARN Queue Name

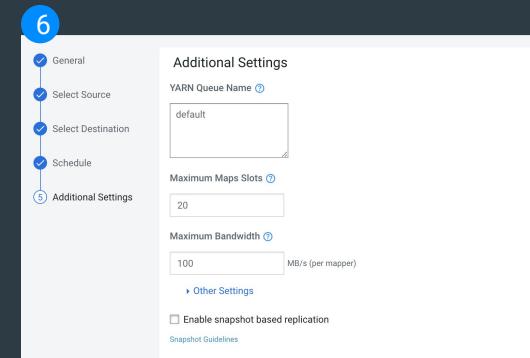
Maximum Maps Slots

Maximum Bandwidth

Other Settings

Enable snapshot based replication

Snapshot Guidelines



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

1

General

Policy Name \* hbase-1

Description Enter a description for the policy

Type

Hive (radio button)

HDFS (radio button)

HBase (radio button, selected)

2

Select Source

Source Cluster \*

Select... Cluster 1 (shailesh-c7-src-1)

3

Select Source

Source Cluster \*

Cluster 1 (shailesh-c7-src-1)

Replication Type

Async Replication Set up an async replication on source namespace and tables to replicate any future data change to destination.

Snapshot Replication Create a new snapshot and replicate selected existing data to destination.

We will automatically restart and configure your source HBASE service to setup an async replication.

Source Namespaces and Tables Namespace Table Name or Regular Expression

Run As Username (on source)

4

Select Destination

Type \* Data Hub

Destination Data Lake \*

Select... od-shailesh (dmx-shailesh)

5

Additional Settings

Maximum Bandwidth 100 MB/s (per mapper)

# ACCELERATE UPGRADES WITH WORKLOAD MANAGER

CLOUDERA Workload Manager

- Summary
- Workloads

ENGINES

- Impala**
- Spark
- Hive
- Oozie
- MapReduce

File Size Report

Help

Raman Rajasekhar

«

### Trend

Count      Concurrency

By Status    By Statement Type

Total Queries: 285K   Failed Queries: 9.8K (3.43%)   Query Active Time > SLA (5s): 18.6K (6.52%)

Total Failed Slow

12 AM 4 AM 8 AM 12 PM 4 PM 8 PM

### Resource Consumption

Average CPU Core Hours: 1.5K per hour   Average Memory Usage: 134.8 GiB·h per hour

CPU Core Hours   Memory Usage

3.9K 2.9K 1.9K 488.3 GiB·h 366.2 GiB·h 244.1 GiB·h

### Queries

Top Queries by: Duration

Status	Query	Duration	User	Pool
Green	insert into u_stephenf.prod...	2h 33m 34s	zs23fce3	root.edh-read...
Green	insert into u_stephenf.prod...	1h 47m 8s	zs23fce3	root.edh-read...
Green	insert into u_stephenf.prod...	1h 43m 46s	zs23fce3	root.edh-read...
Red	SELECT SUM(1) AS `q_nlp...	1h 36m 59s	pd2djhj6	root.pd2djhj6
Green	CREATE TABLE sfdc_stagin...	1h 29m 27s	vd3oddrl	root.default
Green	CREATE TABLE cldr_enrich...	1h 10m 47s	rl8dkej5	root.edh-read...
Green	CREATE TABLE support_st...	1h 9m 38s	zs23fce3	root.edh-read...
Green	SELECT `ira_w_detail` chan...	1h 8m 54s	pd2djhj6	root.pd2djhj6

Keepalive

### Workloads

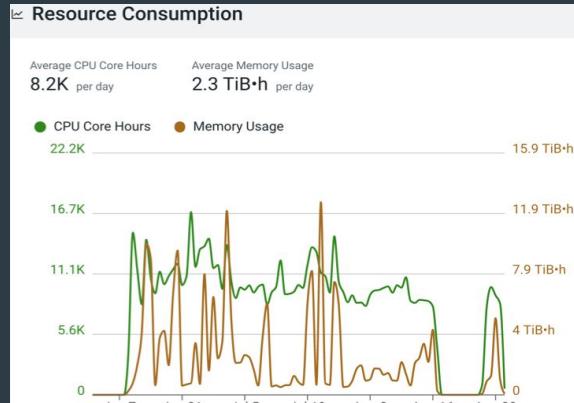
Status	Workload	Missed SLA %	Failure %	Total Queries
Red	TB-Table_wf	100%	13%	24
Green	dcxa_cops_impala	96%	4%	222K
Red	dcxa_Impala	71%	31%	27.9K
Red	Bill_Ad-Hoc_View	68%	0%	256
Red	ETL	37%	1%	2.8K

CLOUDERA

© 2020 Cloudera, Inc. All rights reserved.

56

## 01 - Analyze Resource Consumption



## 02 | 1 - Identify Rogue Users

### Usage Analysis

Users   Pools   Databases

Top Users by: # Queries

User	# Queries	%
zxLKfheu1	175.8K	67%
fkjh3fk45	24.7K	9%
9f8h3dd6	14.4K	6%
zv3dckh4	12K	5%
w38hckj3	7.8K	3%
3xksle8z	7.4K	3%

## 02 | 2 - Identify Rogue Users

### Usage Analysis

Users   Pools   Databases

Top Users by: CPU

User	CPU
3xksle8z	647h 5m 2s
zv3dckh4	445h 9m 31s
zxLKfheu1	386h 9m 46s
9f8h3dd6	177h 32m 8s
w38hckj3	112h 26s
fkjh3fk45	104h 9m 31s

## 02 | 3 - Identify Rogue Users

### Usage Analysis

Users   Pools   Databases

Top Users by: Memory

User	Memory
3xksle8z	5.9 PiB × s
fkjh3fk45	2.9 PiB × s
9f8h3dd6	355.4 TiB × s
zv3dckh4	87.6 TiB × s
w38hckj3	82.6 TiB × s
zxLKfheu1	68 TiB × s

## 04 - Analyze Rogue User's Queries / Jobs

Summary   Trend    Succeeded: 12/11/2020 4:21 AM PDT    User: 3xksle8z    DML: root.default

Joins: 18   Duration: 34h 47m 29s   Rows Produced: 38834350   Aggregate CPU Usage: 212h 16m 47s   Aggregate Memory Usage: 859.8 TiB × s   Peak Memory Usage: 19.9 GiB

Basic   Operators   Hosts   HDFS Tables Scanned

SQL Statement

```
1: insert
2:   overwrite flex_web.aem_dcxa_visitor_events
3: select
4:   concat (
5:     cast(a.post_visid_high as string),
6:     '*****',
7:     cast(a.post_visid_low as string)
8:   ) as visitor_id,
9:   a.date_time as visit_time,
10:  a.visit_num as visit_number,
11:  a.visit_page_num as visit_page_number,
```

Expand

Performance Issues

Potential SQL Issues   Health Check Violations

Optimal Configuration

Aggregation Spilled Partitions: 861

+ Details

Metadata/Statistics

Missing Table Statistics: 14 tables

## 05 | 1 - Plan Optimization Pre/Post Upgrade

### ④ Performance Issues

#### Potential SQL Issues 2

#### Health Check Violations 2

##### High Risks

- ⚠ >=5 table joins or >=10 join conditions found.  
Denormalize tables to eliminate need for joins.

##### Medium Risks

- ⚠ >=10 columns present in GROUP BY list.  
Evaluate memory requirements for the query.

## 05 | 2 - Plan Optimization Pre/Post Upgrade

### Diagnosis

This stage had **poor parallelization** as 3 (out of 27) tasks took abnormal amount of time to finish.



### Recommendation

The following are the possible action items to resolve this:

- If "Task Input Data" health check is also failing then try to partition input data on a different set of partition keys so that input data is more uniformly distributed.
- Try increasing RDD/Dataframe partitions using `repartition` function. However, increasing the partitions may not work if relatively fewer keys are dominant in the data.
- If the job contain joins and one of the join table is small in size, try increasing `spark.sql.autoBroadcastJoinThreshold`. This will increase the likelihood of Spark engine choosing broadcast join over short merge join eliminating the shuffle all together.
- Caution:** While increasing `spark.sql.autoBroadcastJoinThreshold` value, make sure that driver and executors are having sufficient memory to hold broadcasted table.
- If the job contain joins and relatively fewer keys are dominant in the data then key salting should help distribute the join data in a more uniform manner.

## 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 Statistic
12/07/2020 1:03 PM PST	✓ Succeed...	7m 57s	tserver	Corrupt Table Statistic
12/07/2020 9:03 AM PST	✓ Succeed...	3m 37s	tserver	Corrupt Table Statistic
12/07/2020 5:02 AM PST	✓ Succeed...	8m 14s	tserver	Corrupt Table Statistic
12/07/2020 1:11 AM PST	✓ Succeed...	3m 7s	tserver	Corrupt Table Statistic

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

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

---

# 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
<b>Flume</b>	Port to CFM (Apache NiFi)	No
<b>Storm</b>	Port to CSA (Apache Flink)	No
<b>Sqoop2</b>	Port to Sqoop 1 (or Apache NiFi)	No
<b>Pig</b>	Port to Apache Spark	No
<b>Crunch</b>	Port to Apache Spark	No
<b>YARN Fair Scheduler</b>	Configure Capacity Scheduler in CM UI	Yes
<b>Sentry</b>	Migrate to Ranger	Yes
<b>Navigator</b>	Migrate to Atlas	Yes
<b>Ambari</b>	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	<ul style="list-style-type: none"><li>WXM (workload centric approach) or Replication manager</li></ul>
	Druid	<ul style="list-style-type: none"><li>Manual migration to Public Cloud (only)</li></ul>
	SOLR	<ul style="list-style-type: none"><li>Backup your data and schema etc.</li><li>Use the scanner to figure out what in your schema needs to change.</li><li>Use Replication Manager to copy your data to CDP cluster.</li><li>Redesign your schema for Solr 8 and reindex using MRindexer</li></ul>
	HUE	<ul style="list-style-type: none"><li>Export &amp; import the Hue database</li></ul>
	Kudu	<ul style="list-style-type: none"><li>Kudu backup tool to backup &amp; restore to new cluster. Manually migrate configs</li></ul>
	Hive LLAP	<ul style="list-style-type: none"><li>Public Cloud and Private Cloud</li></ul>
OP DB	HBase Phoenix	<ul style="list-style-type: none"><li>Replication manager to migrate data to CDP Public Cloud or HBase native replication to migrate data to CDP D/C</li><li>Migrate configurations manually (&amp; retune)</li><li>Manually migrate ACLs to Ranger or continue to use native HBase ACLs</li></ul>
	Accumulo	<ul style="list-style-type: none"><li>n/a – Will be supported in CDP DC 7.2</li></ul>

# 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	<ul style="list-style-type: none"><li>• Replication Manager will convert policies and migrate policies to Ranger as part of Hive/Impala replication (for CDP-PC)</li><li>• Kafka/Solr permissions need to be manually converted into Ranger policies</li><li>• HDFS ACLs which are automatically set-up by Sentry will need to be manually converted into Ranger policies</li></ul>
Navigator	<ul style="list-style-type: none"><li>• Navigator Managed metadata tags &amp; any manually ported to Atlas Business Metadata Tags</li><li>• Audit data kept in Read Only mode in CM until expired</li></ul>
Ranger	<ul style="list-style-type: none"><li>• Ranger Policy Import/Export feature can be used to migrate existing Ranger policy into CDP-PC and/or CDP-DC</li></ul>
Ranger KMS	<ul style="list-style-type: none"><li>• Use distcp to copy data into Cloud Native encrypted storage (CDP-PC) or into another HDFS encryption zone (CDP-DC)</li><li>• Data re-encryption will take place during copy</li></ul>
Atlas	<ul style="list-style-type: none"><li>• Managed In place upgrade to CDP-DC from CDH (as of 7.1) / HDP(roadmap)</li><li>• CDP Public cloud will have Atlas wired up to all workloads, and replication manager and porting of jobs will recreate lineage. ?</li><li>• Atlas Import / Export can be used to port legacy atlas data to new deploy.</li></ul>
Key Trustee Server Key Trustee KMS Key HSM HSM KMS	<ul style="list-style-type: none"><li>• Use BDR/Replication Manager to migrate encrypted data to Public cloud or to CDP-DC</li></ul>
NavEncrypt	<ul style="list-style-type: none"><li>• Migrate data from encrypted volumes to Cloud native encrypted storage (for CDP-PC) or to another NavEncrypt encrypted volume (in CDP-DC)</li><li>• Data re-encryption will take place during migration</li></ul>
Knox	<ul style="list-style-type: none"><li>• n/a – No migration required</li></ul>

# 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

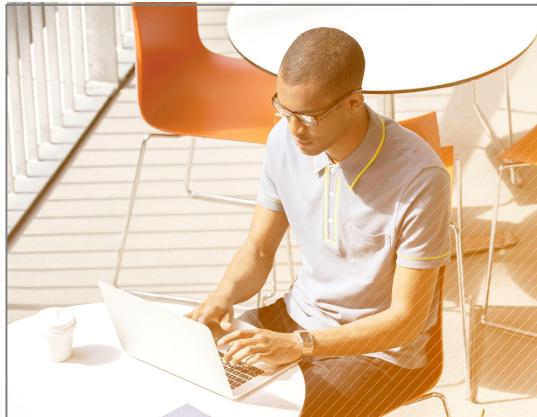
# OVERALL IDEA OF MIGRATION UPGRADE

- 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**<sup>®</sup>

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 EDUCATIONAL SERVICES

## Supporting Customers' Move to the Enterprise Data Cloud



[cloudera.com/training.html](http://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>

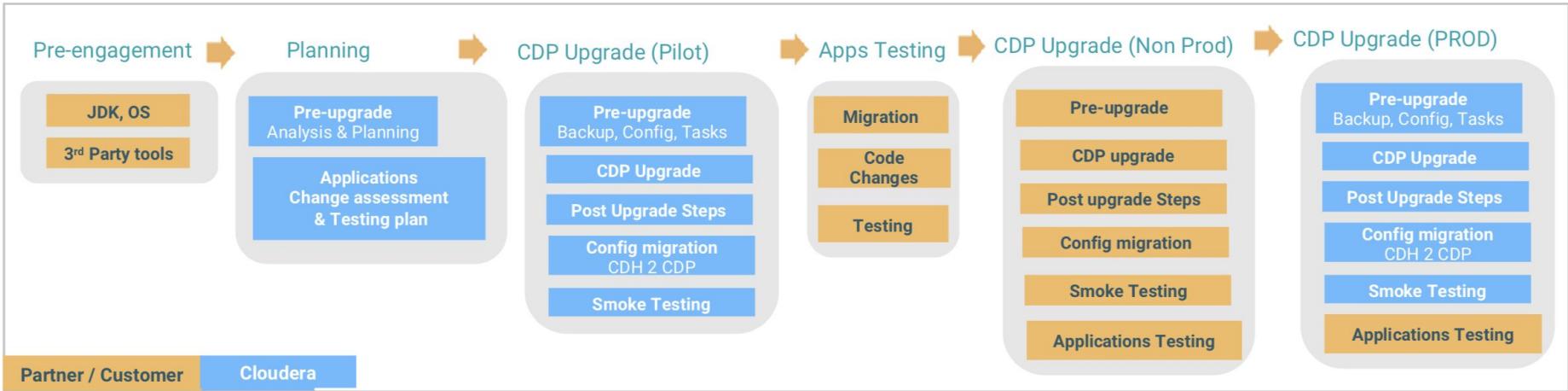
---

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