CLOUDERA

Cloudera Data Platform - CDP

Presales Deep Dive

Suresh MR

Director – Channels & Alliances

sureshmr@cloudera.com

Vinay Rayker

Partner Technology Lead, India

vrayker@cloudera.com

AGENDA

- Cloudera at a glance
- Enterprise Data Cloud
- Cloudera Data Platform
- CDP Demo
- More on CDP Technicals
- Packaging, Pricing & Value Proposition
- Q&A

CLOUDERA AT A GLANCE

CLOUDERA

THE ENTERPRISE DATA CLOUD COMPANY

Any Cloud

Data Lifecycle

CLOUDERA SDX

Secure & Governed

Open

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

January 2021

Cloud Data Ecosystems

Gartner

January 2020

Enterprise Intelligence Platforms

December 2019

Cloudera Data Platform (CDP)

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..."

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

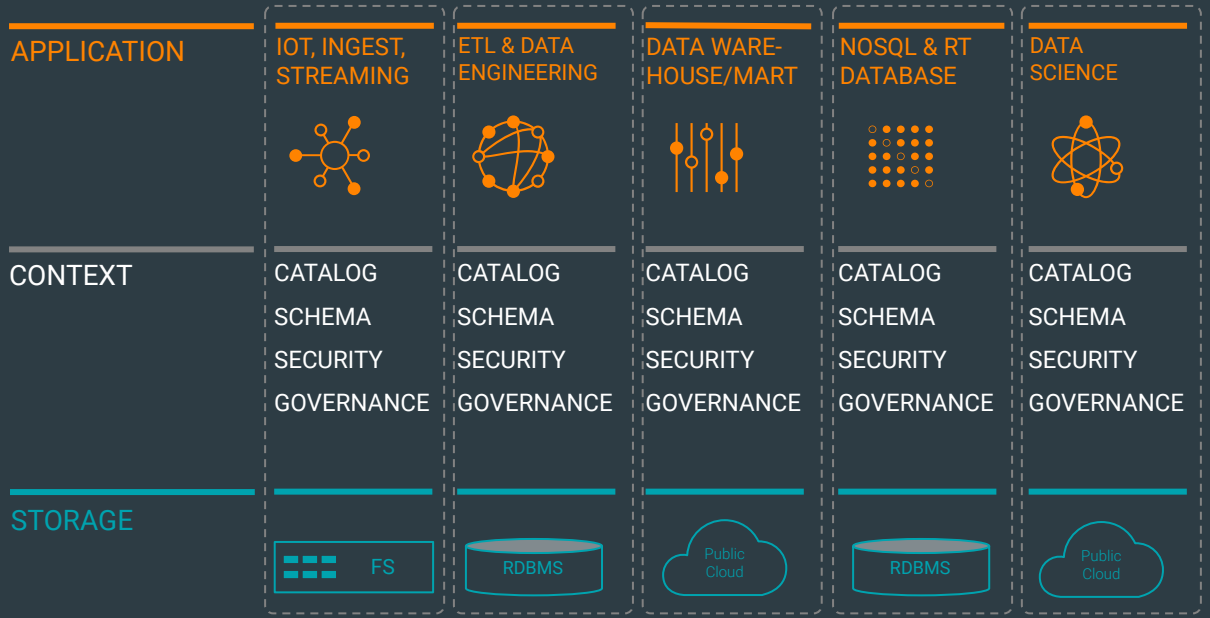
52%

Of repatriated workloads move to private clouds

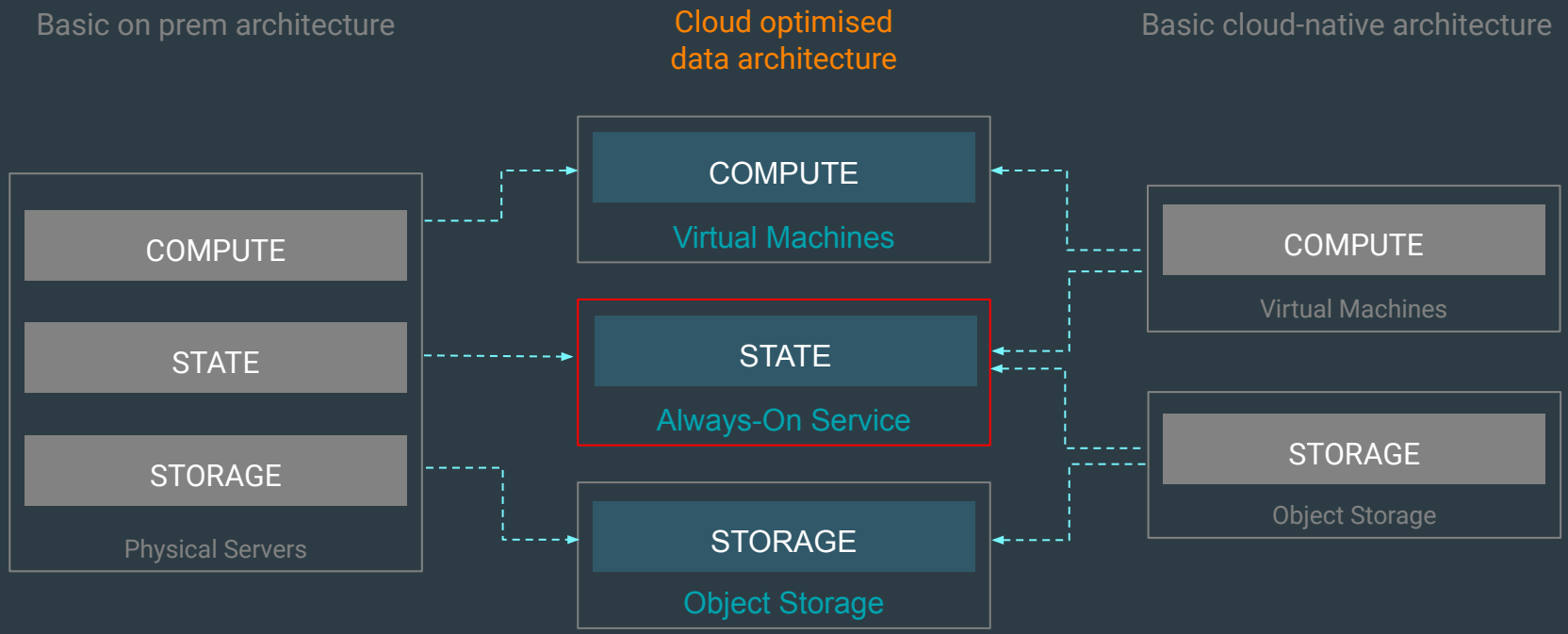
"As enterprise customers gain cloud expertise they're placing investments in private cloud solutions for increased security, compliance, performance, control and cost savings. Private clouds often act as a stepping-stone in the hybrid cloud journey."

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

DATA AND INSIGHT SILOS

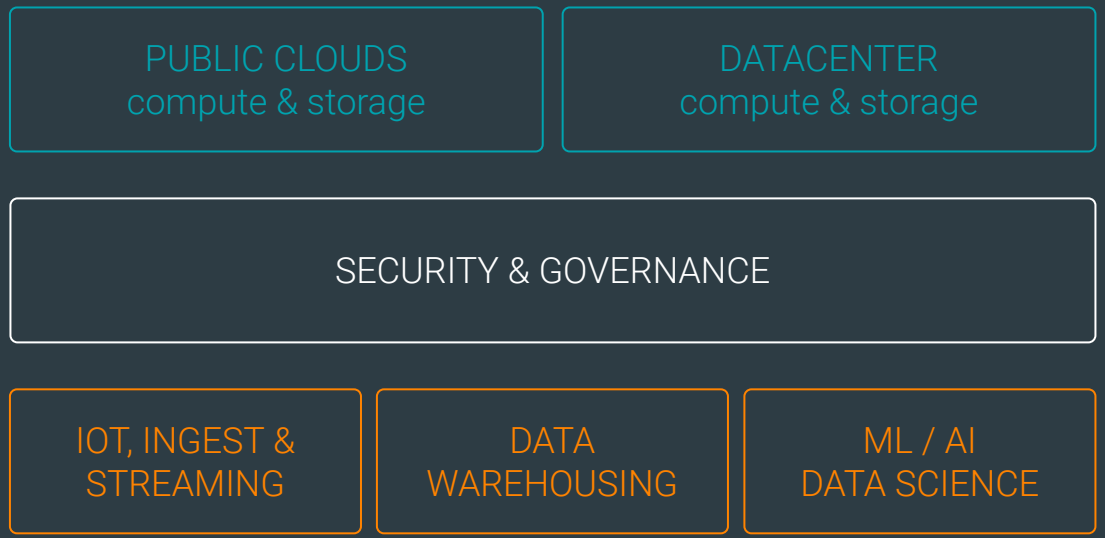


SEPARATE STORAGE AND COMPUTE *AND STATE*



ENTERPRISE DATA CLOUD DESIGN PRINCIPLES

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

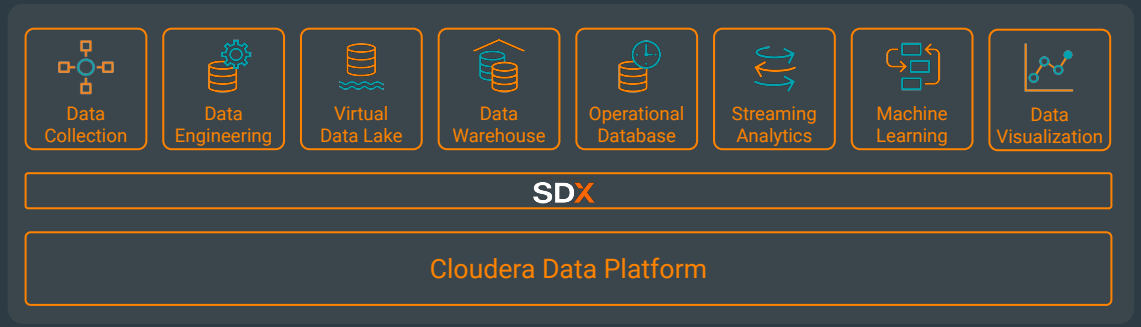


CLOUDERA DATA PLATFORM

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

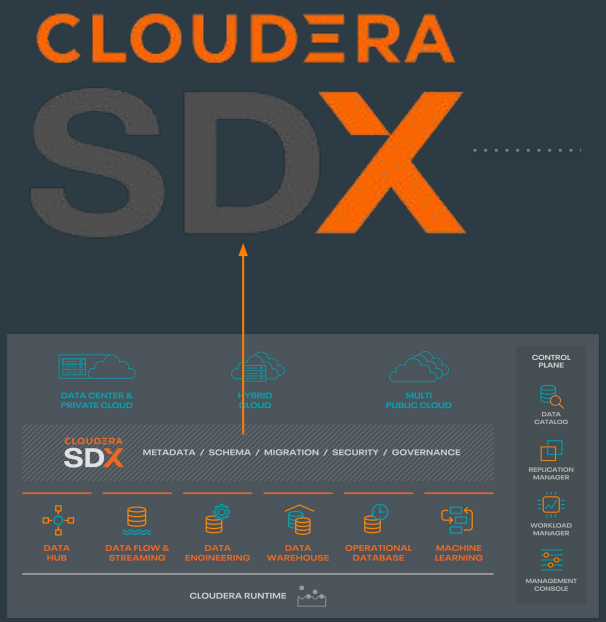
Hybrid & Multi-Cloud
to leverage existing investments and reduce risk

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

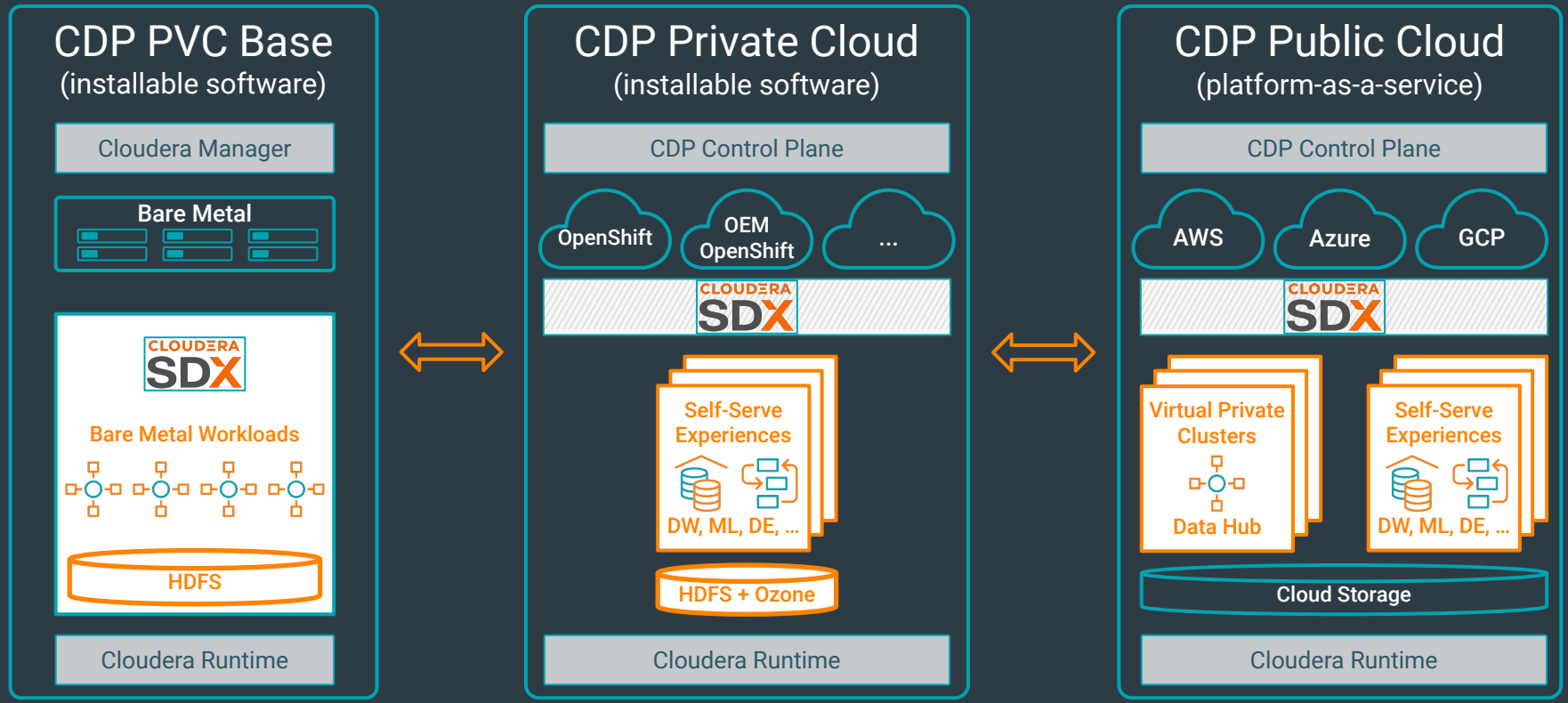
Schema: automatic capture and storage of any and all schema and metadata definitions as they are used and created by platform workloads

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

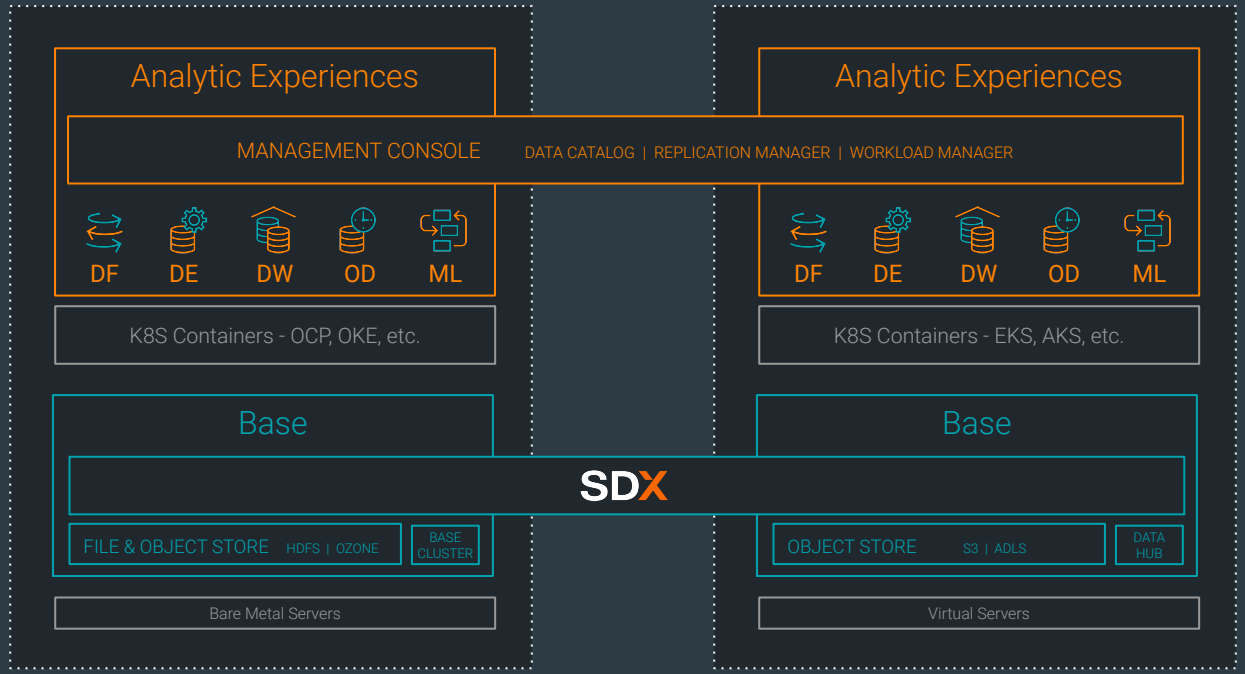
CDP - ONE PLATFORM, THREE FORM FACTORS



CDP HYBRID CLOUD

Consistent operations and analytics experiences across private and public clouds

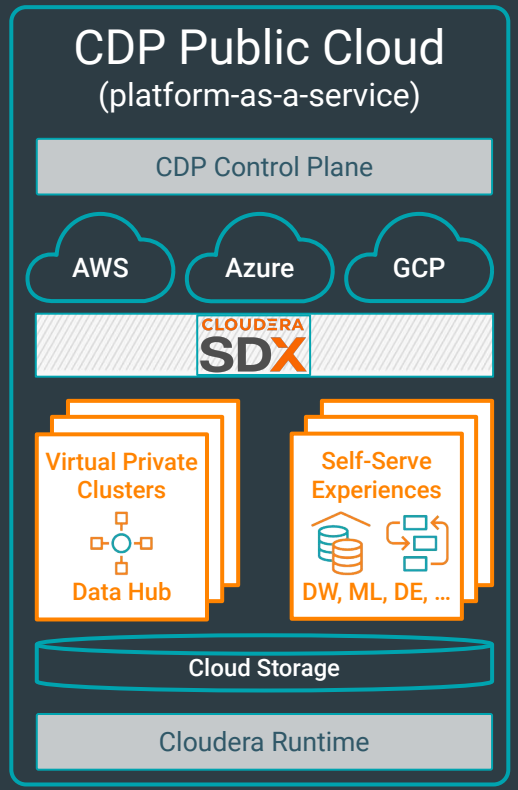
CDP Private Cloud



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

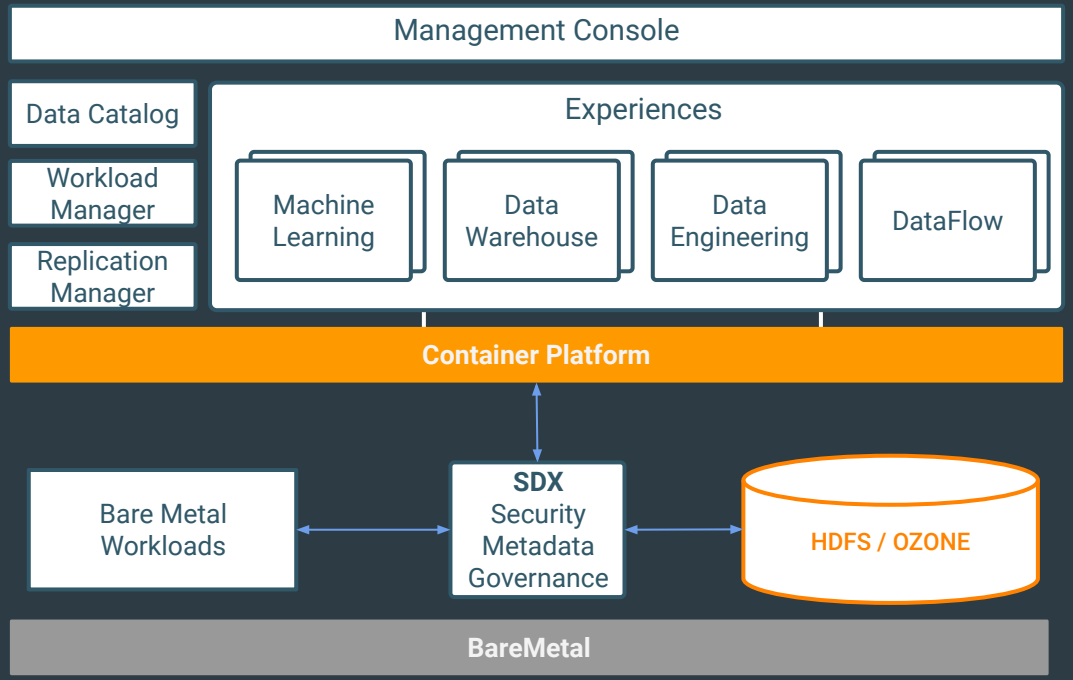
Burst to Cloud

- 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 EXPERIENCES

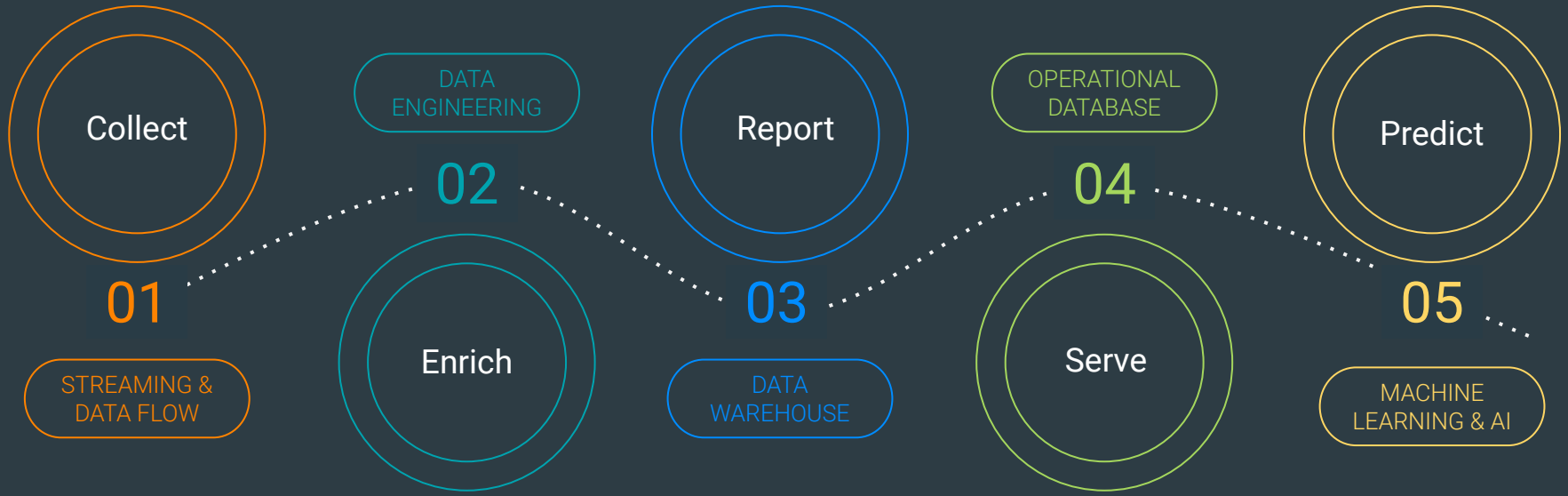
What are the consumption options?

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

CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter

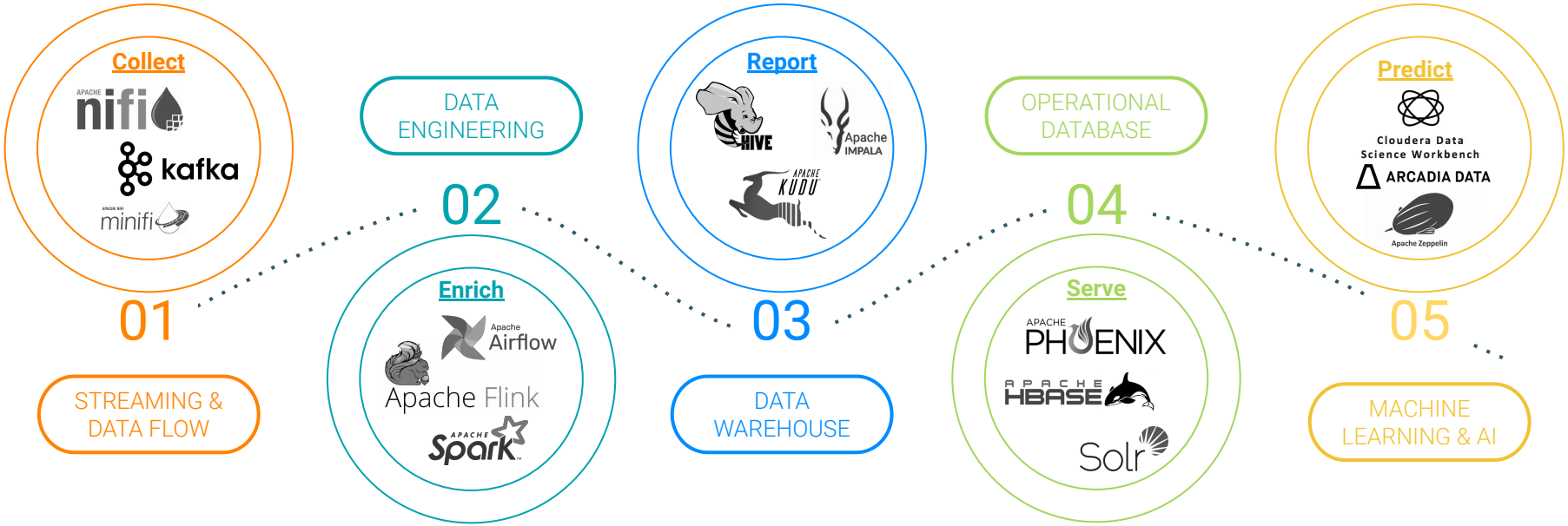


CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

… A RUNTIME FOR THE ENTERPRISE DATA LIFECYCLE

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



SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

Apache Ranger Apache Atlas

CONNECTING THE DATA LIFECYCLE

Starting the data lifecycle journey - solving the “first mile” problem

CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

CLOUDERA DATAFLOW DATA-IN-MOTION PLATFORM

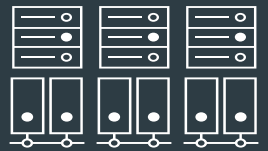
CLOUDERA DATAFLOW SERVICE “EXPERIENCE” (CDF)

A complete SDLC experience for data pipelines from dev through prod

Cloudera DataFlow

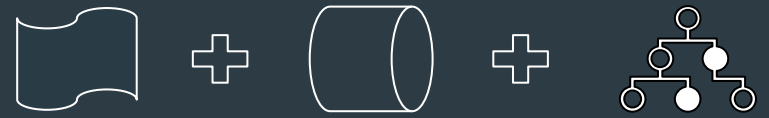
Servers/Clusters/Runtimes

(NiFi, Kafka, Flink)



Cloudera DataFlow Experience

Flows (H1)	Topics (H2)	Stream Apps (H2)
Schema Aware	Schema Aware	Schema Aware



Data Pipelines

Simplifies app development by enabling developers to focus on business logic

Deployment Dashboard

Dashboard
Topics
Catalog
Environments

Help
George Vetticaden

Filter Clear

METRICS WINDOW 30 min

UPDATED: 24 seconds ago

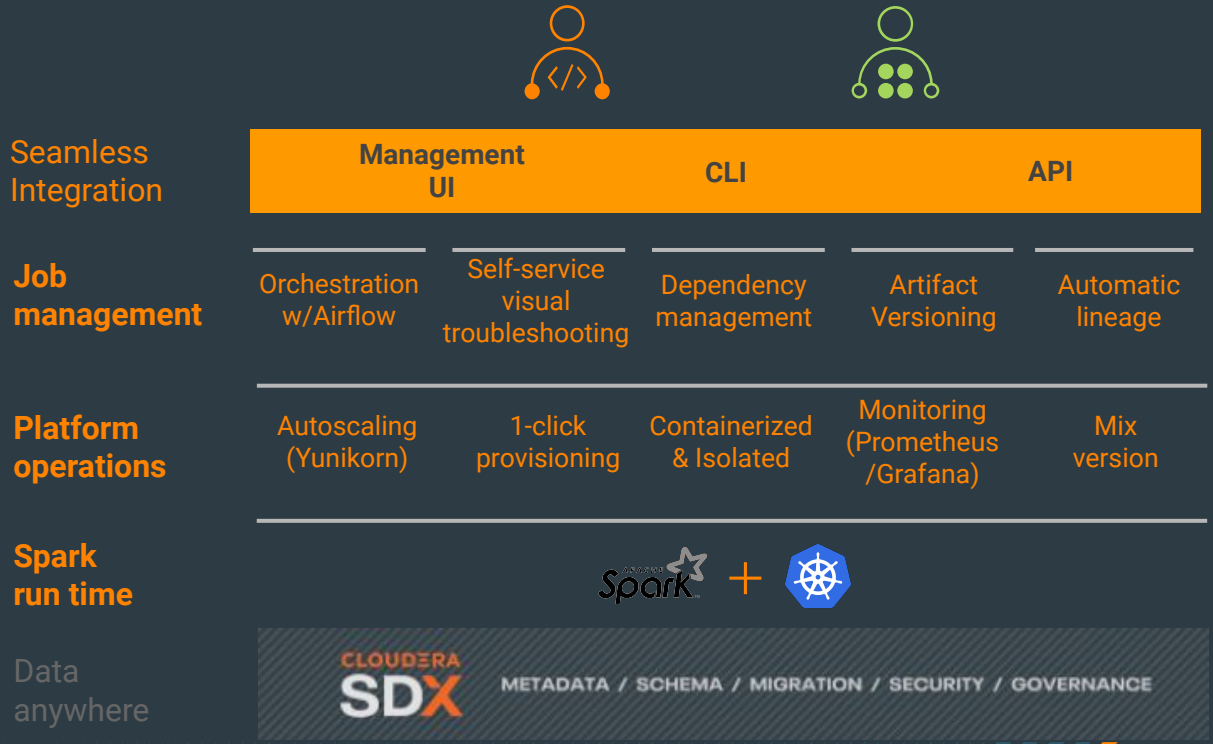
Status	Name	Entity Type	Data Received	Data Sent
Running	Kiosk Critical Event Detection SFO sit-awareness-kiosks-usWest-prod	Data Flow	5 MB/s Records: 83 Files: 97	10 MB/s Records: 168 Files: 126
Running	Kiosk Critical Event Detection ORD sit-awareness-kiosks-usEast-prod	Data Flow	20 MB/s Records: 208 Files: 267	32 MB/s Records: 456 Files: 789
Running	Kiosk Critical Event Detection LHR sit-awareness-kiosks-usEast-prod	Data Flow	20 MB/s Records: 208 Files: 267	32 MB/s Records: 456 Files: 789
Running	Kiosk Critical Event Detection SIN sit-awareness-kiosks-usWest-prod	Data Flow	20 MB/s Records: 208 Files: 267	32 MB/s Records: 456 Files: 789
Running	Kiosk Streaming Analytics sit-awareness-kiosks-usEast-prod	Streaming App	20 MB/s Records: 208 Files: 267	32 MB/s Records: 456 Files: 789
Running	Kiosk Wait Times sit-awareness-kiosks-usEast-prod	Streaming SQL App	20 MB/s Records: 208 Files: 267	32 MB/s Records: 456 Files: 789

CONNECTING THE DATA LIFECYCLE

Enriching the data lifecycle journey

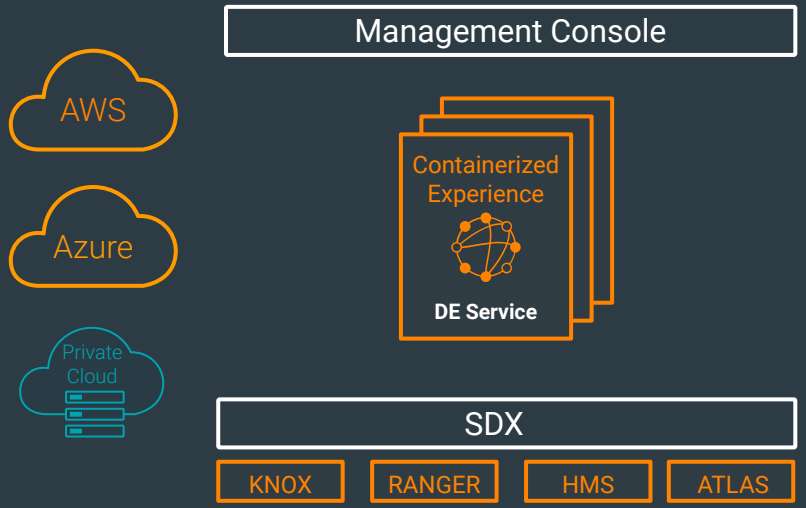
CLOUDERA SDX SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

CLOUDERA DATA ENGINEERING



TARGET PERSONAS

MANAGING RESOURCES & MANAGING JOBS



Platform Admins

- Quickly provision new workloads
- Ensure isolation across LoB
- Resource guardrails
- Control costs through on-demand autoscaling
- Show resource usage over time
- Centralized Access controls & governance

Data Engineer

- Easy deployment & monitoring of jobs
- Self-service troubleshooting with rich visual analysis
- Powerful workflow scheduling
- Automatic lineage capture
- Multiple versions of Spark

CONNECTING THE DATA LIFECYCLE

Enriching the data lifecycle journey

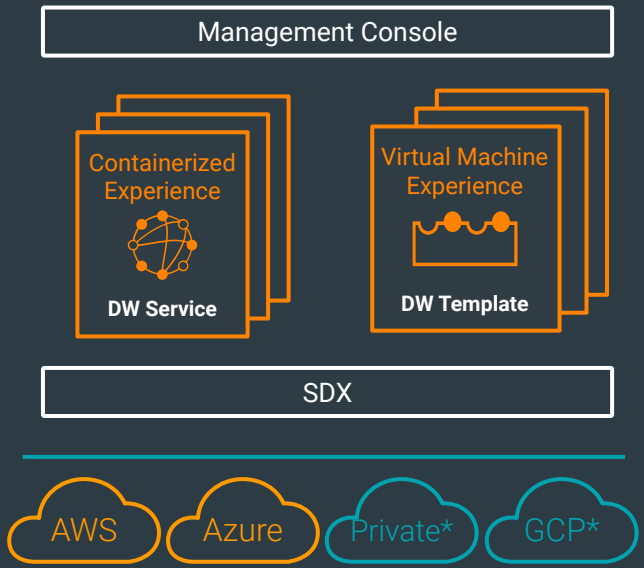
CLOUDERA SDX SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

CDW is a managed data warehouse service that runs Cloudera's **powerful engines** on a **containerized architecture** to let you **meet SLAs**, **onboard new use cases** with zero friction, and **minimize cost**

Two Cloud-Native Solutions for CDW

DW Service

- Kubernetes orchestration of container-based compute for agile clusters
- Opinionated and packaged provisioning / scaling
- Commonly administered by Line of Business
- Simplicity and ease of use over customization and control

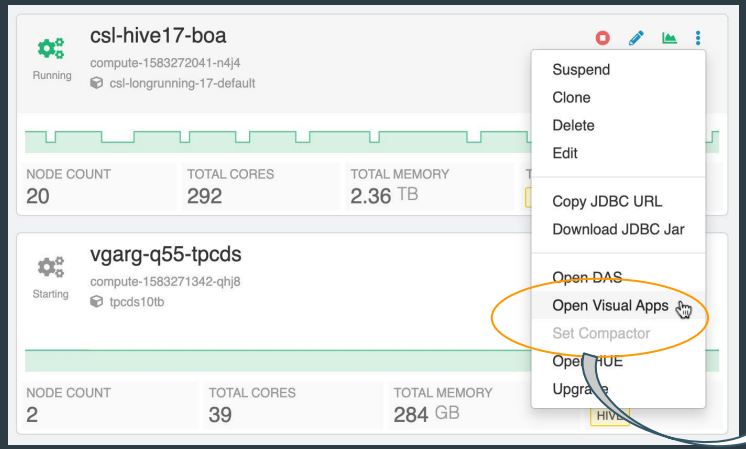


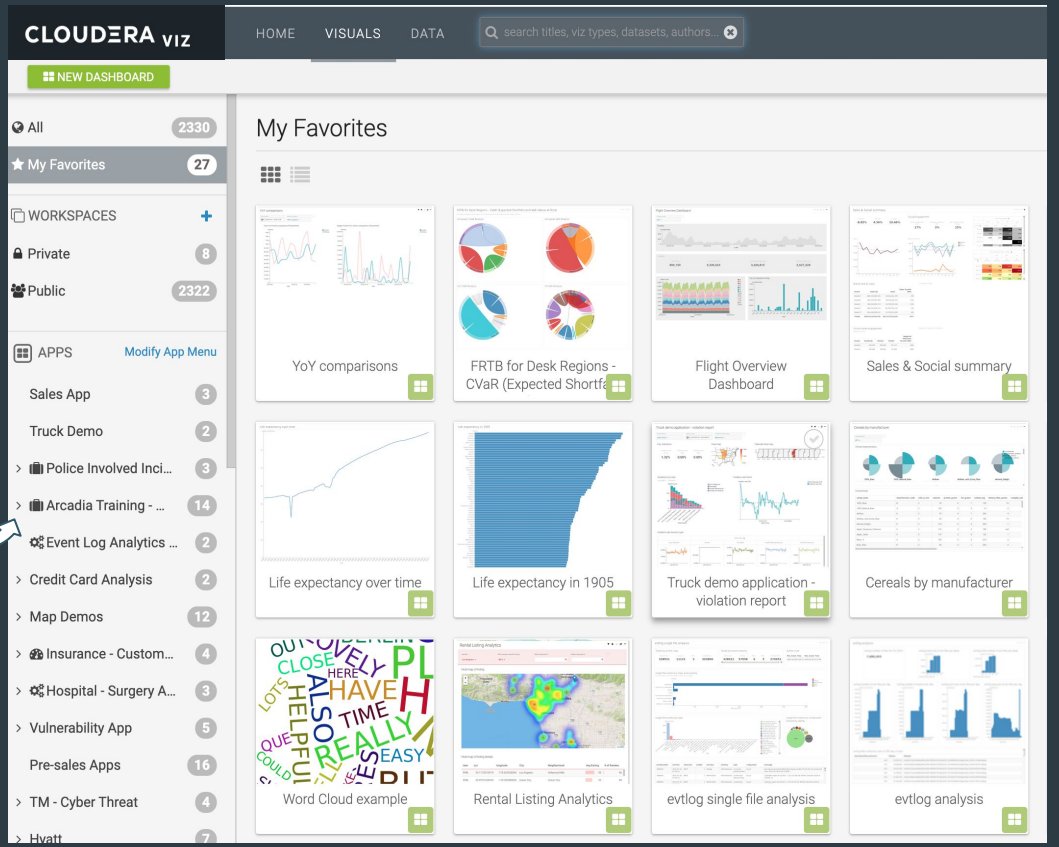
DW Template

- Native VM clusters for complex long running workloads (BI, ETL)
- Bespoke and flexible provisioning / scaling
- Typically administered by Central IT
- Customization and control over simplicity and ease of use

* Future Release

DW Viz in CDW



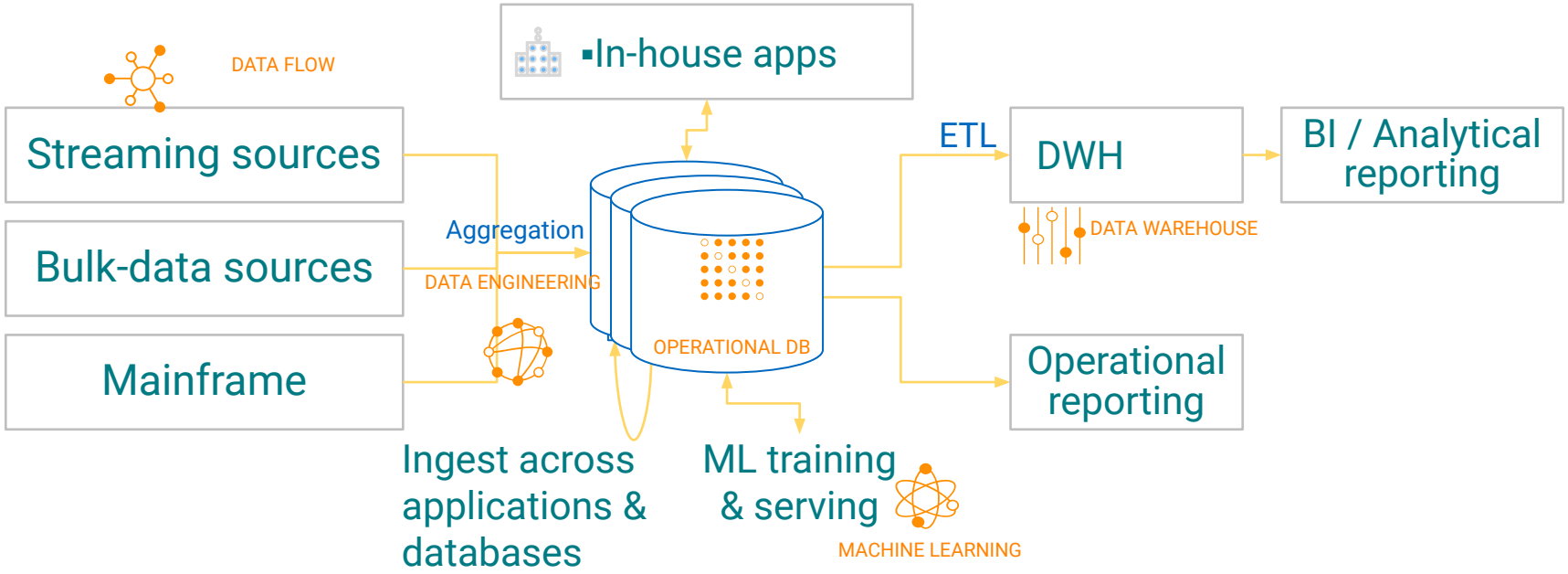


CONNECTING THE DATA LIFECYCLE

Completing the data lifecycle journey - solving the “last mile” problem

CLOUDERA SDX SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

HOW DOES OPDB FIT IN YOUR ENVIRONMENT



WHAT TO EXPECT IN CDP PUBLIC CLOUD

Allow developers to spend time where it matters

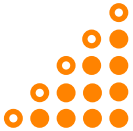
Easy and quick deployment for developers

3 Clicks

20 Minutes

Reduces deployment time to minutes from weeks/months on legacy databases

Autonomous management for admins



Auto Scale

Optimizes cloud utilization

Auto Tune

Improves performance

Auto Heal

Resolves operational failures

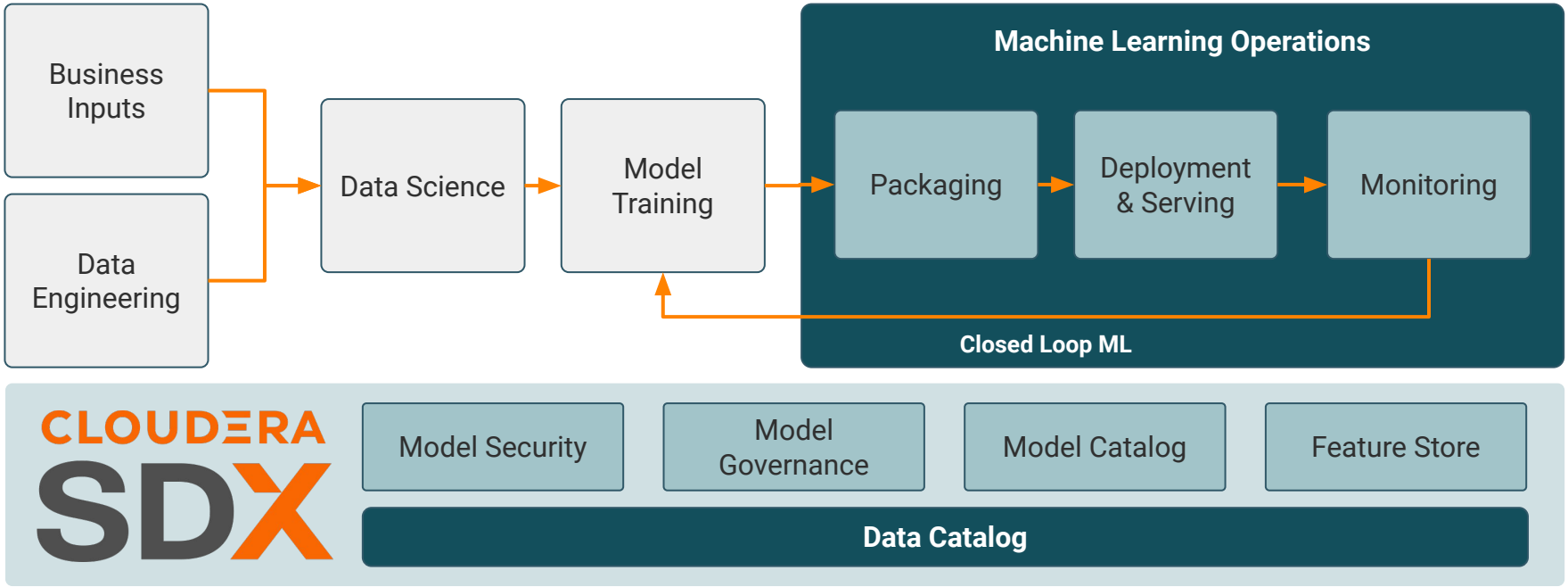
Eliminates operational management

CONNECTING THE DATA LIFECYCLE

Completing the data lifecycle journey - solving the “last mile” problem

CLOUDERA SDX SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

MACHINE LEARNING IN PRODUCTION

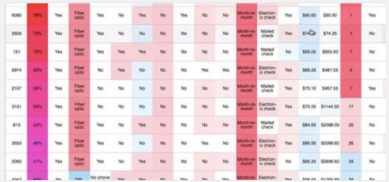


Applied ML Prototypes Catalog

Projects
Sessions
Experiments
Models
Jobs
Applications
Settings
Prototype Catalog
Admin

Search prototypes

Tags: all

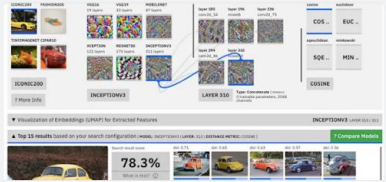


Churn Modeling with XGBoost

EXPLAINABILITY, XGBOOST

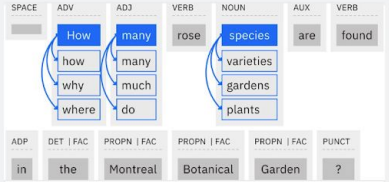
Airline Delay Prediction

XGBOOST



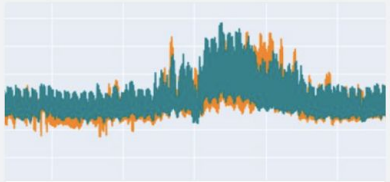
Deep Learning for Image Analysis

COMPUTER VISION IMAGES



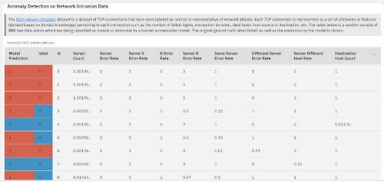
Neural Question Answering

NLP



Structural Time Series

TIME SERIES

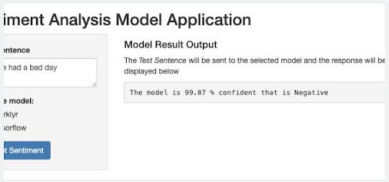


Deep Learning for Anomaly Detection

ANOMALY DETECTION DEEP NEURAL NETS

Fraud Detection

FRAUD DETECTION ANOMALY DETECTION



Sentiment Analysis with Sparklyr and Tensorflow

SENTIMENT ANALYSIS SPARK

< 1 2 >

License expires in 59 days

CONNECTING THE DATA LIFECYCLE

Completing the data lifecycle journey - solving the “last mile” problem

CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

UNIFIED MANAGEABILITY, SECURITY AND DATA GOVERNANCE

Identity & Perimeter

Validate users in enterprise directory

Technical Concepts:
Authentication
User/group mapping

Kerberos, Apache Knox

Access

Defining what users and applications can do with data

Technical Concepts:
Permissions
Authorization

Apache Ranger

Visibility

Reporting on where data came from and how it's being used

Technical Concepts:
Auditing
Lineage

Apache Atlas

Data Protection

Shielding data in the cluster from unauthorized visibility

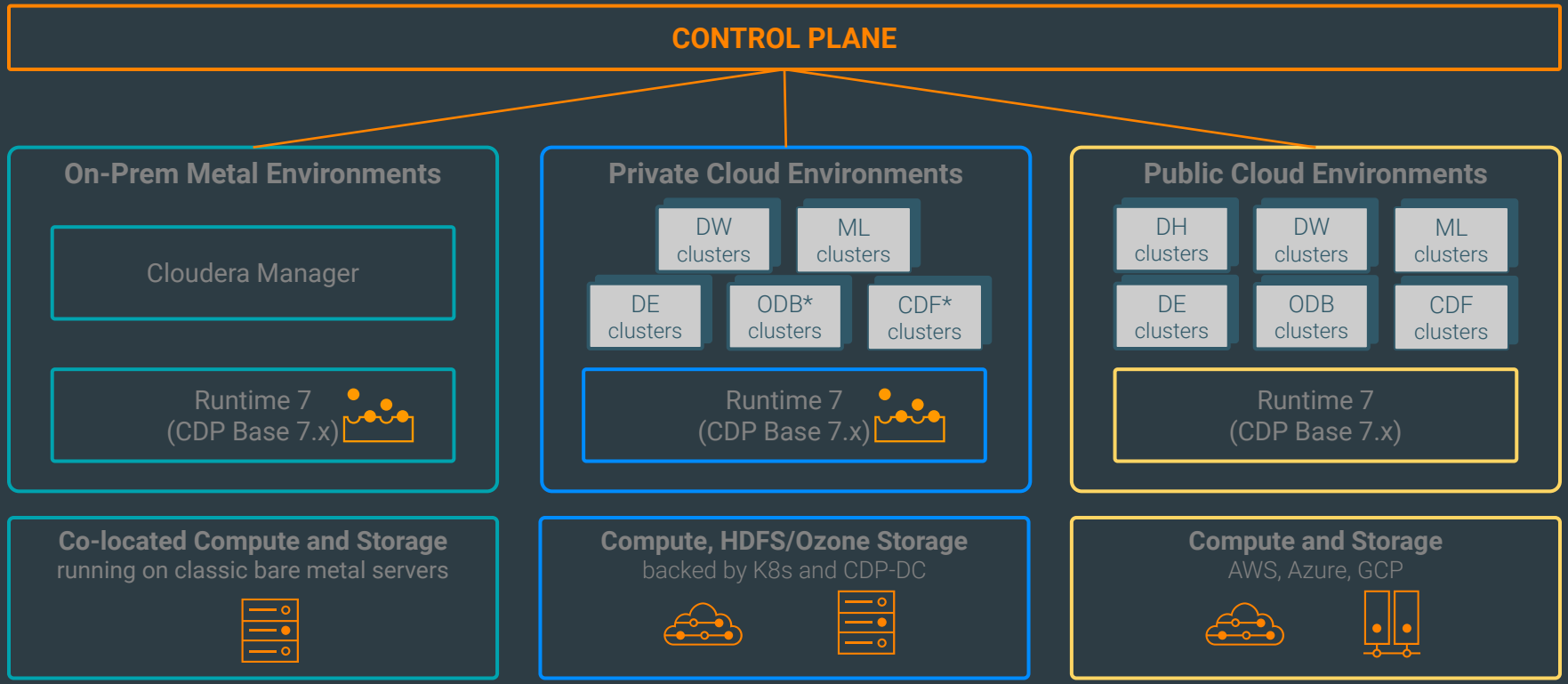
Technical Concepts:
Encryption, Key Management

SSL/TLS, HDFS TDE, Ranger (KMS, Masking, Filtering)

Cloudera Data Platform Demo

More on CDP Technicals

3 FORM FACTORS

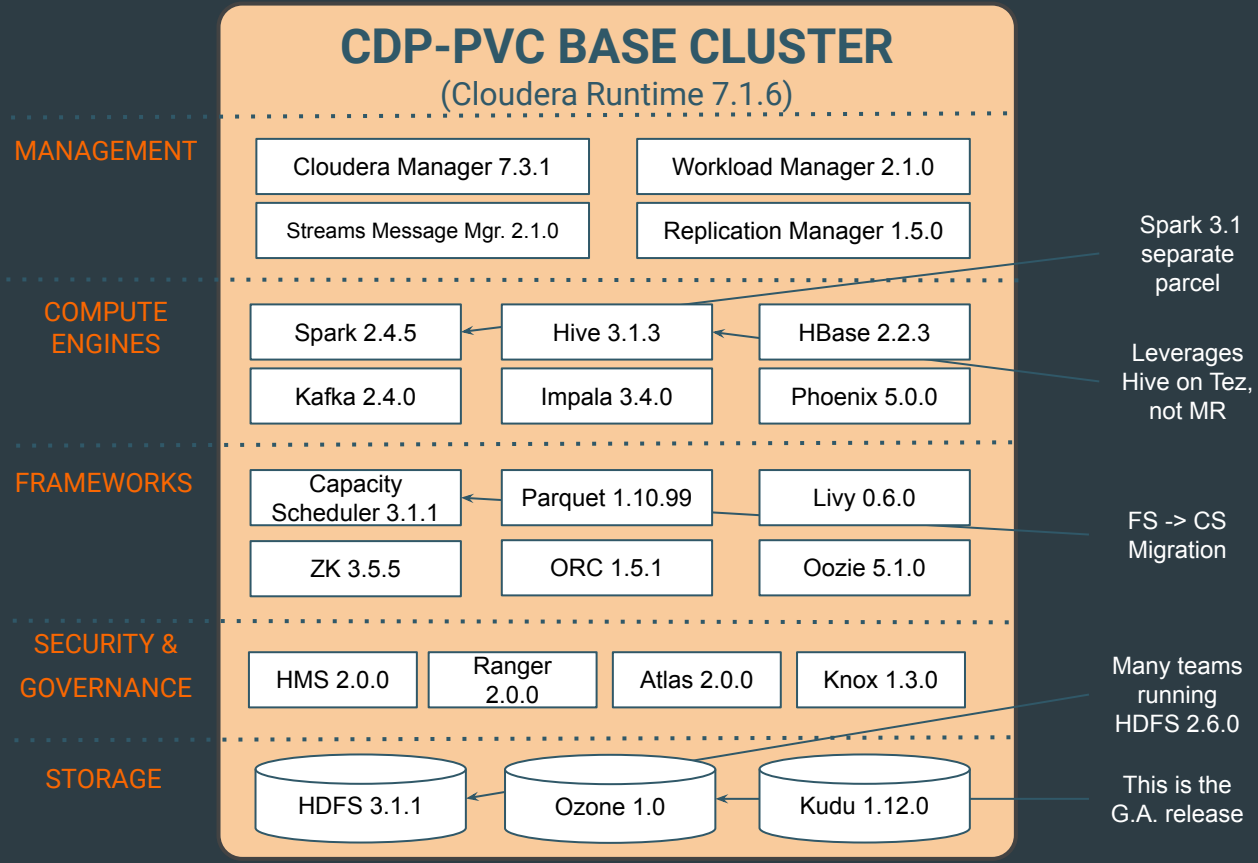


CDP RUNTIME 7.1

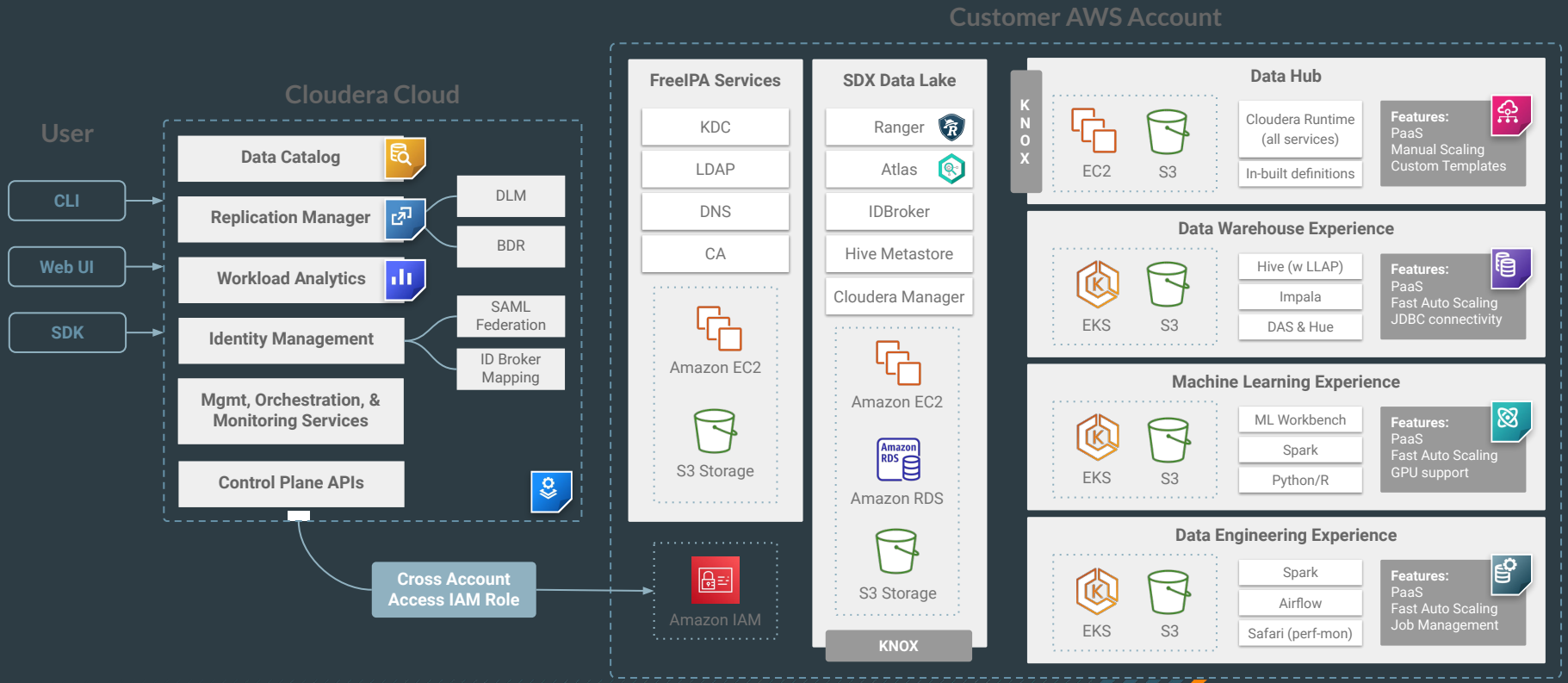
A CDP Private Cloud Base cluster (powered by Cloudera Runtime), can serve as a traditional "data lake" (storage & compute) cluster, or as a "base storage cluster" (storage only) serving compute workloads running on Kubernetes.

This image shows the component versions in Cloudera Runtime 7.1.6.

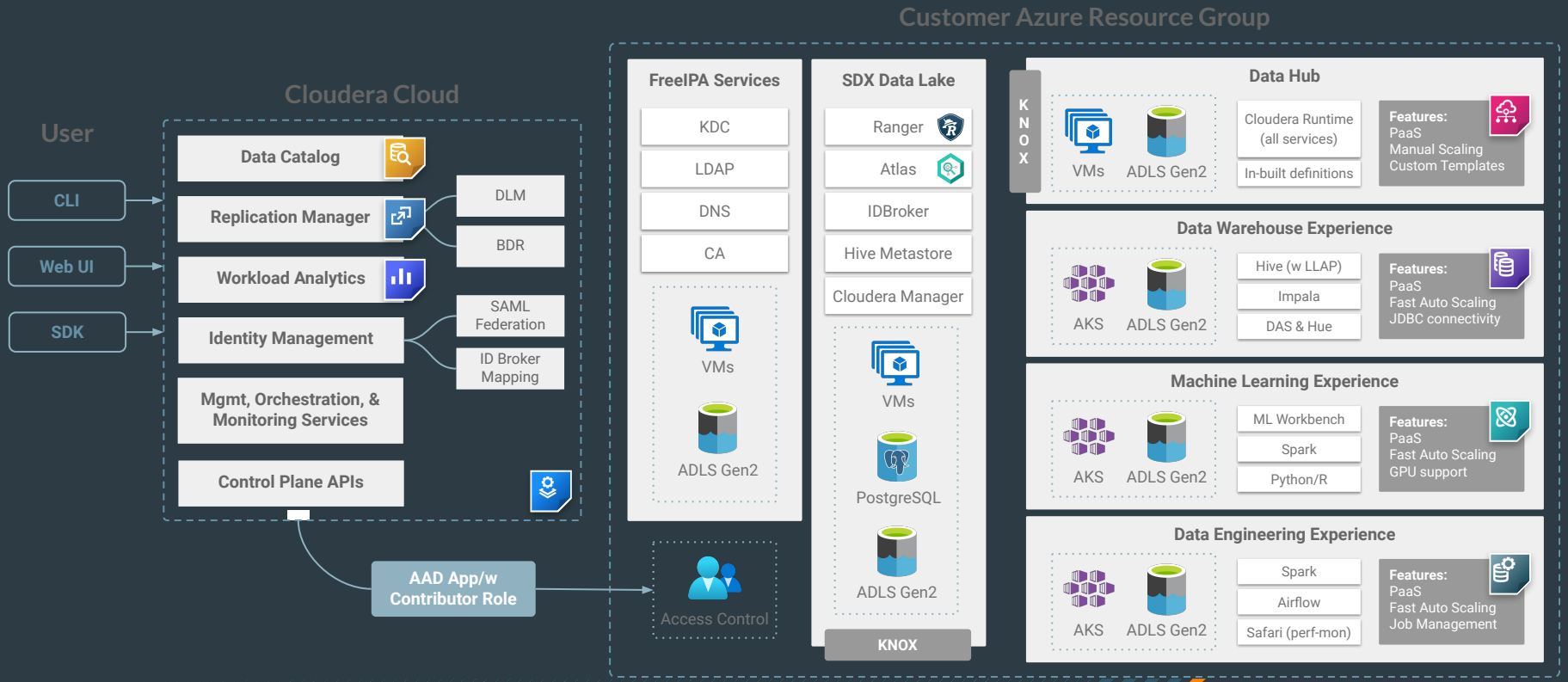
Click HERE for the complete list of supported components.



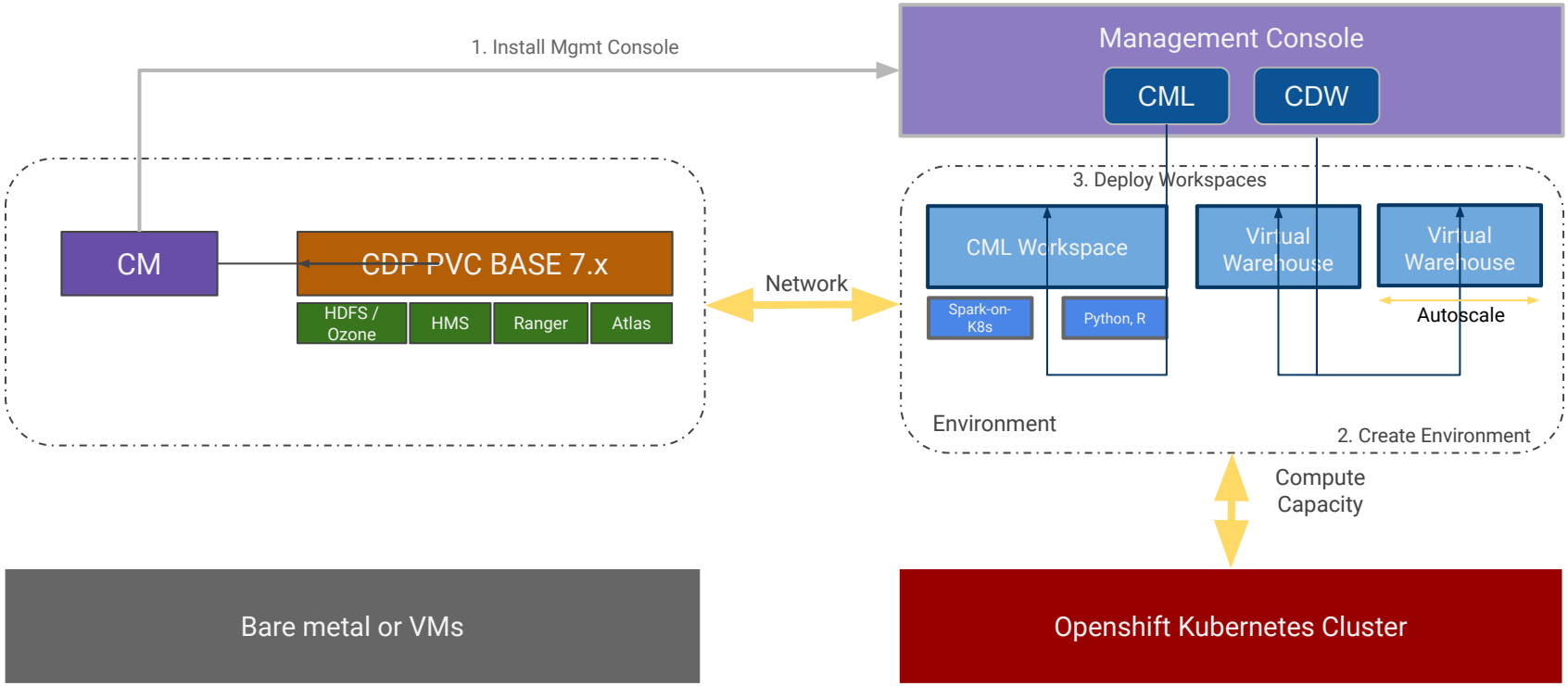
CDP - AWS HIGH LEVEL ARCHITECTURE



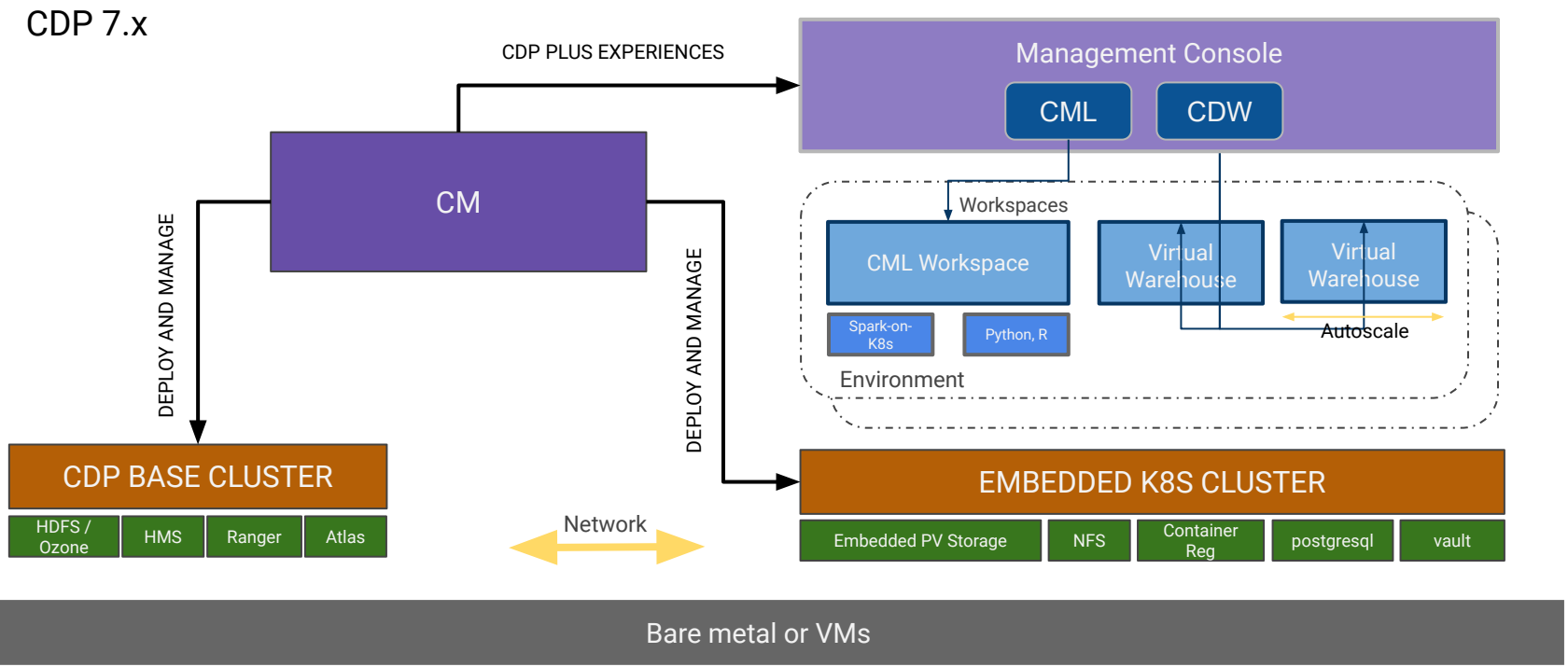
CDP - AZURE HIGH LEVEL ARCHITECTURE



CDP Private Cloud – Dedicated Kubernetes

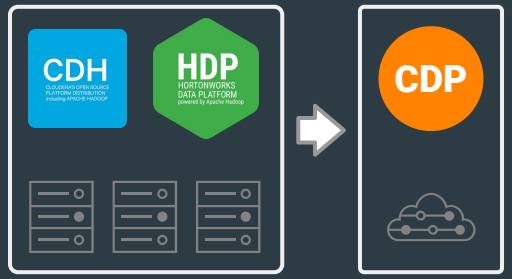


CDP Private Cloud – Embedded Version



THREE PATHS TO CDP

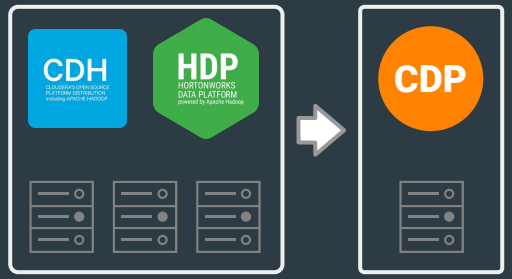
Migrate to Public Cloud



Copy data and metadata to a public cloud; implement new, or migrate existing workloads on CDP Public Cloud.

Small initial investment

Migrate to CDP PVC-Base



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

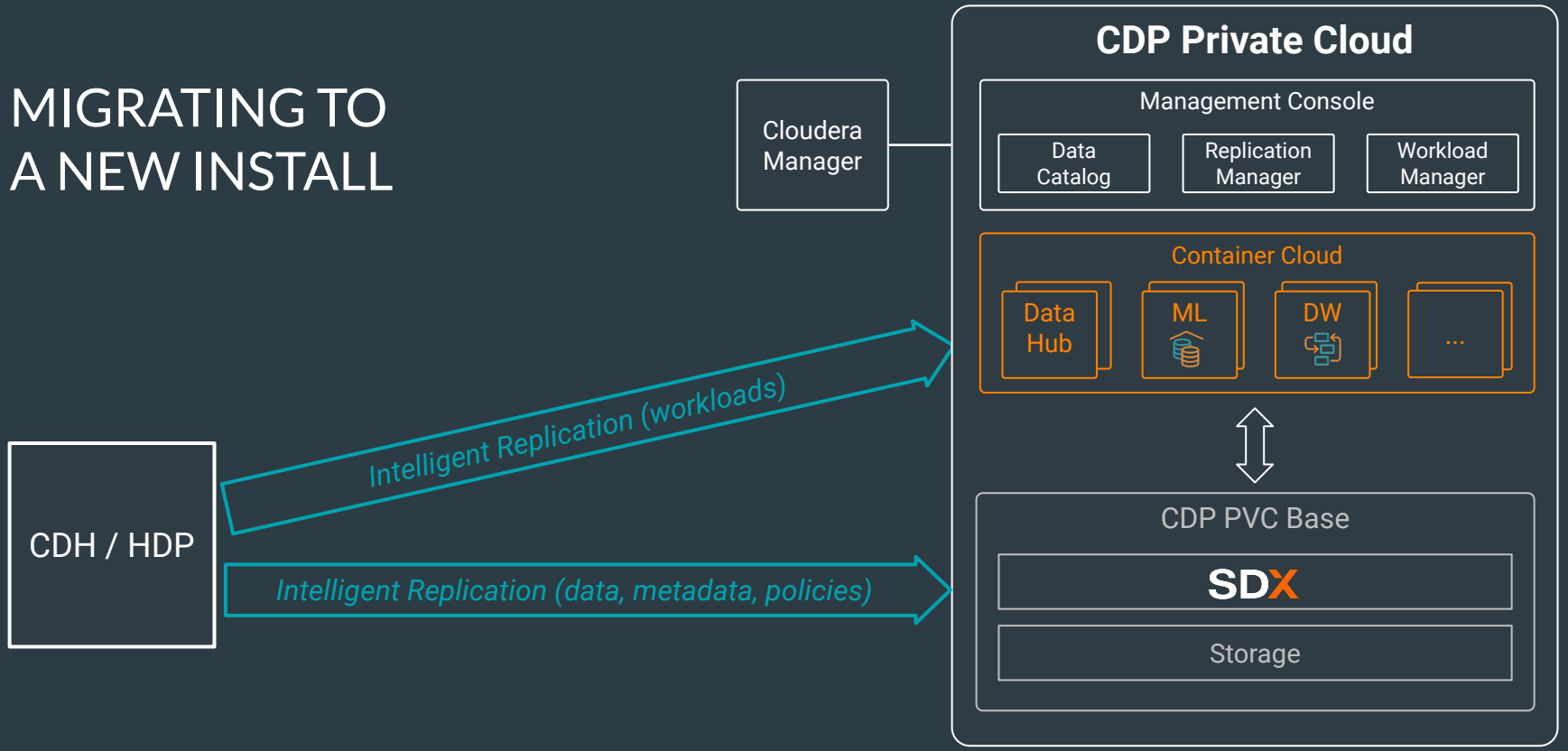
Higher initial investment

Upgrade to CDP PVC-Base

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

Single cutover, lower capital investment

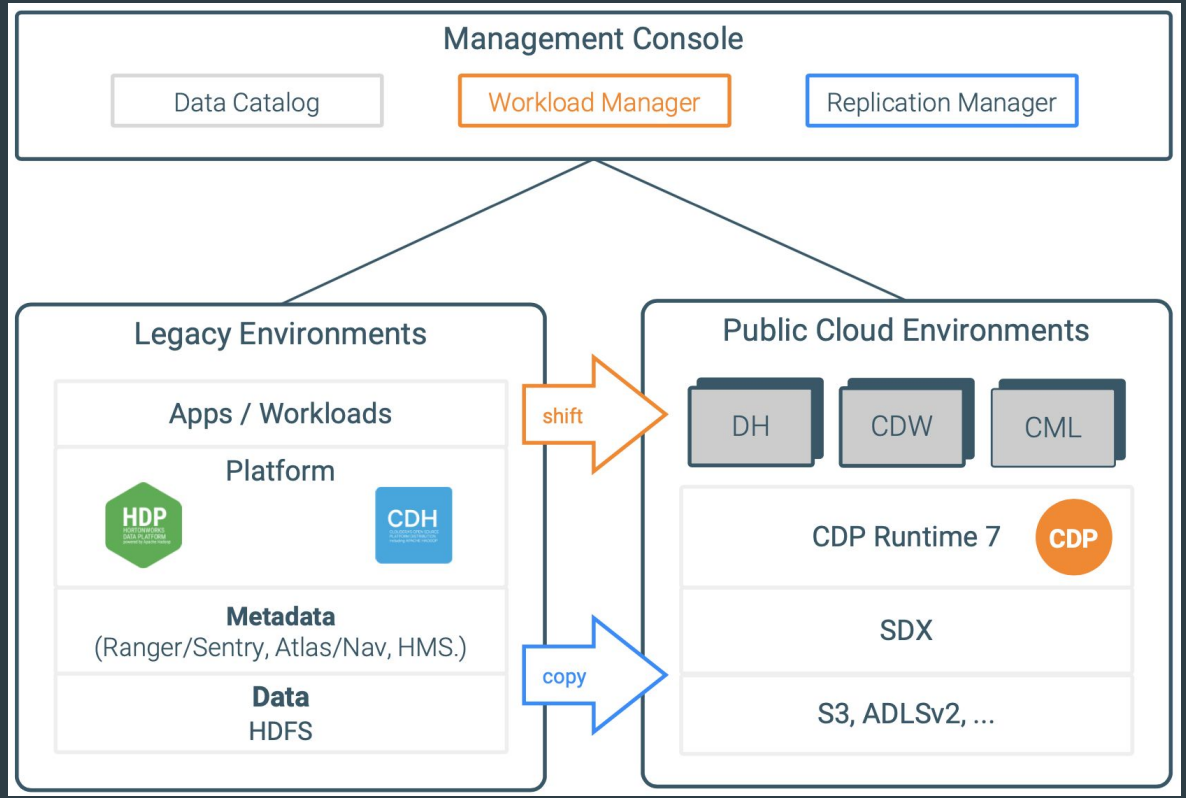
MIGRATING TO A NEW INSTALL



MIGRATE TO PUBLIC CLOUD

Process

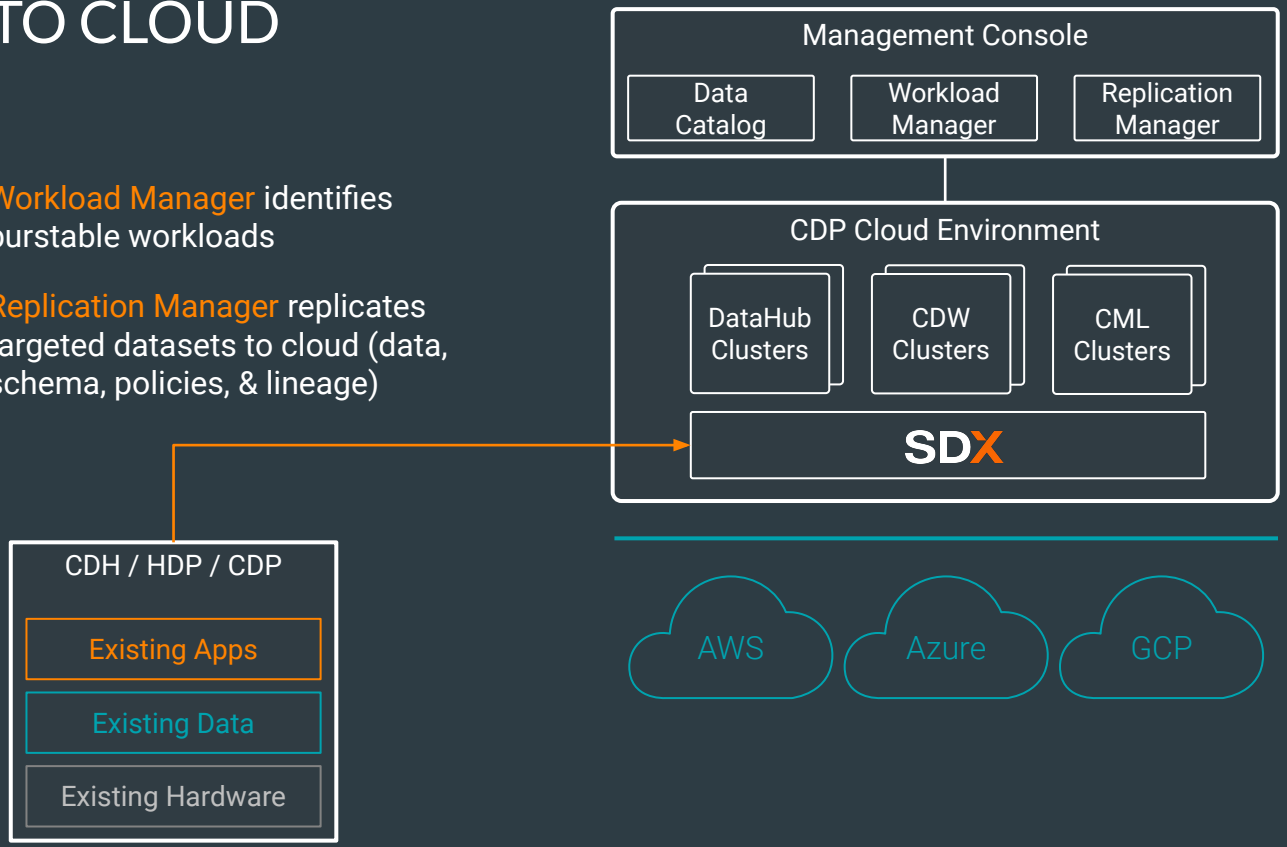
- Set up public cloud environments
- Register classic cluster(s)
- Identify candidate workloads
- Migrate workload data and metadata using Replication Manager (“Burst to Cloud”)
- Test and promote to production



BURST TO CLOUD

Workload Manager identifies burstable workloads

Replication Manager replicates targeted datasets to cloud (data, schema, policies, & lineage)



Packaging, Pricing & Value Proposition

CDP CUSTOMER ADOPTION

Early customer feedback is positive on 3 dimensions...Security | Cost Savings | Performance

CDP eliminates the need to compromise. Its workload management enables an optimal customer experience without overprovisioning. And with SDX, we don't have to sacrifice governance to deliver a great customer experience. With CDP, we get security, customer experience and lower costs.

CDP became a clear choice for two reasons. First, enterprise security and governance, at which Cloudera has always excelled. But also lower costs. CDP's Data Warehousing auto-scales up and down. This means we only use and pay for what we need -- Cloudera's service and cloud infrastructure.

CDP's cloud-native Machine Learning is a great example of an innovative analytics service. It helps our data scientists experiment faster and collaborate better. Our IT organization likes it because it keeps them in control and embraces shadow IT. This is crucial to delivering innovation while ensuring security and governance.

PRODUCTS

- **CDP Core Products** — Today's products
  - CDP PVC Base — Upgraded data and analytics platform
  - SDX — End-to-end security and governance
  - CFM — Flow Management
  - CSA — Streaming Analytics
  - CDSW — Machine Learning
- **CDP Data Hub** — Cloudera's cluster-as-a-service for AWS, Azure, GCP
  - 10 Templates - Optimized for migrating apps to cloud without rewrite
- **CDP Experiences (Improved)** — Cloudera's next generation cloud-native data services optimized for practitioner experience
  - DF **(New)**, DE, DW, OD, ML — Ease of use & management
  - CDP PaaS — Customer choice and control
  - CDP SaaS **(Coming Soon)** — Customer self-service & simplicity
- **CDP PVC Experiences (New)** — Optimized for modernizing apps on prem to accelerate time to value and improving practitioner experience

CDP EXPERIENCES

Practitioner-grade experience, for building and operating multi-function applications

PVC BASE & DATA HUB		CDP EXPERIENCES
Servers	→	*Services*
Monolithic	→	*Modular*
Co-Located Storage & Compute	→	*Separated Storage & Compute*
HW Dependent	→	*SW Defined*
Operator Focused	→	*Practitioner Focused*
Optimized for Existing Applications	→	*Optimized for New Applications*
Static Workloads	→	*Portable Workloads*

CDP Public Cloud Services

Initially AWS; Azure and GCP coming soon**

CDP Data Hub

$0.240/hr*

A service for creating general purpose data management and analytic clusters that enable developers to create custom business applications, includes 24x7 technical support

*Pricing for m5.2xlarge instance type - other instance types available

CDP Data Warehouse

$0.72/hr*

A service for creating self-service data warehouses and the underlying compute clusters for teams of business analysts, includes 24x7 technical support

*Pricing for r5d.2xlarge instance type

CDP Machine Learning

$0.68/hr* + $399/user/mo

A service for creating self-service machine learning workspaces and the underlying compute clusters for teams of data scientists, includes 24x7 technical support

*Pricing for m5.2xlarge instance type - other instance types available

Data Warehouse Experience on AWS - Minimum Hardware Requirement

Data Warehouse Experience (Hosted on EKS)				
Resource Type	Minimum Count	Scaling	Persistent	Purpose
m5.2xlarge	3+	Auto	Yes	Database Catalog - Shared Services Nodes (Max = 20 Nodes)
r5d.4xlarge	2	Auto	No	Reserved Capacity - Compute Nodes (Default = 2, see notes below)
r5d.4xlarge	user defined	Auto	No	Virtual Warehouse Compute Nodes (T Shirt sizing)
db.r4.large	1	N/A	Yes	RDS DB Instance (PostgreSQL)

CDP PRIVATE CLOUD PACKAGING

MAX edition (roadmap)

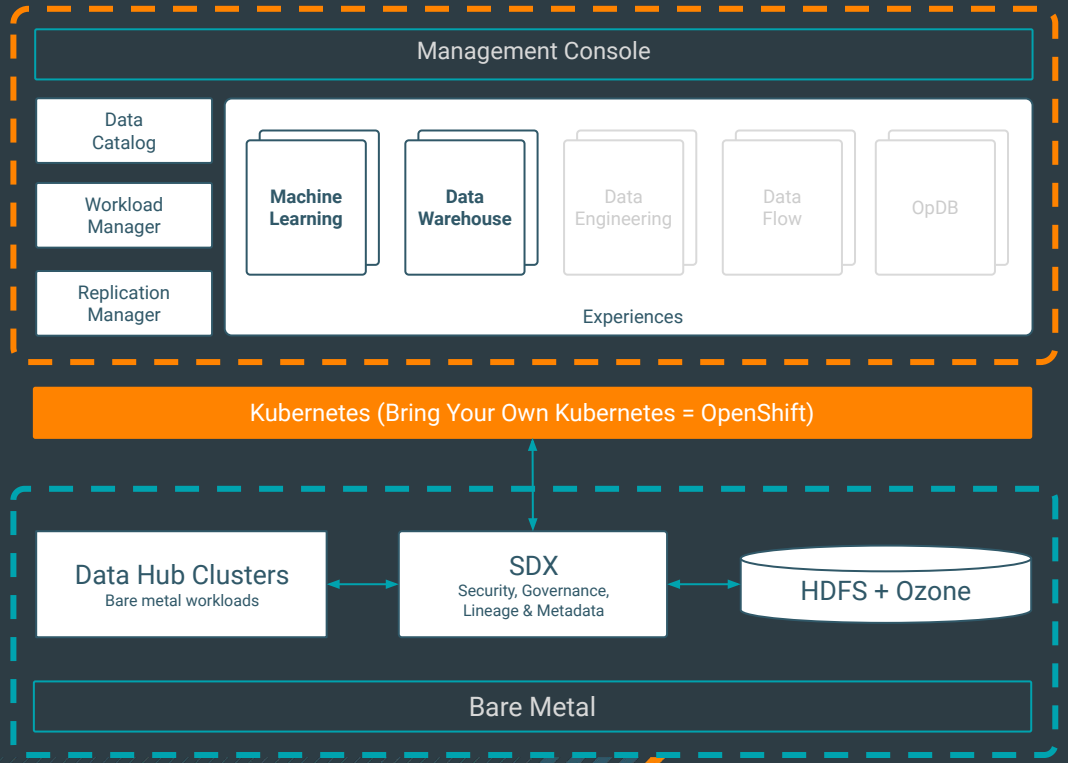
- PLUS +
- All Experiences
- Additional features (TBD)

PLUS edition

- BASE +
- Control Plane (DC, WM, RM)
- Any two (2) Experiences
- Analytic compute multiplexing
- Burst to Public Cloud experiences

BASE edition (aka DC)

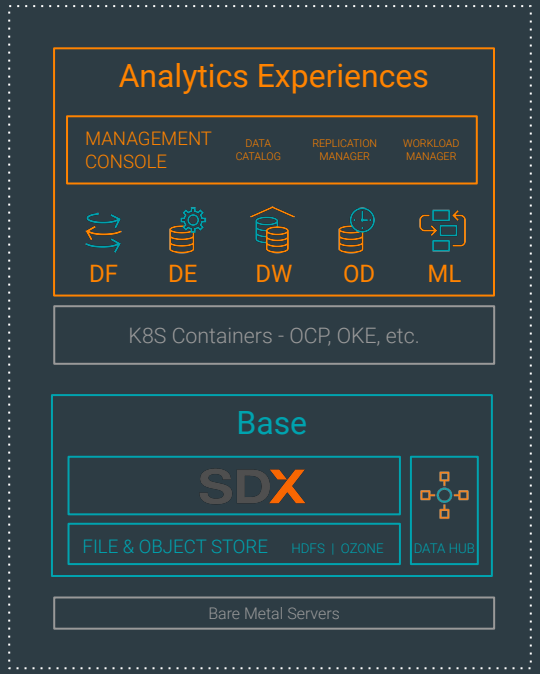
- Data Hub clusters
- Storage management
- SDX



CDP PRIVATE CLOUD – HARDWARE REQUIREMENTS

	Bare metal or VMs		OpenShift 4.3 Cluster	
Component	**Minimum**	**Recommended**	**Minimum**	**Recommended**
Number of servers	6	8	14	20
CPU (cores)	16 cores	24 cores	16 cores	32 cores
RAM (GB)	32GB	128GB	128GB	384GB
Storage HDD (Num x Capacity)	12 x 2TB	12 x 4TB	2TB (SATA)	4TB (SSD/NVMe)
Network	**Minimum**: 1Gbps guaranteed bandwidth between any 2 nodes in the cluster (this is worst case bandwidth, under full load). Recommended: 10Gbps guaranteed bandwidth. Topology Recommendation: 25+GbE Spine/Leaf, max 4:1 oversubscription between the spine and leaf switches.			

CDP PRIVATE CLOUD: WHAT IS PRICED & LICENSED?



Analytic Experiences

- Cloudera Compute Unit
  - CCU = 1 core + 8 GB RAM
  - 'Container' cores & memory
- Storage
  - By TB as needed (HDFS & Ozone)

Base

- Node
  - 16 cores, 128 GBs RAM, 48 TBs
- CCU & TB above node cap
- 'Physical' node, cores & memory

Helpful Notes

Red Hat Openshift required

No caps in Plus; count cores and memory

Plus includes access to Base software

CCU defined the same in Base & Plus -- the price & what is counted is different

CDP PRIVATE CLOUD INFRASTRUCTURE GUIDE

DELL Technologies CLOUDERA intel

Dell EMC and Intel Infrastructure Guide for Cloudera Data Platform Private Cloud

Abstract

This white paper provides infrastructure configuration and strategy guidance for customers planning new or upgraded data center deployments of Cloudera Data Platform Private Cloud on Intel architecture.

May 2020

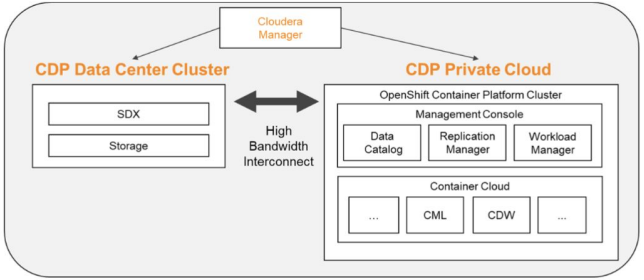


Figure 2 CDP Private Cloud components

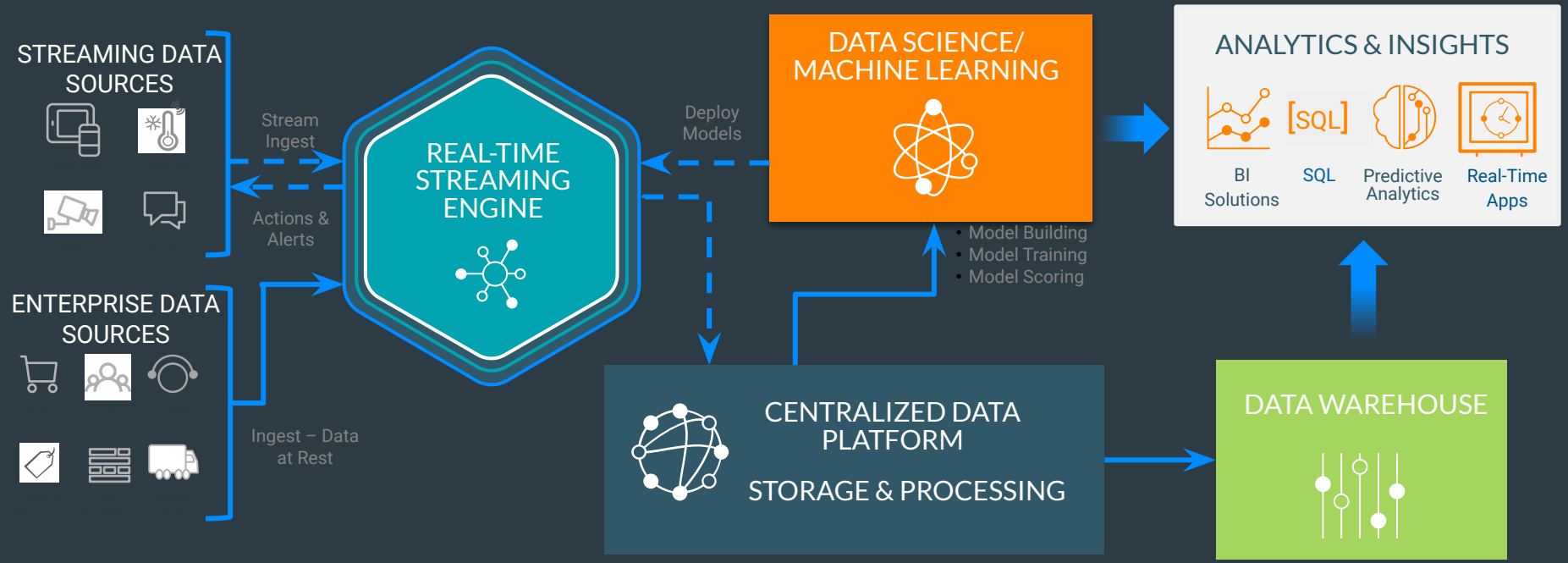
A full installation consists of two cooperating clusters – CDP Data Center, and a private cloud platform running OpenShift. The primary functions are:

1. **Cloudera Manager** provides management, monitoring, and deployment for both clusters.
2. The **CDP Data Center** cluster provides the core Cloudera **SDX** services and **HDFS/Ozone storage.**
3. The **OpenShift Container Platform** provides the core capabilities necessary for the containerized cloud environment, including Kubernetes and its control plane.
4. The **Management Console** is the control plane for the container cloud, and provides the **Data Catalog**, **Replication Manager**, and **Workload Manager** services.
5. The **Container Cloud** runs the workloads, including **CDW and CML.** More workloads will be added over time.
6. A **High Bandwidth Interconnect** connects the two clusters to provide access from the container platform to storage on the CDP Data Center cluster.

https://infohub.delltechnologies.com/section-assets/infrastructure-guide-cdp-private-cloud

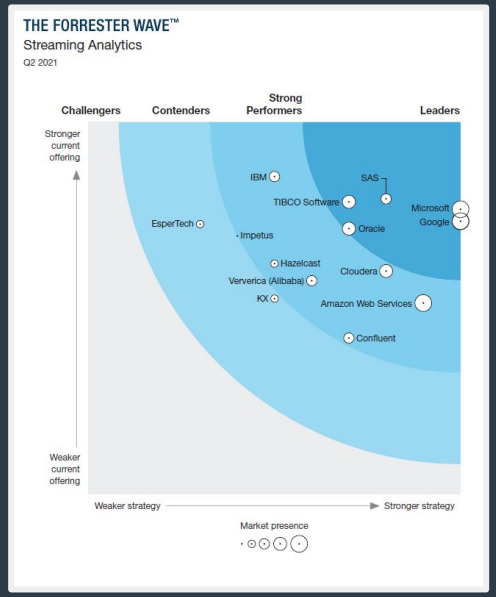
ENABLING ANALYTICS & INSIGHTS ANYWHERE

Driving Enterprise Business Value

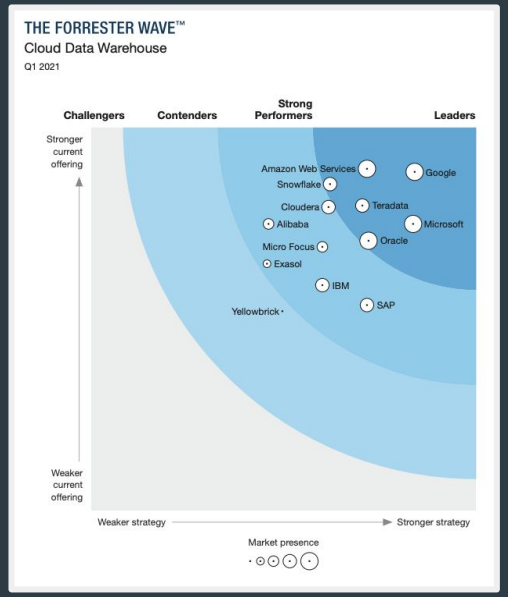


CLOUDERA RECOGNIZED IN FORRESTER WAVES

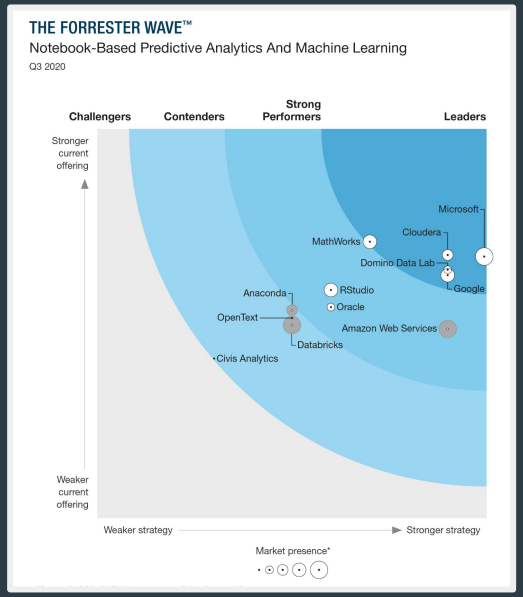
Streaming Analytics



Data Warehouse

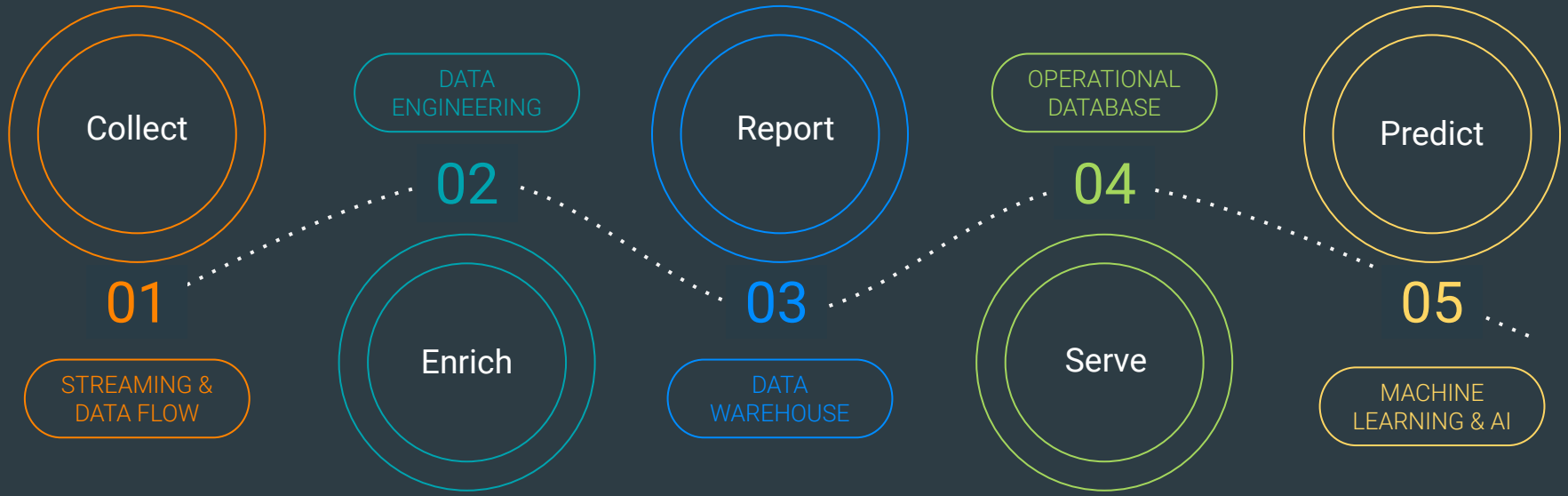


Machine Learning



CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter

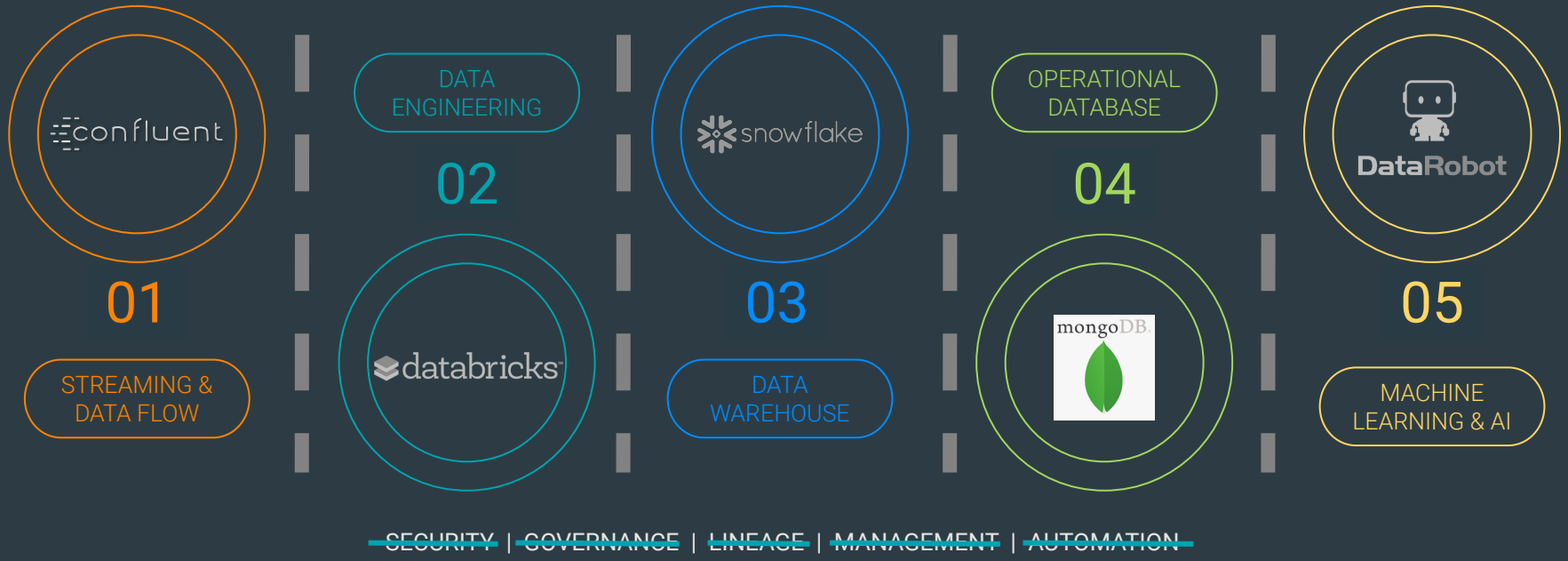


CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

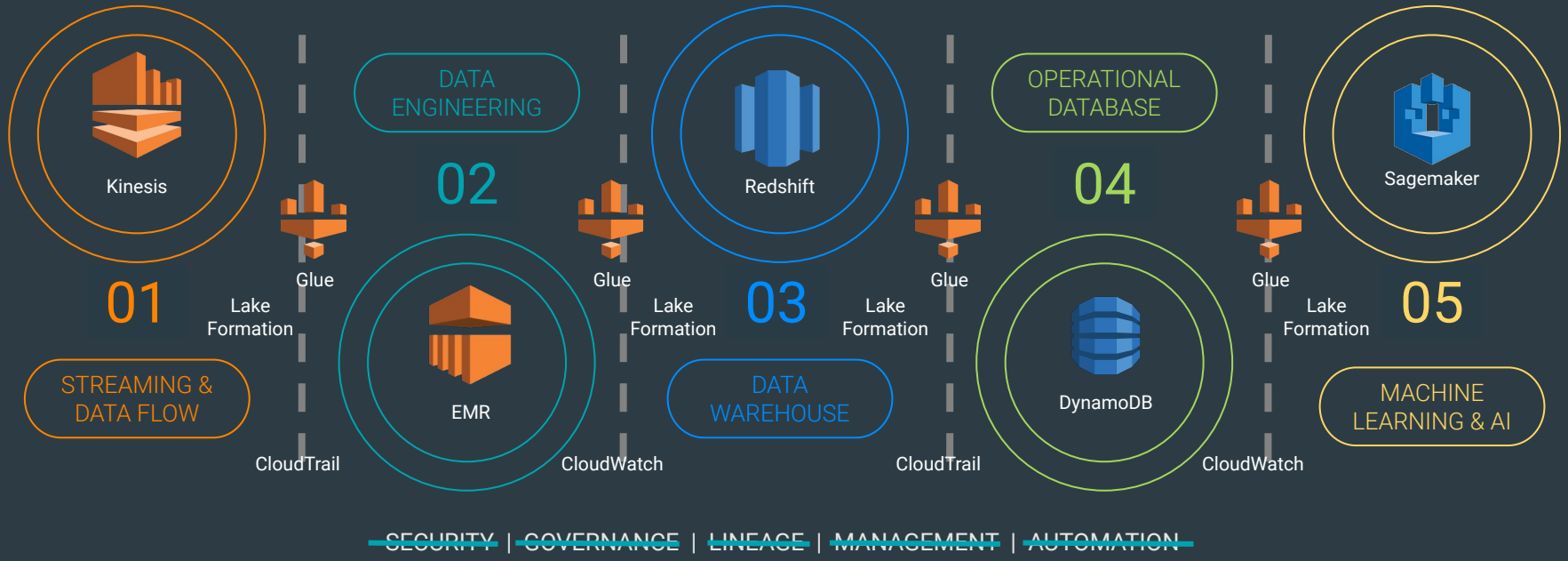
POINT SOLUTIONS HAVE AN INTEGRATION TAX

Security & governance is an afterthought



POINT SOLUTIONS FROM PUBLIC CLOUD PROVIDERS

Building blocks



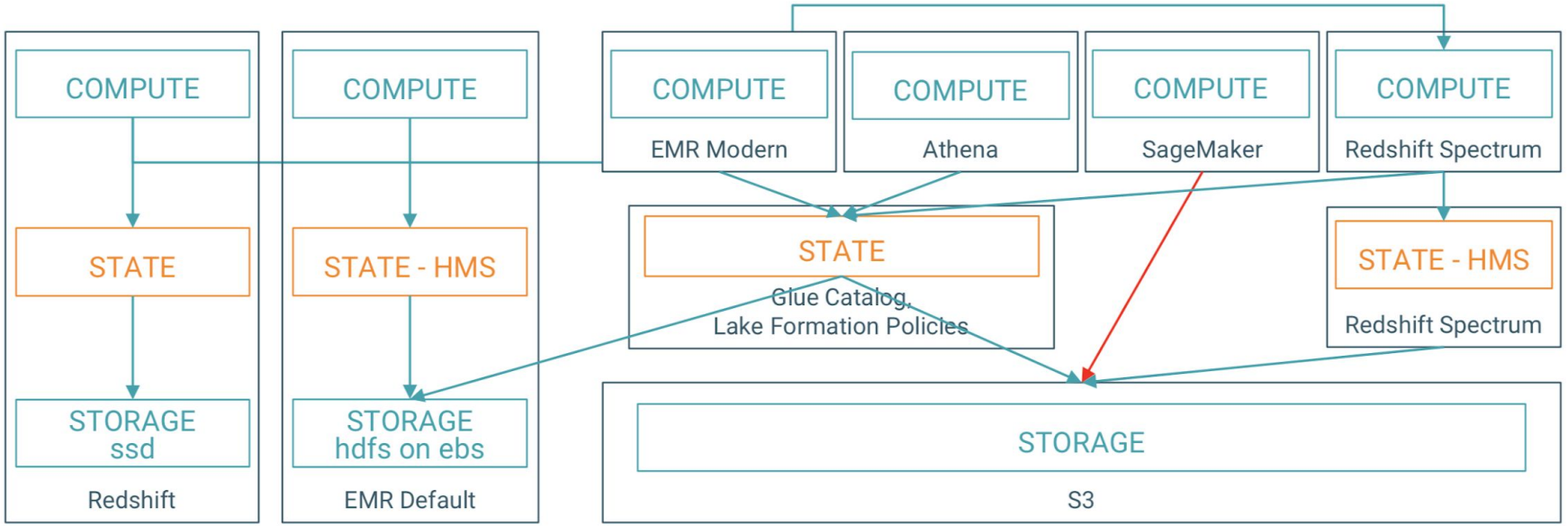
Key points to consider – CDP Vs AWS Native Offerings

- **Hidden Costs** - Customers can accrue significant costs associated with EMR especially for long running applications – engineering, compute, storage
- AWS offerings are optimized for a small to mid-sized cost-conscience company with no Hadoop expertise in-house.
- **Security** - Ranger, Atlas, Knox and Kerberos are not available in EMR without a lot of manual configuration and using a third party to mask sensitive data or track data lineage. GDPR-like functionality is not possible to address in EMR.
- **Portability** - lift&shift from an on-prem to EMR is very impractical and needs a lot of hacking around
- **No "STOP" feature** - EMR's biggest pitfall is the inability to shut-down and restart when needed. It needs to be reprovisioned.

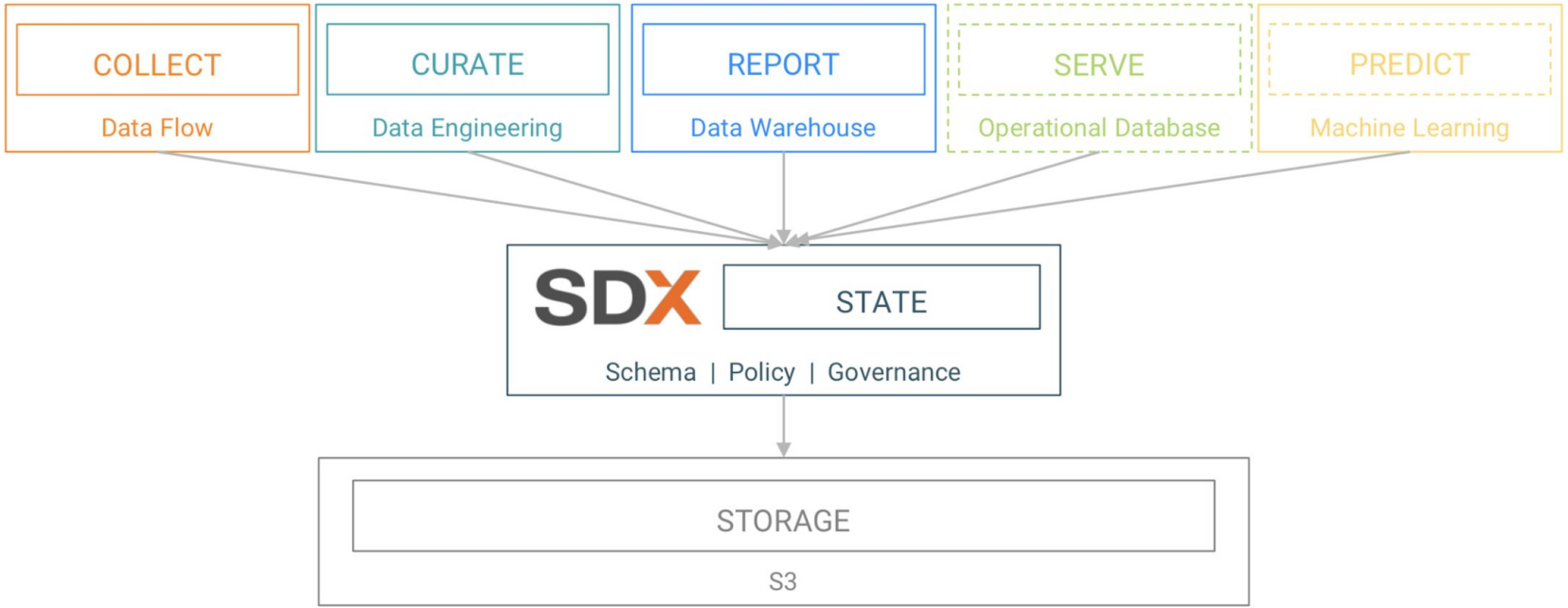
	Cloudera	AWS EMR
Unified Security / Governance	✔	Lake Formation*
Data Engineering	✔	EMR*
Machine Learning	✔	Sagemaker*
Data Warehouse	✔	***Redshift****
Ingest / Streaming	✔	Kinesis*
Unified Data Platform	✔	
Hybrid Cloud	✔	Outpost*
Open Source	✔	
Long Running Cost Optimized	✔	

CORE PROBLEM WITH POINT CLOUD SOLUTIONS

Security and governance is an “after thought”



CDP ARCHITECTURE WITH SDX



CLOUD DATA WAREHOUSE PERFORMANCE TESTING

Cloudera Delivers Better Price Performance

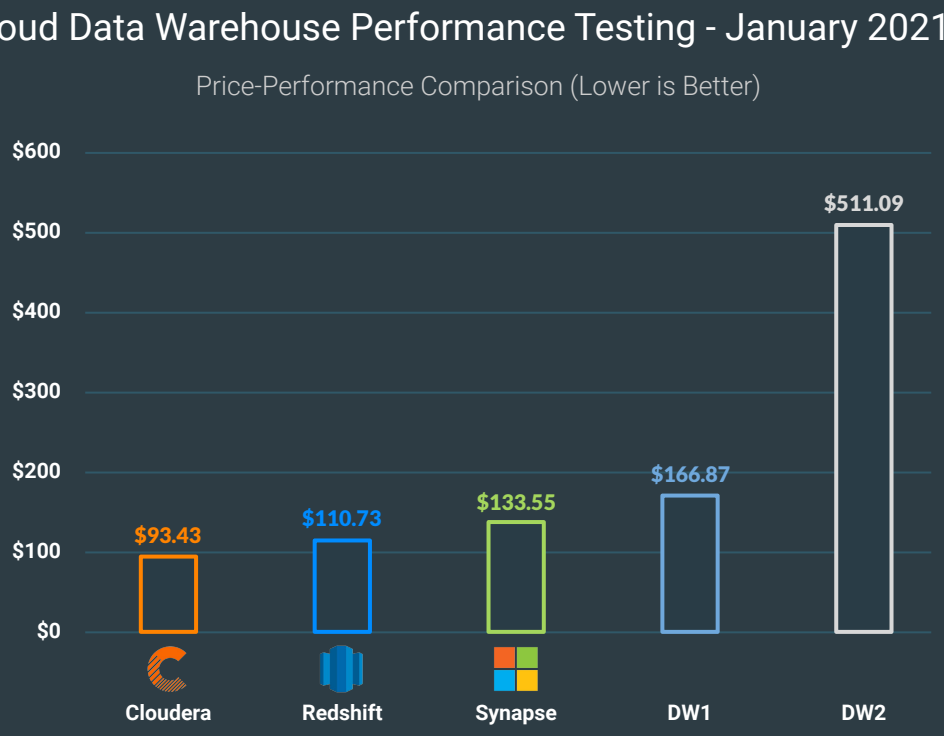
Industry standard TPC Benchmark

20% lower costs than Amazon Redshift

40% lower costs than Microsoft Synapse

80% lower costs than “DW1”

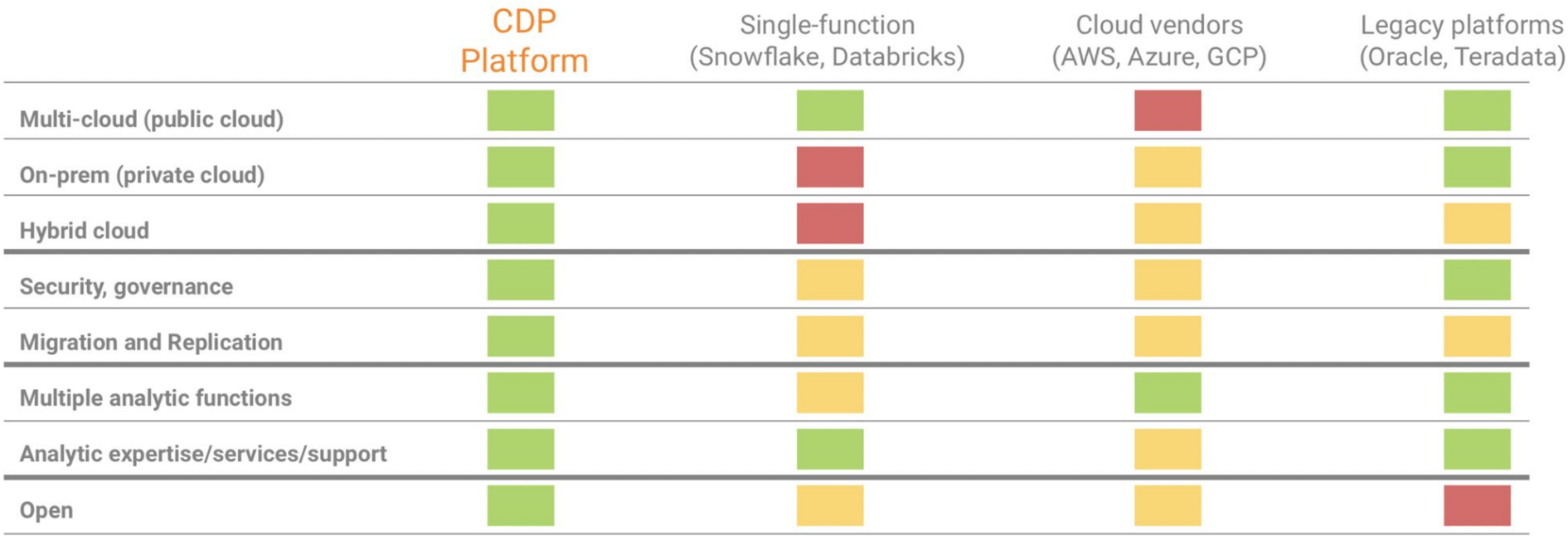
550% lower costs than “DW2”



More can be learned about the TPC-DS benchmark at http://www.tpc.org/tpcds/.

Prepared by: McKnight Consulting Group, www.mcknightcg.com January 2021

COMPARING CDP WITH OTHER VENDORS



What's in it for you?

DATA LIFECYCLE	vs	SINGLE FUNCTION
HYBRID & MULTI	vs	CLOUD ONLY
SECURE & GOVERNED	vs	BASIC SECURITY
OPEN SOURCE	vs	PROPRIETARY LOCK-IN
PLATFORM	vs	POINT SOLUTION

DIGITAL TRANSFORMATION

Key Bookmarks

CDP Upgrade/Migration
https://docs.cloudera.com/cdp/latest/upgrade.html

Reference Architectures
https://docs.cloudera.com/documentation/other/reference-architecture.html

Pricing Related https://www.cloudera.com/products/pricing.html

Partner Portal
https://my.cloudera.com/partner-portal.html

Call to Action

Register onto Partner Portal
https://my.cloudera.com/partner-portal.html

CDP On-boarding Setup your CDP Demo & working environment

- Fill-in form for enabling licenses
- Get your AWS/Azure account for infrastructure
- Get your pre-requisites ready
- Use Starter kit to setup CDP

Joint Customer Pursuits & Pro-active engagements on CDP Upgrade

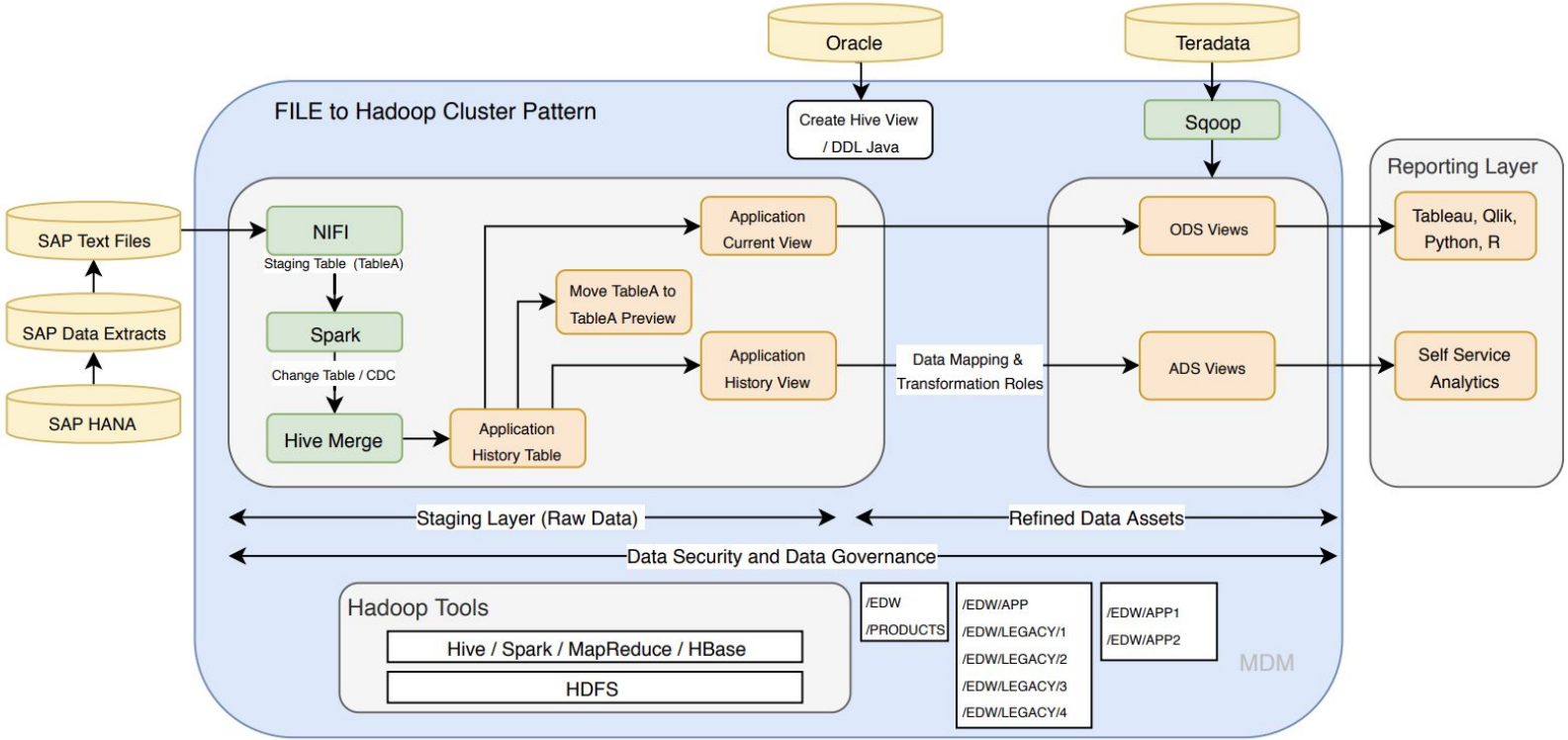
More Enablement & Hands-on Workshops

Q&A

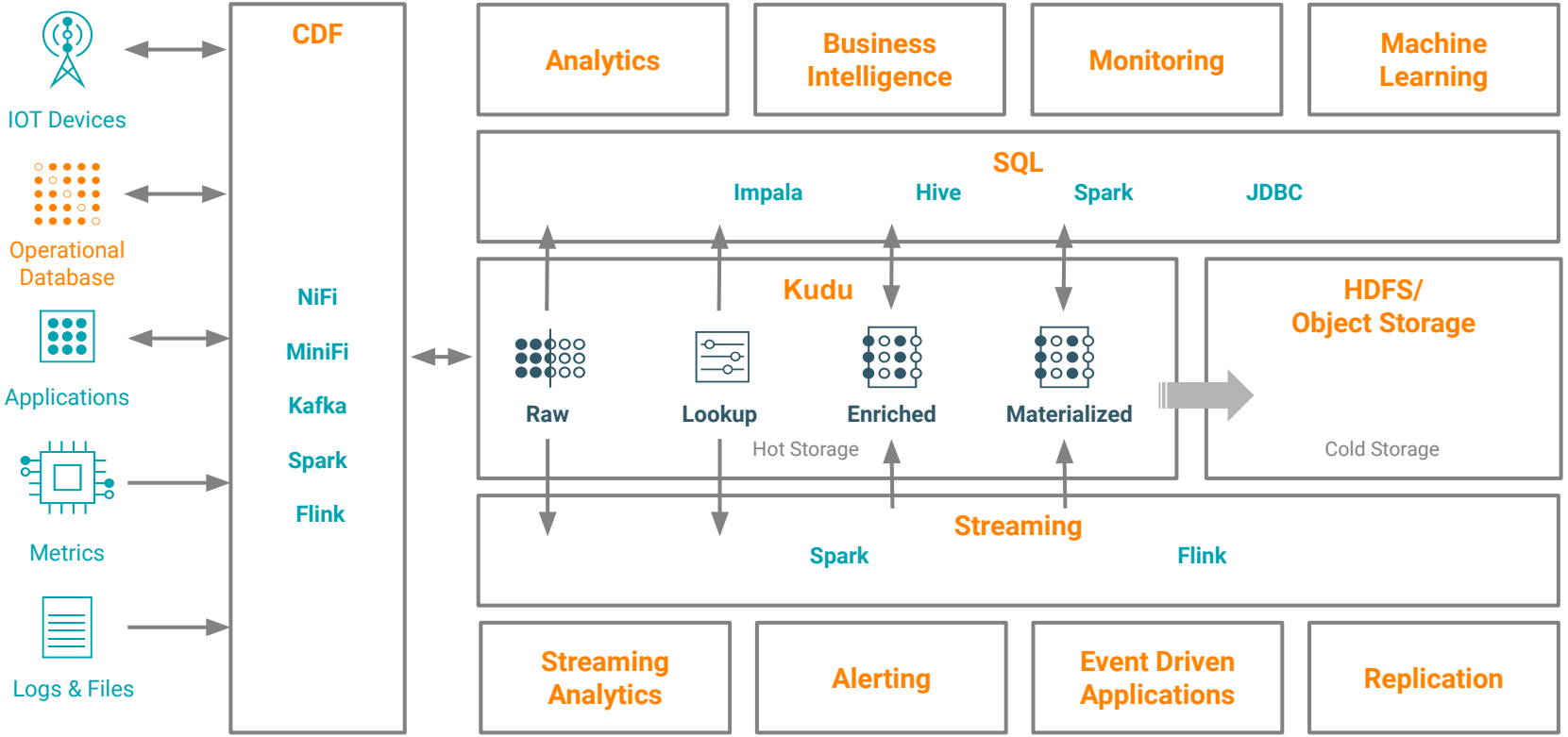
THANK YOU

CLOUDERA

ENTERPRISE DATA WAREHOUSE (DATA LAKE)



REAL TIME DATA MARTS



STORAGE

HERE ARE YOUR STORAGE OPTIONS WITH SQL USE CASES

HDFS
Best Performance via local reads, +100TB working datasets, and years of Production workloads with strong data durability. Review block size based on workload needs. Reduce small files to increase scale, performance, and stability

OBJECT STORE / OZONE
Great for slow transforming data and archival use-cases, but also enables decoupled compute / storage architecture enabling Compute only Clusters (Cloud Burst)

KUDU
Enables Ad-hoc analytics with pipeline latencies of 10 seconds, data mutability, eliminate HDFS small file problems, and Lambda architectures. Pairs well with Impala, Nifi, and Spark(Hive)

ENGINES and FILE FORMATS

HERE ARE YOUR FILE FORMAT OPTIONS WITH SQL USE CASES

ENGINE SELECTION
Use Hive/Spark for ETL, Impala for BI

FILE FORMATS and COMPRESSION
Prefer ORC / Parquet compressed formats. Compressed Text can be used for staging data.
Avro helps for data formats that might change.
Start with ORC/Zlib, Parquet/Snappy, but as data and use cases vary dramatically, adjust as necessary if the tradeoff of disk space vs cpu for (de)compression is acceptable (zstandard)

THIRD PARTY TOOLS
Tableau, Arcadia Data, Microstrategy, Cognos, Tibco Spotfire, and the popular BI tools in this space support JDBC and ODBC SQL connectivity for interactive Visualizations and Dashboard creations

HUE INTERFACE
Hue is a web-based interactive query editor that enables you to interact with data warehouses

CLI
Impyla supports Impala and Hive. Impala-shell, beeline standalone (WIP)

NDA

	1H 2021 (Released)	2H 2021 (Coming Soon)
Platform Readiness	• VPC Endpoints • Azure Single Resource Group • Control Plane Audit & RBAC • Transparent Tunnel • YARN Scheduler Migration • HBase Replication • Cloud Resource Tagging • Physical HW redundancy • SSO over JDBC/ODBC • Custom AMIs (SDX, Azure)	• In-Place Upgrades • High Availability • Simple storage authorization • Simplified onboarding and automation
Regional Control Plane Availability	• US	• Germany, Australia
Compliance Certifications	• FIPS, SOC2 Type II, TISAX, ISO27001	• Pre-work for FIPS 140-2, FedRAMP, and HITRUST
GCP	• Data Hub on GCP	• All Data Hub Templates on GCP
CDP Experiences	• COD • CDF Data Service on AWS	• CDF Data Service on Azure
Hybrid Cloud Use Cases	• CDH & HDP Replication to cloud	First 2 hybrid use cases (Phase 1) • Data Syndication • Application Migration

NDA

	PVC Base 7.1.6 (Released Mar '21)	PVC Base 7.1.7 (Released Aug '21)	PVC Experiences 1.3 (Coming Soon)
General enhancements & upgrade blockers	• Accumulo support • Kafka-only clusters • Phoenix transactions support • RHEL 7.9 • MySQL 8 • PostgreSQL12 • CDS 3.0	• In-place upgrade with Isilon • In-place upgrade in FIPS mode • CVE remediations & security fixes • Kafka topic lineage via Atlas • RHEL/CENTOS 8.2 • Ubuntu 20 • MariaDB 10.3, 10.4 • Oracle 19c (19.9) • Oracle 11 JDK	• CDE Experience • Experiences Compute Service (ECS) • Control Plane Private Experiences CLI • Grow / Shrink / Upgrade ECS cluster • Secure Logs / Log Redaction • MySQL / Oracle DB in Base cluster • WXM Integration
CDH-specific upgrade blockers	• Rollback for CDH 5 upgrades • Ranger audit filters improvements • YARN compatibility enhancements (pt 1)	• In-place upgrade from CDH 6.1 - 6.3 • Rollback for CDH 6 upgrades • YARN compatibility enhancements (pt 2) • HDFS Audit improvements with Ranger • Impala row filtering via Ranger	
HDP-specific upgrade blockers	• In-place upgrade from HDP 3.1 • Ambari to CM Migration • Rollbacks for HDP 2 upgrades • Hive compatibility enhancements	• Rollbacks for HDP 3 upgrades • Replication Manager - Knox Integration • Sidecar migration from HDP (via HMS Mirror Tool)	