CLOUDERA

CDP PRIVATE CLOUD Overview

SkillUP Series

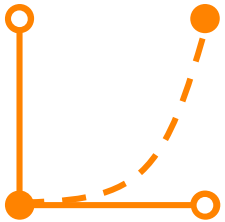
Chakra J, Private Cloud Specialist

AGENDA

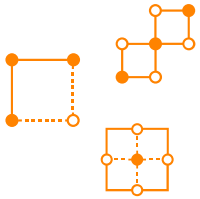
- *Why/What/How of CDP Private Cloud (PvC)*
- *PvC Architecture*
- *Running a POC*
  - *Minimum hardware requirements*
  - *Prerequisites*
  - *Setup process*
- *PvC Use cases*
- *Customer success*

HARD TO KEEP UP MANAGING ENTERPRISE DATA

And getting harder all the time

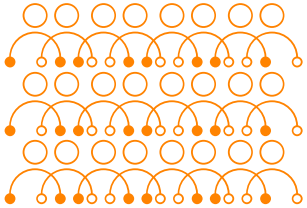


Data Growth



Data Variety

Analytics



Tenant Growth

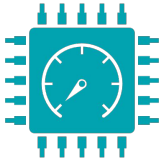
IMPACT OF CHALLENGES IN EXISTING SOLUTIONS

Noisy neighbors

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

Complex upgrades

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



Lack of elasticity

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

Cluster sprawl

"I don't have time to work on new use cases, I'm just focused on keeping the light on"

Time to value

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

Decouple Data and Compute

CDP Private cloud

ANSWER: MODERNIZATION IN-PLACE TO DATA SERVICES

Practitioner-focused data services for building & operating multi-function data applications

DATA CLUSTERS

SPARK, HIVE, IMPALA, CDSW

- Servers
- Monolithic
- Co-Located Storage & Compute
- HW Dependent
- Operator Focused
- Optimized for Existing Applications
- Forklift Upgrades
- Static Workloads

DATA SERVICES

- Services
- Modular
- Separated Storage & Compute
- SW Defined
- Practitioner Focused
- Optimized for New Applications
- Independent Upgrades
- Portable Workloads

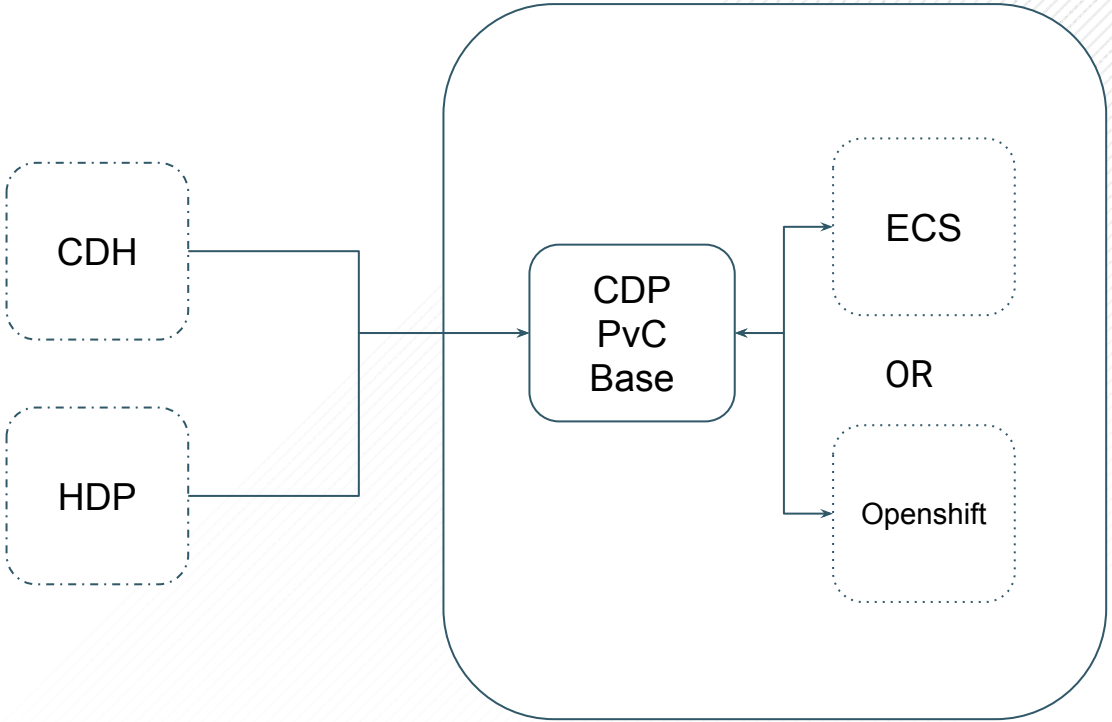
CDP Private Cloud

*An **enterprise data platform** with next generation cloud-native hybrid data architecture, enabling **on-premises deployments** with the agility, flexibility and cost-efficiency of cloud architectures*

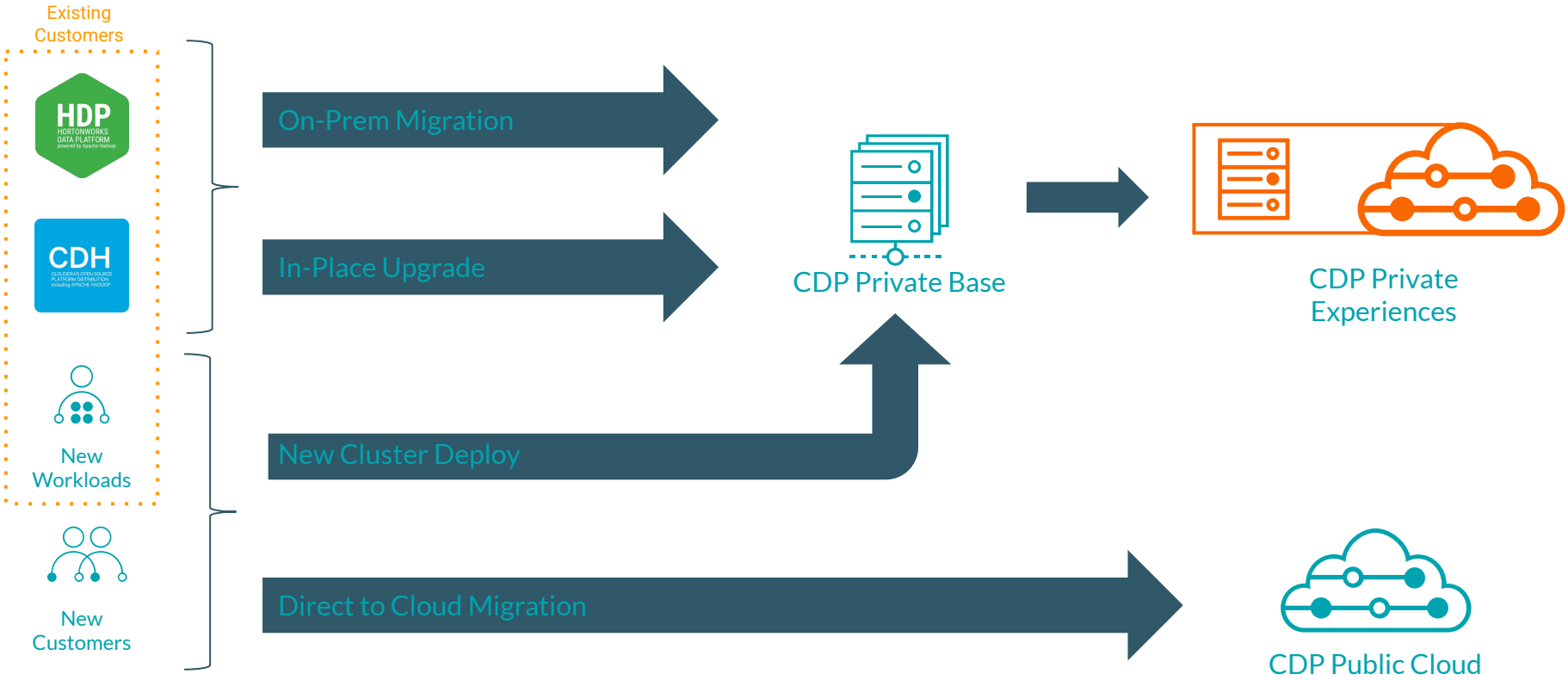
Acronyms… Acronyms..

- CDP - Cloudera Data Platform
- CDP PC - CDP Public Cloud i.e. CDP on AWS/GCP/Azure
- CDP PvC - CDP Private Cloud i.e. CDP on premise
- CDP PvC Base - On premise data lake / Security / Governance
- CDP PvC Experiences - CDE/DWX/CDE on Kubernetes
- CDP PvC ECS - Cloudera provided Kubernetes for running experiences
- CML - Cloudera Machine Learning on containers
- DWX - Cloudera Data Warehouse / hive / impala on containers
- CDE - Apache airflow integrated into CDP on containers

Private Cloud Options

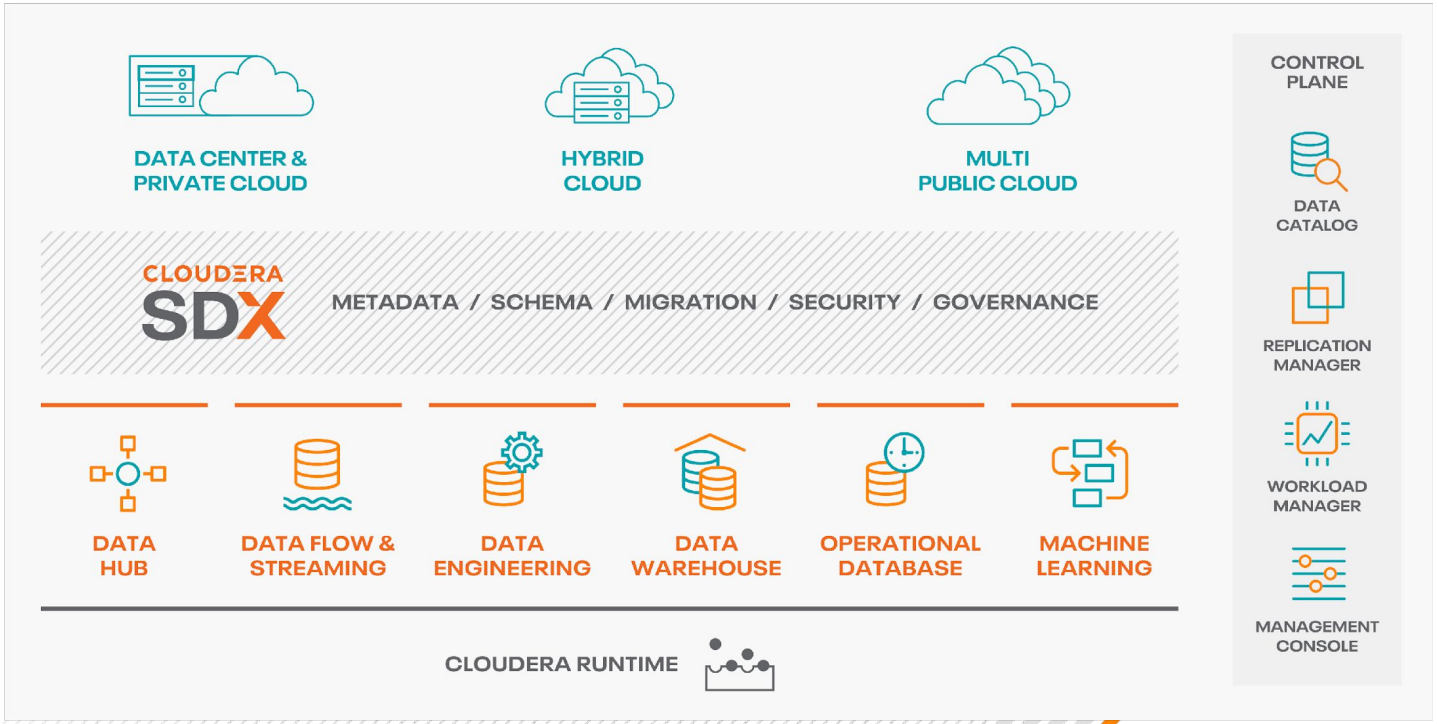


Upgrading & consuming CDP Experiences (path)

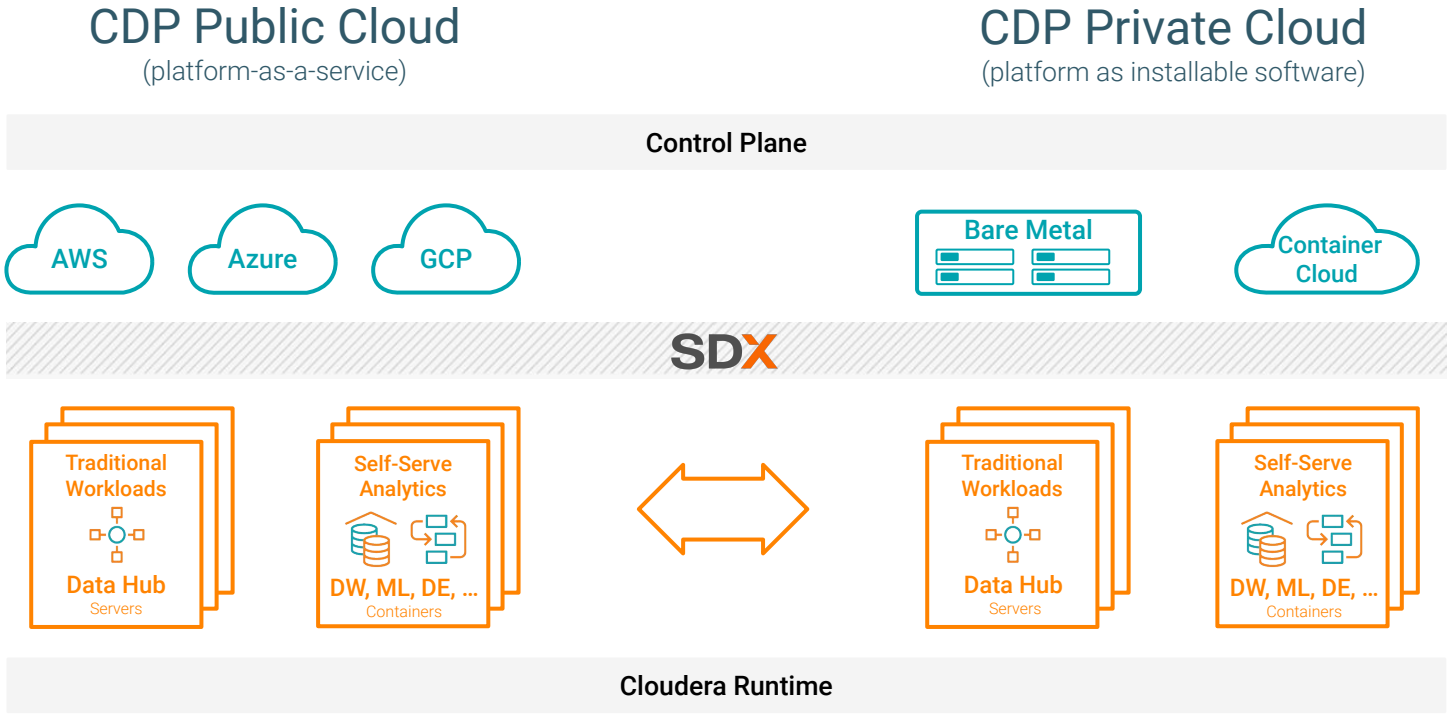


CLOUDERA DATA PLATFORM

World's first enterprise data cloud

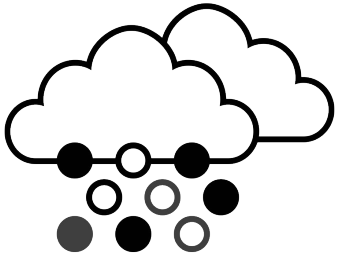


ONE PLATFORM – TWO FORM FACTORS



CLOUDERA

THE ENTERPRISE DATA CLOUD COMPANY

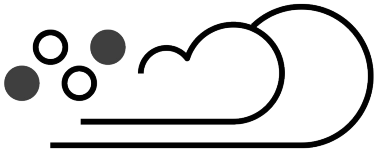


Any Cloud

Data Lifecycle

CLOUDERA SDX

Secure & Governed



Open

CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter

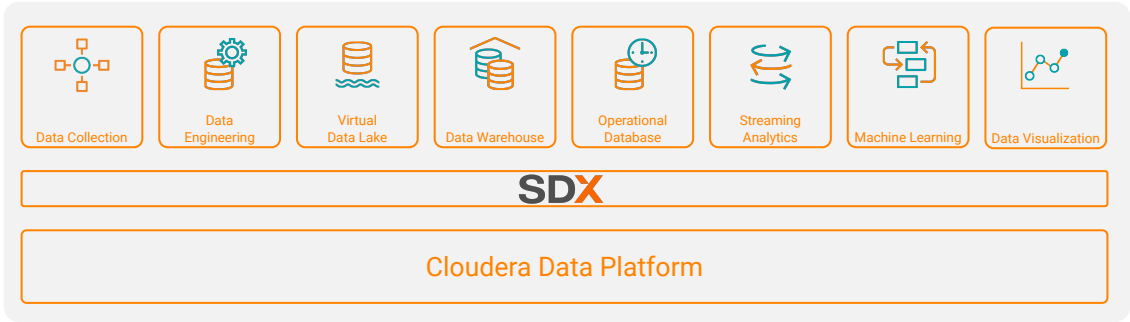
CLOUDERA SDX

SECURITY | GOVERNANCE | INTELLIGENCE | METADATA | CATALOG

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

kubernetes
aws Google Cloud Azure Red Hat

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

CLOUDERA SDX

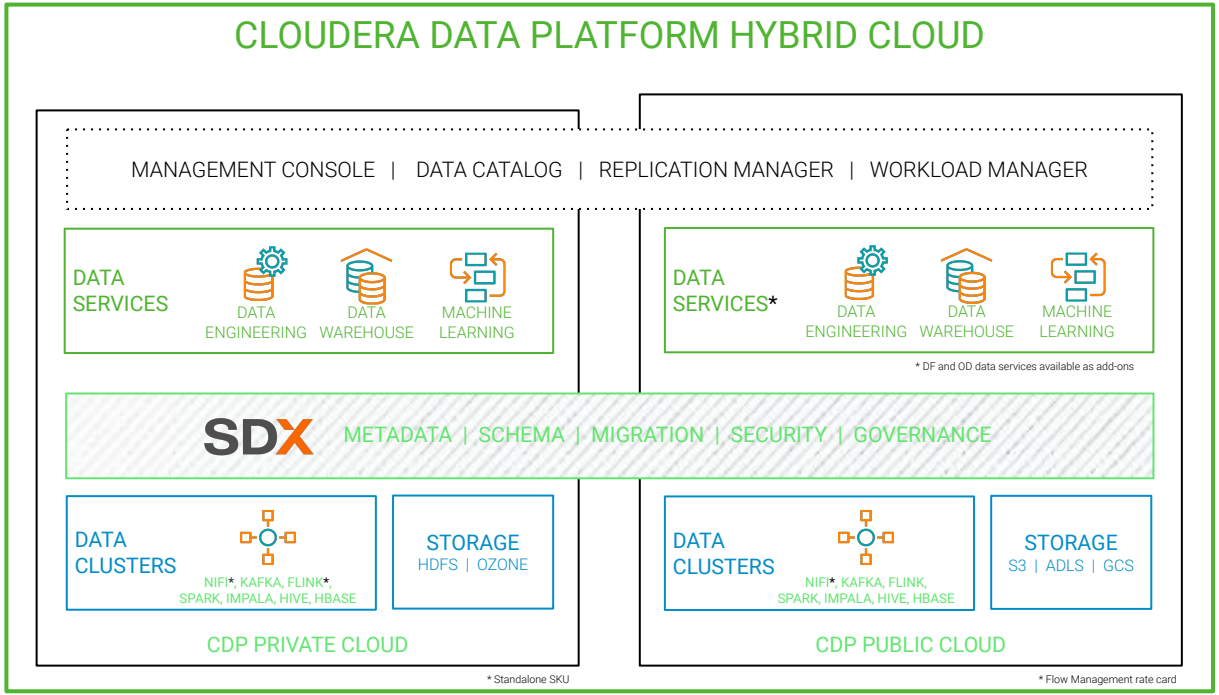
Secure & Governed
to simplify data protection, sharing and compliance

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

CDP HYBRID CLOUD

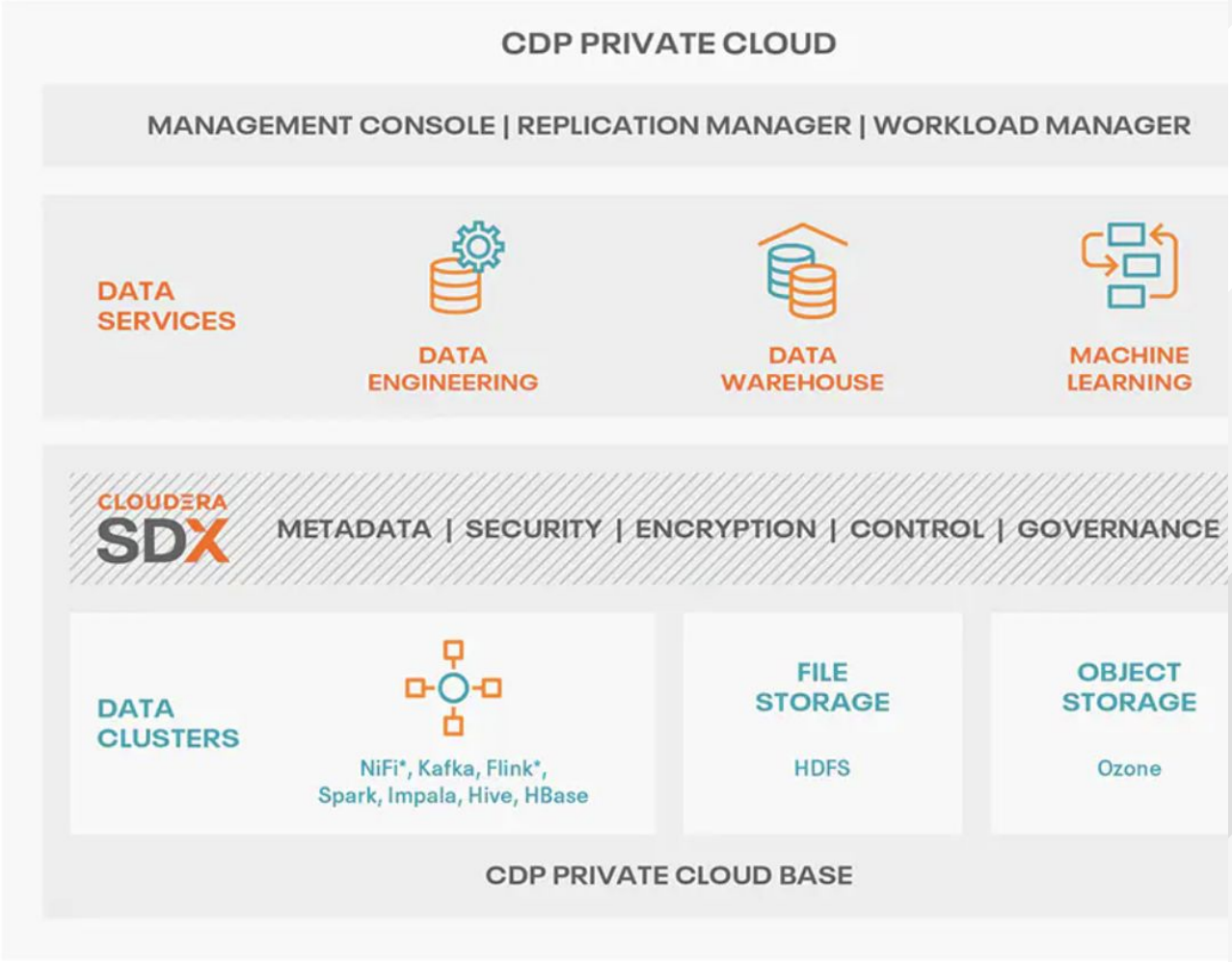
Consistency and flexibility for data analytics across private & public clouds

- Traditional data clusters & purpose-built data services
- Easy to manage and with more control
- Consistent security & governance across environments

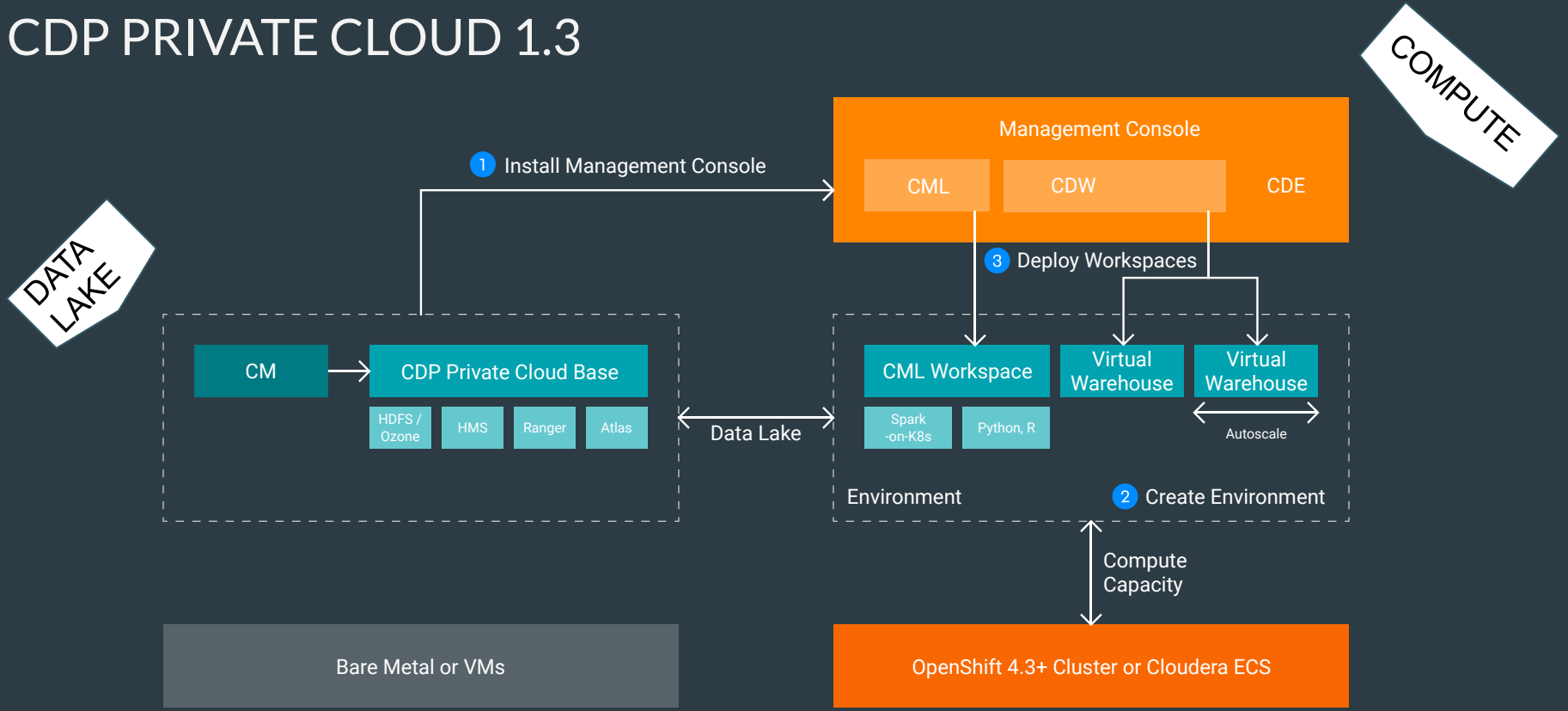


CDP PRIVATE CLOUD

CDP Private Cloud Components



CDP PRIVATE CLOUD 1.3



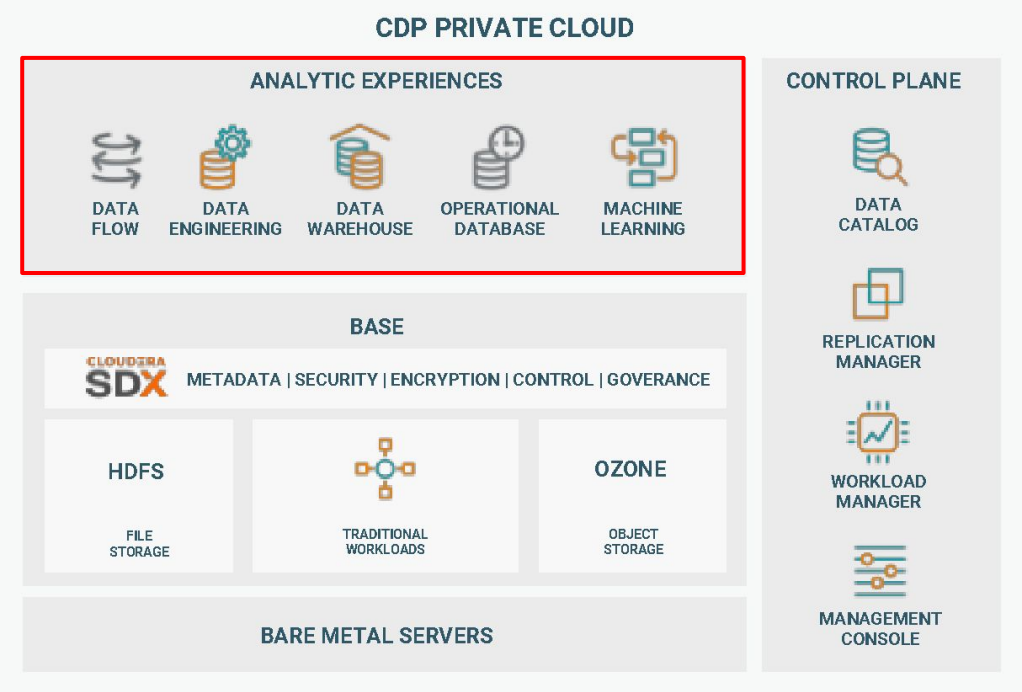
A DEEPER LOOK AT CDP PRIVATE CLOUD

Predictable workload performance with tenant isolation

PREDICTABLE

Avoid Noisy Neighbors

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



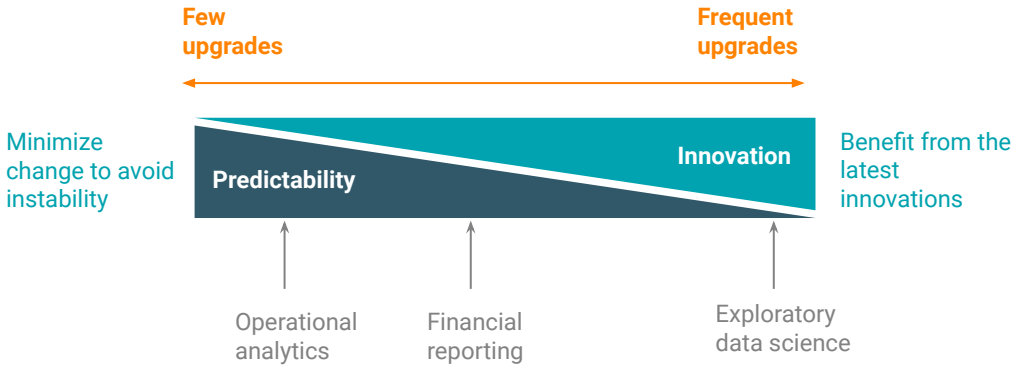
A DEEPER LOOK AT CDP PRIVATE CLOUD

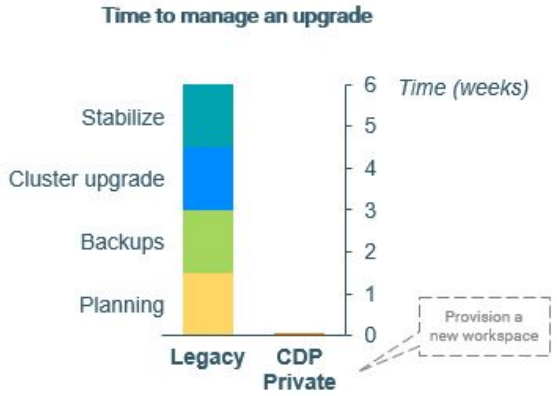
Better multi-tenancy for ‘upgrade agility’

PREDICTABLE

Independent Upgrades

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





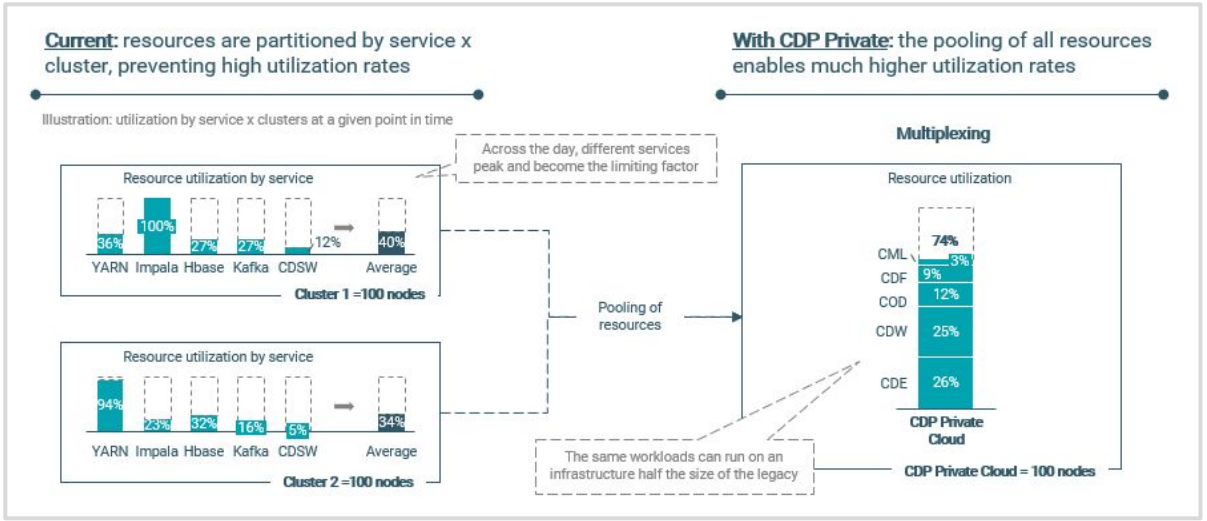
A DEEPER LOOK AT CDP PRIVATE CLOUD

Performance without replicating data or creating silos

COST EFFICIENT

Consolidate Your Clusters

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



A DEEPER LOOK AT CDP PRIVATE CLOUD

Improve cost efficiency with better infrastructure utilization

COST EFFICIENT

Elasticity to Autoscale & Auto-Suspend

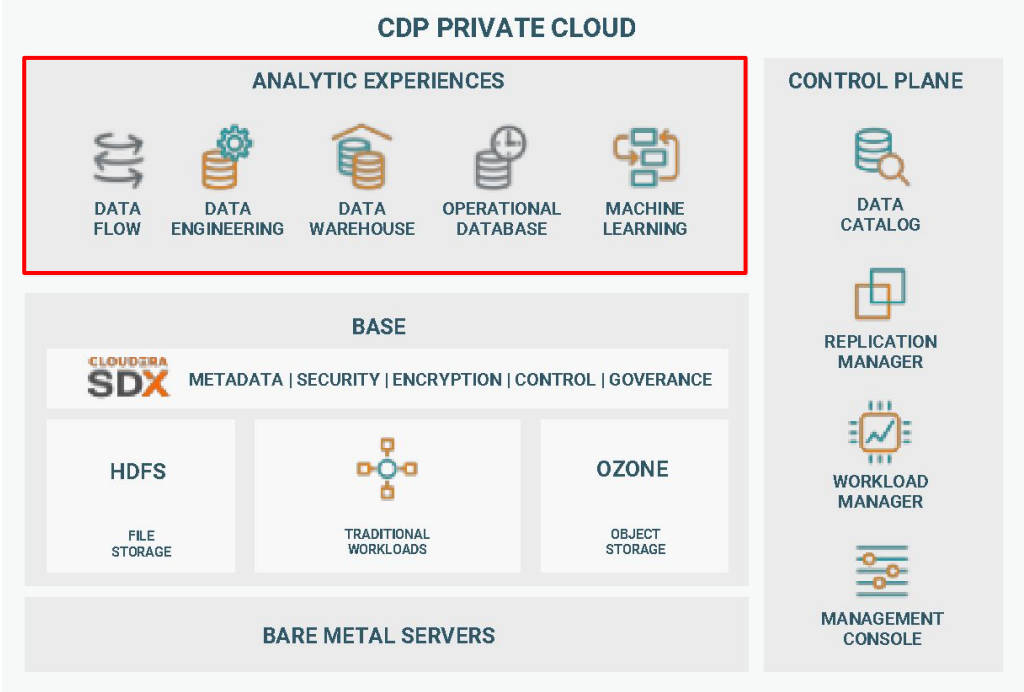
- Use what you need, when you need it
- Shift excess capacity according to demand

Containerized Compute Platform

- Workloads are abstracted from physical infrastructure
- Decoupled compute and storage

Additional Features Coming *

- Share infrastructure with non-Cloudera apps
- Quota management to set mins and max per tenants



A DEEPER LOOK AT CDP PRIVATE CLOUD

Faster time-to-value with simplified onboarding

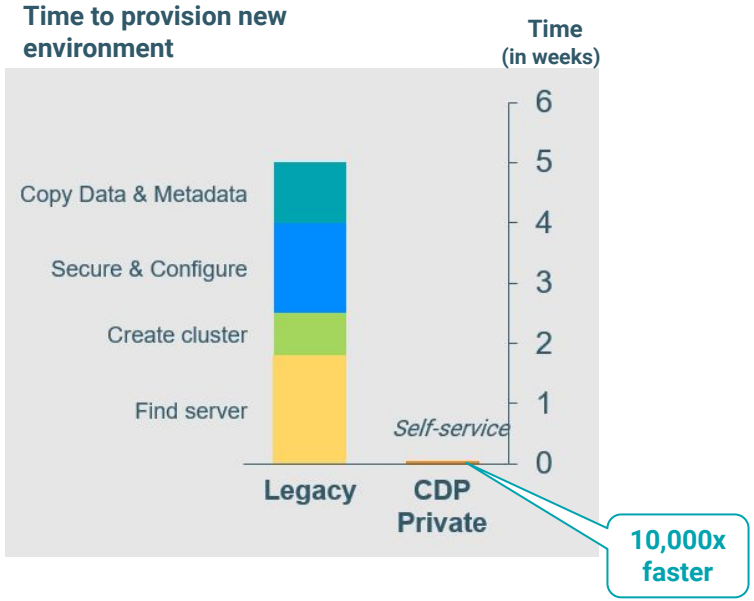
RAPID

Push-button Provisioning

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

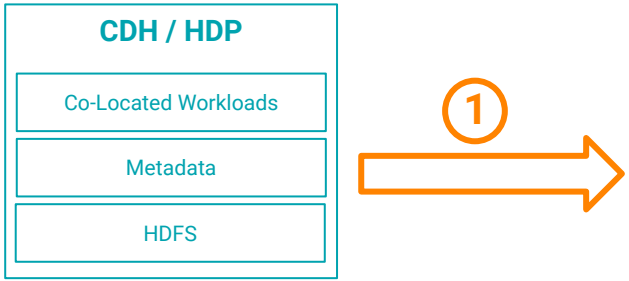
Redesigned User Interfaces

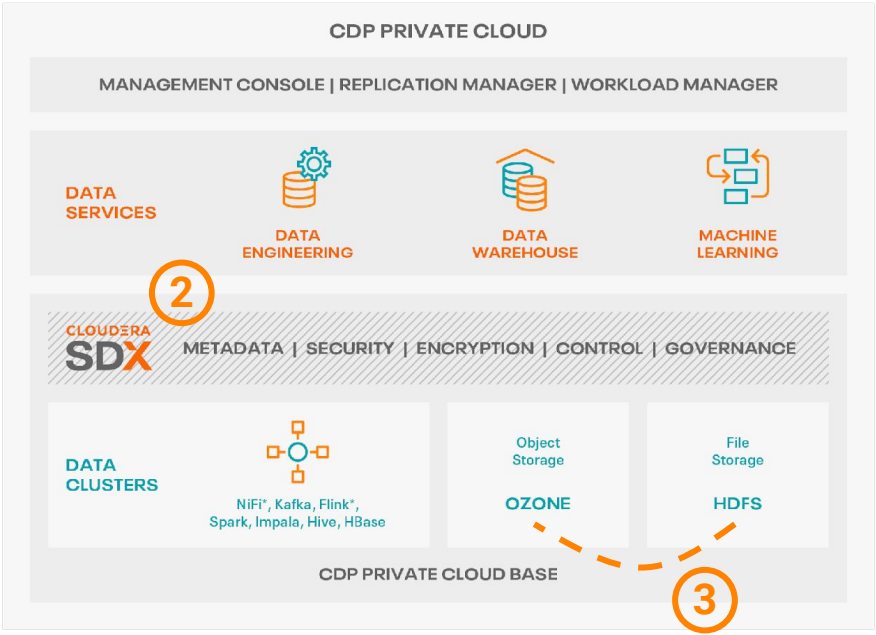
- Workflows optimized for self-service analytical experiences



IMPACT OF IN-PLACE MODERNIZATION

Separation of compute and storage





1. Upgrade cluster on the same hardware, no need to move data
   – reduced risk, new version value
2. Optimize governance model using new SDX controls compliance
   – improved security and governance, improved risk and
3. Optimize storage by shifting data to Ozone
   – higher storage density and scalability, reduced hardware costs

CDP PVC STRATEGIC FEATURES – DEMONSTRABLE

Faster, Easier Administration	Best in Class Data Lifecycle Solutions	Integrated Data Platform	Scalability & Performance	Secure & Governed
Simplify management with a centralized and holistic view of the entire data lifecycle.	Powerful analytics, transactional and ML workloads to accelerate data-driven decisions.	Eliminates data center silos storing and analyzing data, getting to deeper insights faster.	Optimize performance and efficiently match resources to demand for multiple concurrent analytic workloads.	Ensure control and governance with built-in security for the entire data and analytics lifecycle.
Single Pane Management Manage the full data lifecycle with a centralized single pane of glass for a holistic view.	**Purpose-built** Powerful, performant and functional analytics each excelling in their class.	**Single Copy of Data** Run unlimited concurrent workloads against the same, single copy of data with a consistent experience.	**Elastic Scalability** Automatically scale compute and storage resources. Shift excess capacity to workloads that need it.	**Integrated Policy Engine** Depoy consistent security and governance policies across all data, analytics and deployments.
Simplified Upgrades Upgrade tenants independently, for a seamless blend of innovation and stability. In-place upgrades from CDH and HDP.	**Dedicated Experiences** Purposeful, analytics focussed interfaces and capabilities for users.	**Any Data Type** Supports structured or unstructured data, delivered in batch or real-time.	**Resource Optimization** Strict tenant isolation and auto-scaling ensures critical workloads meet SLAs and optimal utilization at all times.	**Governed Access** Manage user access, monitor activity and audit usage of shared data across every workspace.
Advanced Automation Automate deployment and maintenance tasks for faster time to insight and predictable performance.	**Self-Service Access** Find the right data, start the right analytics for faster time to value and more efficient operation.	**Integrated Analytics** Interconnected analytics enable complex data pipelines to deliver value and deeper insight faster.	**Cost Efficiency** Consolidate clusters, minimize data replication for reduced operational overheads, risk and data center costs.	**Regulatory Compliance** End to end governance of sensitive data to meet evolving needs for compliance and the most regulated industries.

PvC Experiences

CLOUDERA

CLOUDERA DATA ENGINEERING (CDE)

Faster time to value with an integrated, purpose-built Data Service for data engineers

OPTIMIZED FOR MODERN DATA ENGINEERING

- Advanced orchestration with Apache Airflow, optimized for Spark jobs
- API-driven Devops automation tooling for Spark and Airflow jobs
- Containerized with elastic compute

CENTRALIZED MONITORING & PIPELINE MANAGEMENT

- Single pane of glass to manage logs, scheduling, configuration, & Spark
- Visual performance profiling & troubleshooting for Spark jobs

ENTERPRISE READY HYBRID ARCHITECTURE

- Faster tenant onboarding with optimized hardware utilization
- Managed multi-version Spark support
- Governed and secure with SDX

CLOUDERA DATA WAREHOUSE (CDW)

Empowers analysts to quickly answer burning business questions with ease

SELF-SERVICE ANALYTICS

- Easy to use and autoscaling
- Automation of DBA workflows
- Better SLA management
- Easy troubleshooting with 2-min workflows

FASTER INSIGHTS FOR FASTER BUSINESS DECISIONS

- Faster queries with Caching, automated MVs
- Faster insights from built-in integrations with CDE and CML
- Unlimited concurrency

SUPPORT FOR NEW BUSINESS USE CASES

- Supports Transactional DW, real time analytics
- New data sources
- Self-service discovery
- End to end lifecycle

CLOUDERA MACHINE LEARNING (CML)

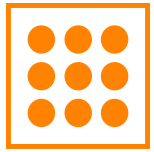
Enabling Production ML At Scale

EXPERIMENT FASTER, POWER MORE USE CASES

- Use any language or IDE including R, Python, or Scala natively
- Collaborate easily across teams
- Access on-demand compute in fully containerized ML workspaces

DEPLOY & MONITOR WITH CONFIDENCE

- Deploy models and apps to production with a few clicks
- Prediction-level monitoring & ground truthing



BRING YOUR DATA SCIENTISTS TO THE DATA

- Governed and granular access to enterprise data with SDX
- Consistent experience across hybrid cloud deployments
- No data silos

CDP PRIVATE CLOUD 1.3 - TODAY

Platform

- OpenShift 4.6 or Cloudera ECS
- Resource groups
- In-place updates
- Custom
  - external alerts
  - TLS certificates
  - diagnostic bundles

CDW

- Reduced minimum HW requirements
- Custom principal for Hive/Impala
- Impala ORC ACID reads
- Hive ACID compaction

CDE

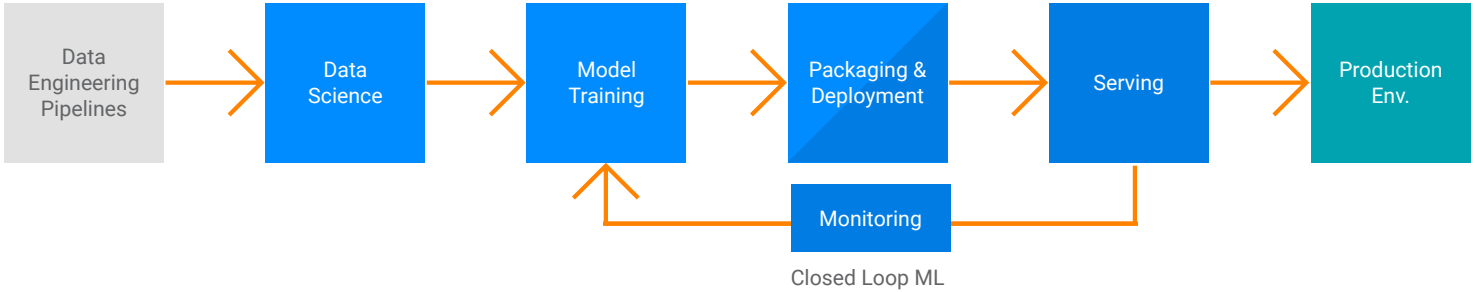
- Apache airflow on Kubernetes
- cli/api based provisioning

CML

- Applied ML Prototypes
- Improved air gapped deployment
- Improved NFS support
- Grafana monitoring

GET TO PRODUCTION, SCALE ML & AI USE CASES

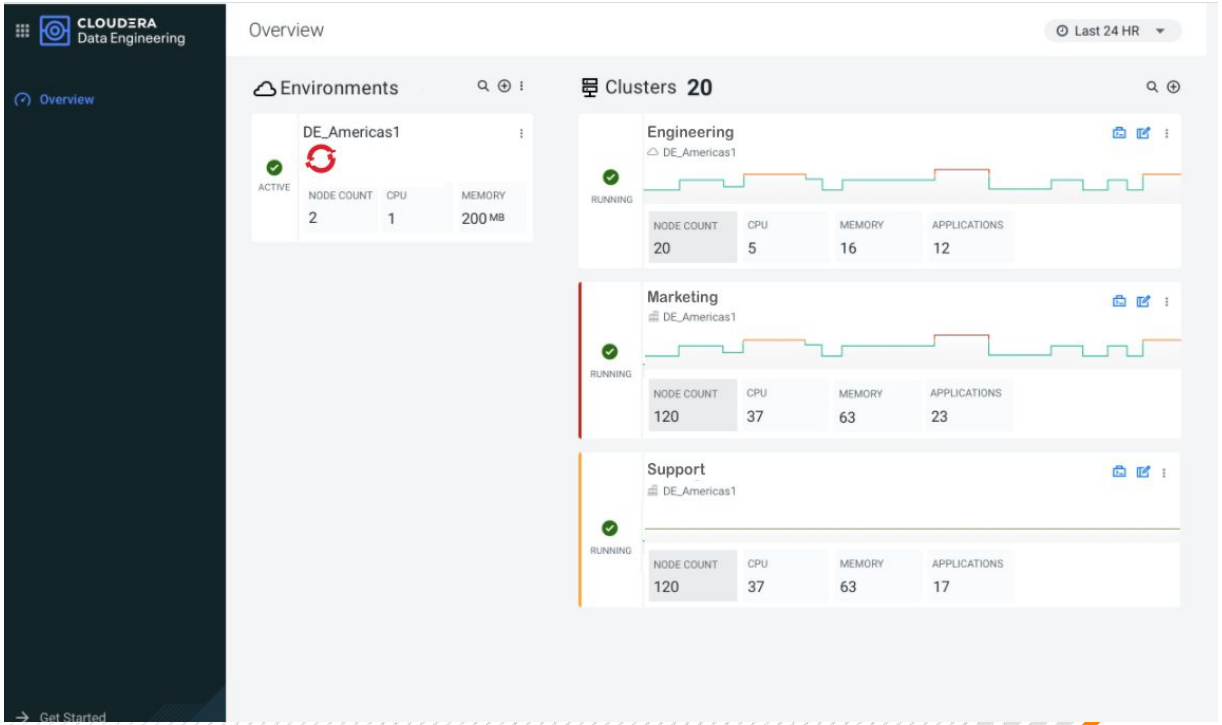
An integrated lifecycle is easier to use, manage and secure



CLOUDERA SDX

METADATA / SCHEMA / MIGRATION / SECURITY / GOVERNANCE

PVC 1.3 INTRODUCES *CLOUDERA DATA ENGINEERING*



CLOUDERA DATA ENGINEERING

Faster time to value with an integrated, purpose-built Experience for data engineers

CONTAINERIZED, MANAGED SPARK SERVICE

- Autoscaling compute
- Governed & secure with Cloudera SDX
- Mix version deployments

APACHE AIRFLOW SCHEDULING

- Open preferred tooling
- Orchestrate complex data pipelines
- Manage & schedule dependencies easily

TUNING & VISUAL TROUBLESHOOTING

- Resolve issues fast with real-time visual performance profiling
- Complete monitoring & alerting capabilities

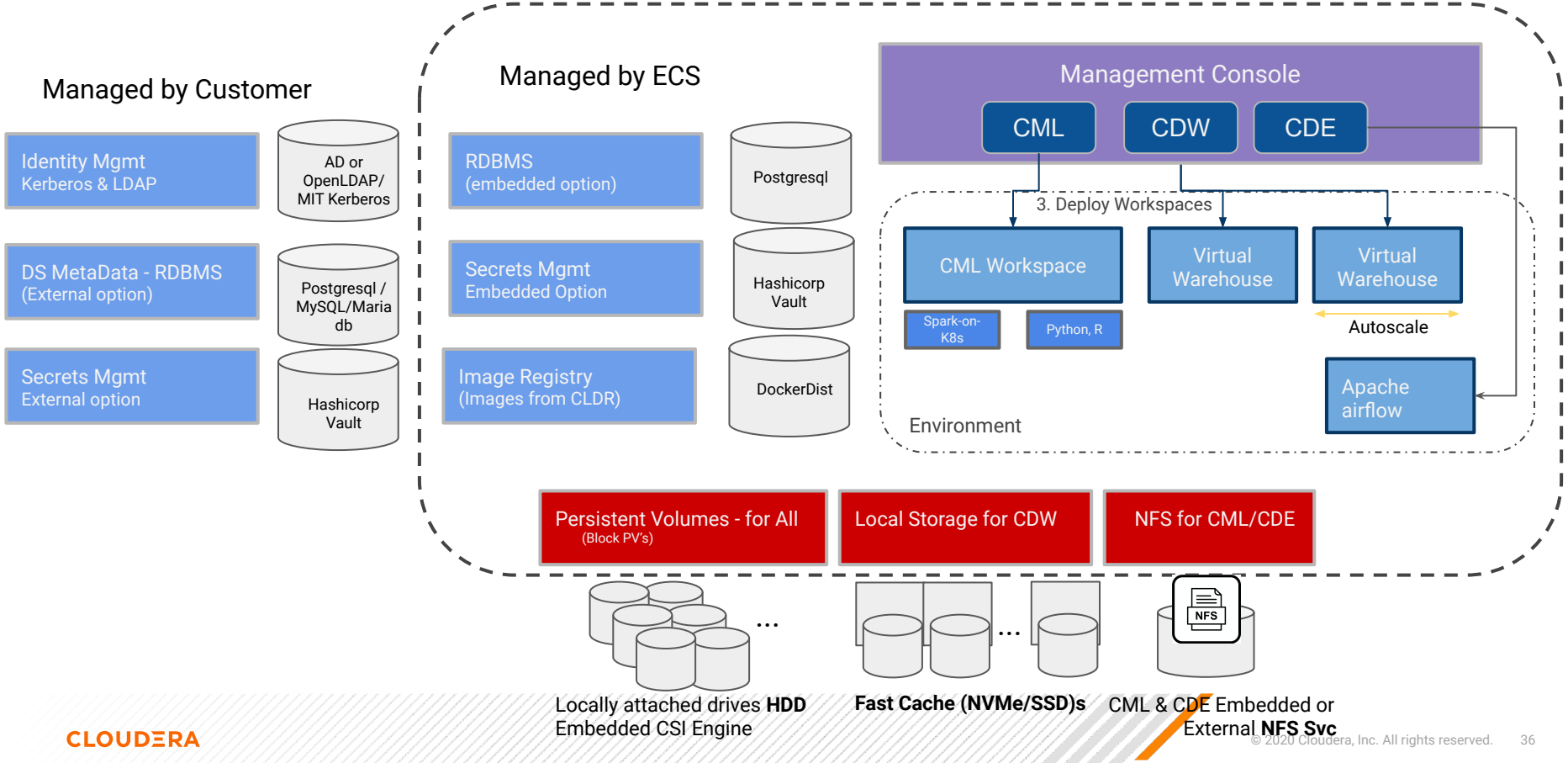
SIMPLIFIED JOB MANAGEMENT & APIS

- Full lifecycle mgmt.
- API-driven pipeline automation for any service
- Any language: SQL, Java, Scala, Python

CDP PvC POC requirements

CLOUDERA

PVC Architecture



	Bare metal or VMs (for Base Cluster)		Bare metal or VMs (for ECS Cluster)	
Component	**Minimum**	**Recommended**	**Minimum**	**Recommended**
Number of servers	6	8	10	16
CPU (cores)	16 cores	24 cores	16 cores	48 cores
RAM (GB)	32GB	256GB	64GB	384GB
Storage HDD (Num x Capacity)	12 x 2TB	12 x 4TB	1 x 2TB (SSD/SATA)	1 x 4TB (SSD/NVMe)
Network	Recommended: 25+GbE Spine/Leaf, max 4:1 oversubscription between the spine and leaf switches.			

Other dependencies

- PostgreSQL database server running version 10.6 or later.
- PostgreSQL TLS enabled
- Wildcard certs for subdomain signed by enterprise ca
- Wildcard subdomain
- Reverse DNS entries for all nodes
- LDAP Bind credentials
- Nexus docker image repository to host air gapped docker images
- External vault if available

CDP pvc Base Cluster Requirements

- System
  - RHEL/CentOS 7.x
  - CM 7.2.4
  - CDP 7.1.4 or 7.1.5.3 HF
  - CDP Private Cloud Plus license entitlements
  - JDK 11
- DB
  - External Postgres 10.6 DB configured for inbound TLS to the HMS DB
- Services - all must be healthy
  - Hive Metastore (HMS)
  - Ranger
  - Atlas
  - HDFS
  - Ozone
- Security
  - Kerberos must be configured w/ AD or MIT KDC
  - Ranger w/ LDAP groups must be configured
  - TLS must be enable on CM cluster
- Network
  - Must be on same network as OpenShift Cluster
  - Base cluster host names must be forward and reverse resolvable in DNS from OCP cluster
  - Load Balancer in front of OCP external API and must allow websocket traffic & HTTPS

See https://docs.cloudera.com/cdp-private-cloud-experiences/1.1/installation/topics/cdppvc-installation-cdp-data-center.html

Important Links

- CDP Test drive
- Product documentation
- CDP Private cloud base
- CDP Private cloud experiences

CLOUDERA

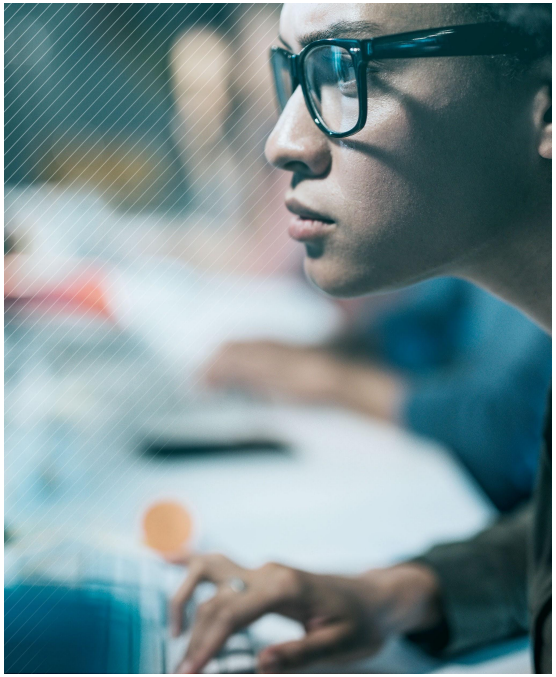
CUSTOMER SUCCESS

GLOBAL INFORMATION TECHNOLOGY COMPANY

$650K

Annual cost savings

Optimizing infrastructure spend and improving scalability, experimentation and customer service



CHALLENGE

- High infrastructure costs and inability to scale for expected growth in search traffic and explosion in data volumes
- Needed a semantic search engine to power the search function for all the apps on its platform

SOLUTION

- Migrated 900+ nodes to CDP Private Cloud Base
- Kafka on CDP advantages include replication, security, and monitoring
- Eliminated legacy data point technologies (i.e. Confluent)

OUTCOMES

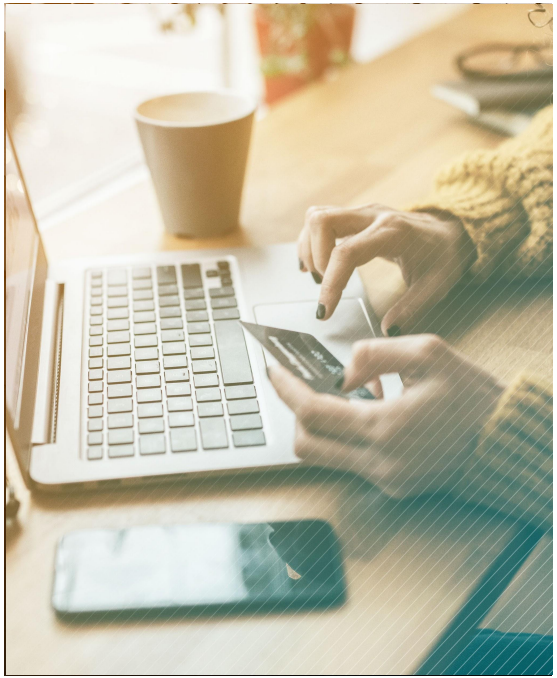
- Meeting SLA requirements with robust security and high availability
- Categorizing data more efficiently to improve relevancy
- Cross data center replication abilities globally

Read our blog to learn more about this customer's journey to CDP!

REGIONAL US BANK

In-Place Upgrade: CDH 5.14.2 to CDP PVC BASE 7.1.2

10	3	1PB
Weeks	Clusters	Data



CHALLENGE

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

SOLUTION

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

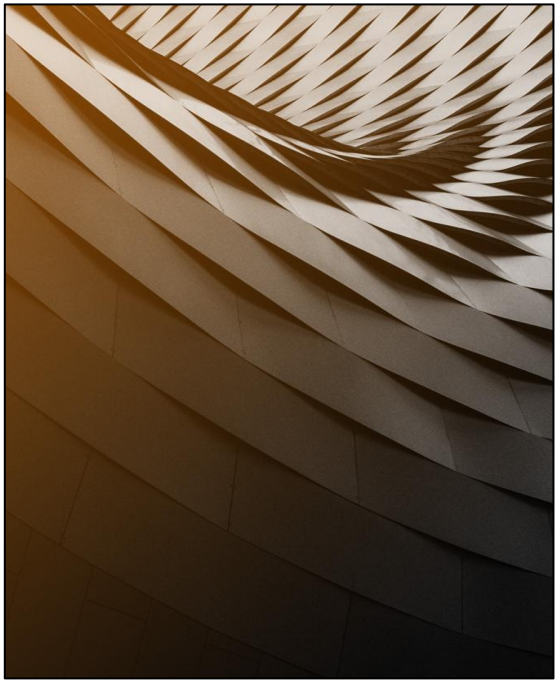
OUTCOMES

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

SOUTH AMERICAN BANK

In-Place Upgrade: CDH 5.15 to CDP PVC BASE 7.1.2

1000	3	1PB
Users	Clusters	Data



CHALLENGE

Customer was reaching end of support and needed to upgrade before December code freeze. Small window to do the PROD upgrade as they have 1000+ business users.

SOLUTION

SMARTUPGRADE for phased upgrade strategy, including doing a sidecar pilot with security and three key environments one at a time (DEV, UAT & PROD)

OUTCOMES

Successful upgrade to CDP PVC Base, transfer of knowledge to leverage Hive 3 ACID tables, Atlas, Ranger, Hbase

THANK YOU

CLOUDERA

CDH/HDP

CDP = Private Cloud + Public Cloud

CDP Private Cloud = Private Cloud Base + Private cloud experiences

Private Cloud Experiences = CML/CDW/CDE on Openshift or CML/CDW/CDE on ECS

ECS = Elastic Container Service by Cloudera