CLOUDERA

Cloudera Data Platform - CDP

Presales Deep Dive

The session will begin shortly

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

Ventana Research — *January 2021*

Cloud Data Ecosystems

Gartner — *January 2020*

Enterprise Intelligence Platforms

451 Research — *December 2019*

Cloudera Data Platform (CDP)

FEIBUSTECH — January 18, 2021

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

Ventana Research — January 22, 2021

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

ENTERPRISE DATA CLOUD

ENTERPRISES ARE EMBRACING PRIVATE CLOUD

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

67%

Of enterprise workloads run on public and private cloud implementations

84%

Of enterprises report repatriating some workloads from public cloud

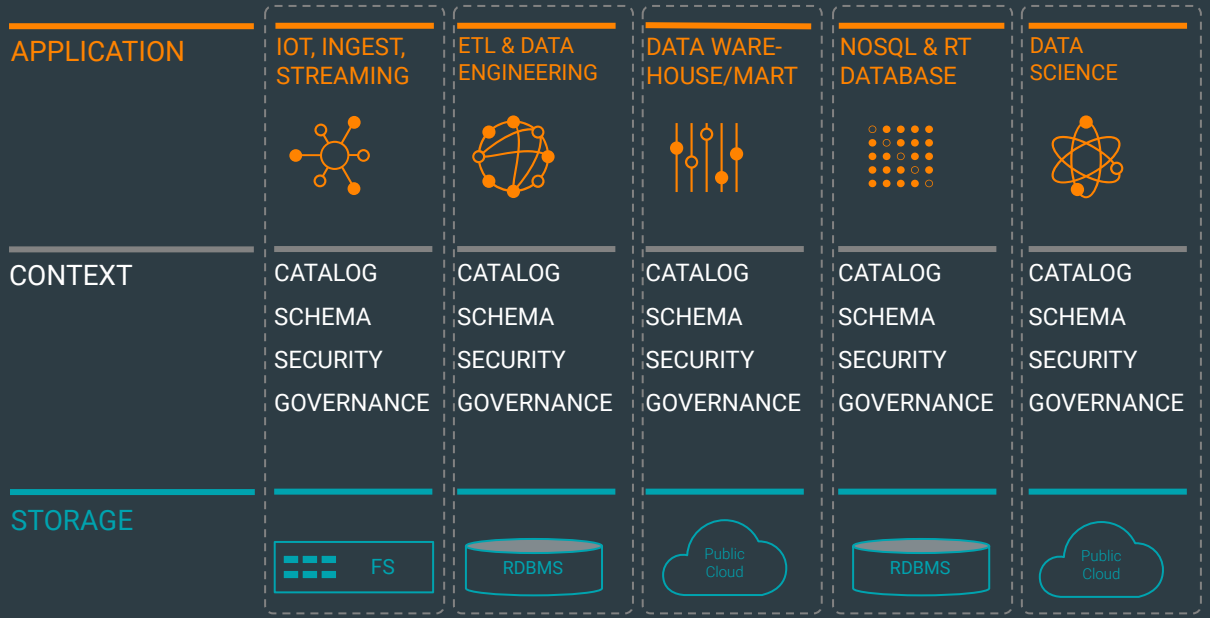
52%

Of repatriated workloads move to private clouds

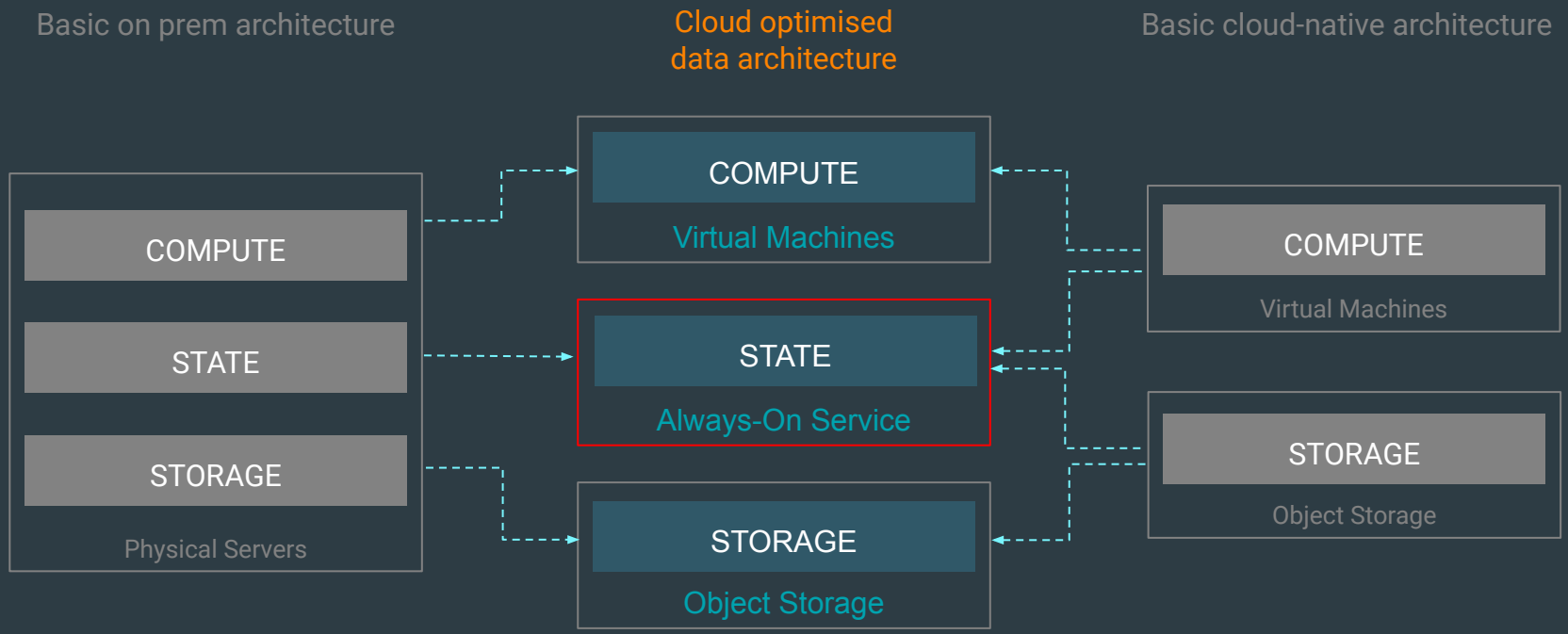
"As enterprise customers gain cloud expertise they're placing investments in private cloud solutions for increased security, compliance, performance, control and cost savings. Private clouds often act as a stepping-stone in the hybrid cloud journey."

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

DATA AND INSIGHT SILOS

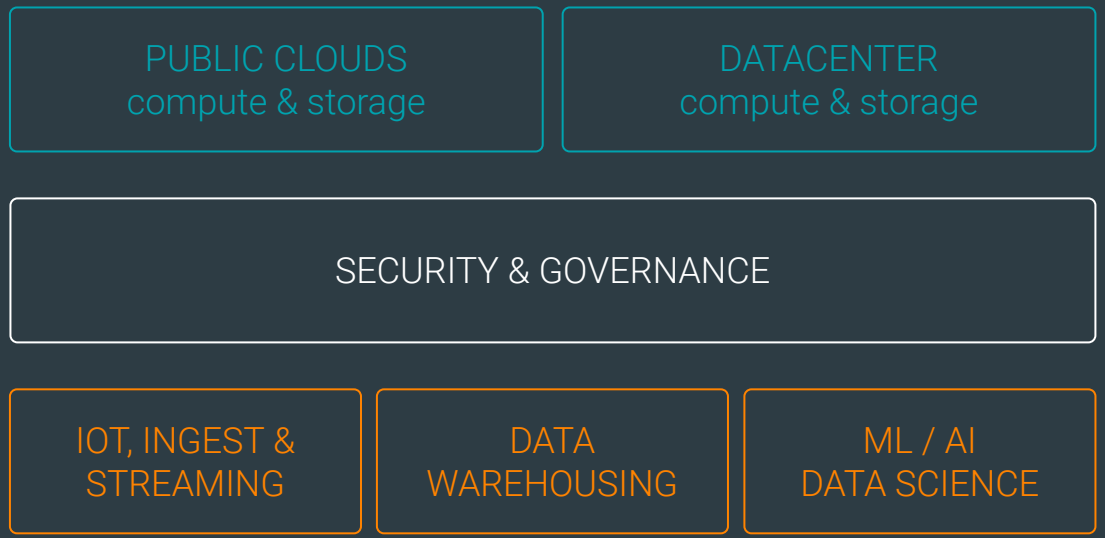


SEPARATE STORAGE AND COMPUTE *AND STATE*



ENTERPRISE DATA CLOUD DESIGN PRINCIPLES

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

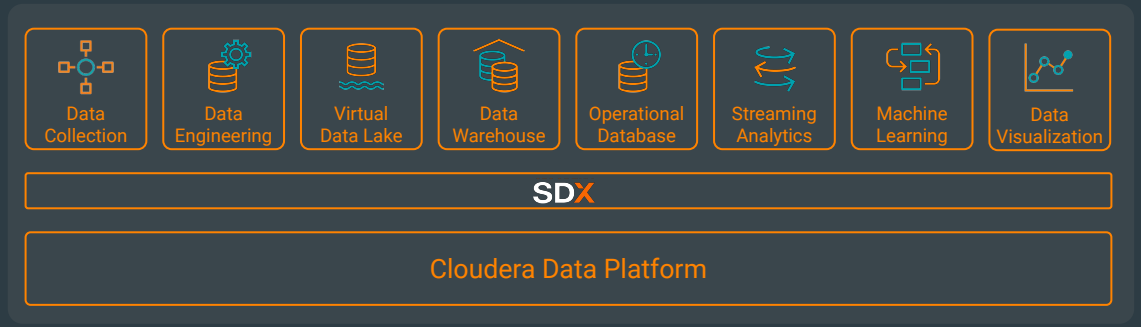


CLOUDERA DATA PLATFORM

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

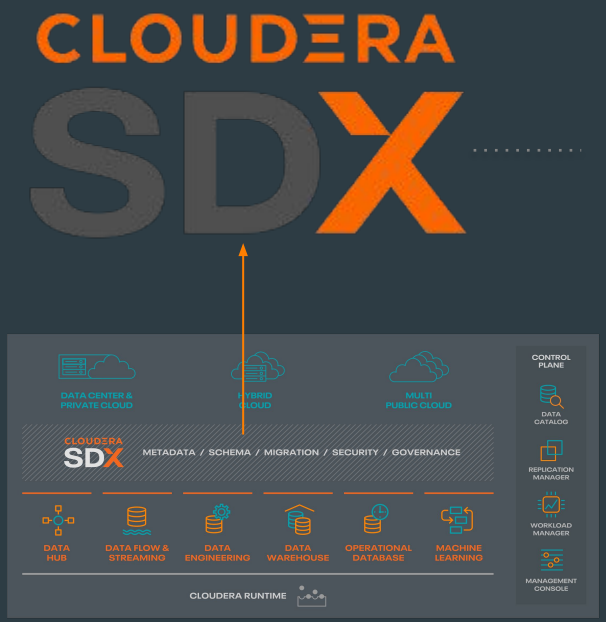
Hybrid & Multi-Cloud
to leverage existing investments and reduce risk

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

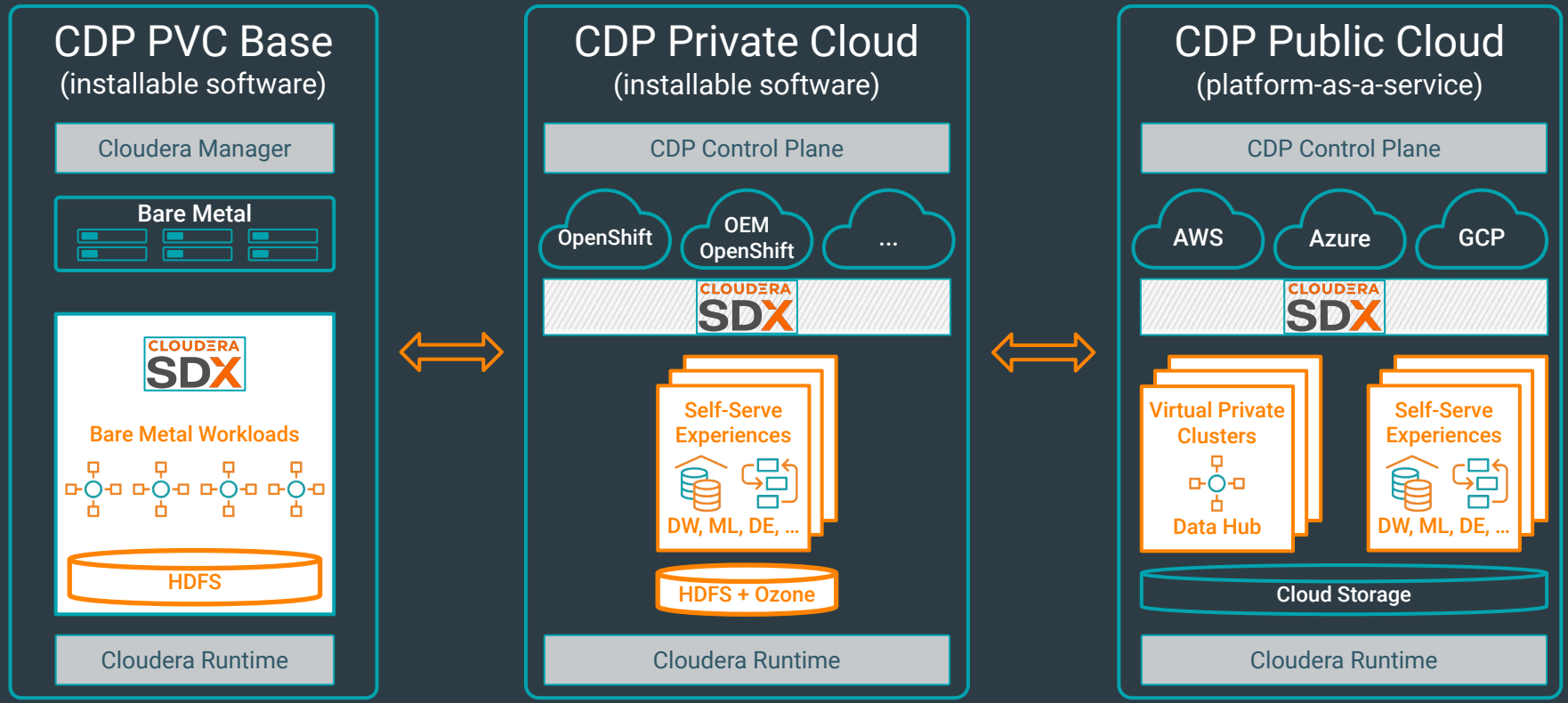
Schema: automatic capture and storage of any and all schema and metadata definitions as they are used and created by platform workloads

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

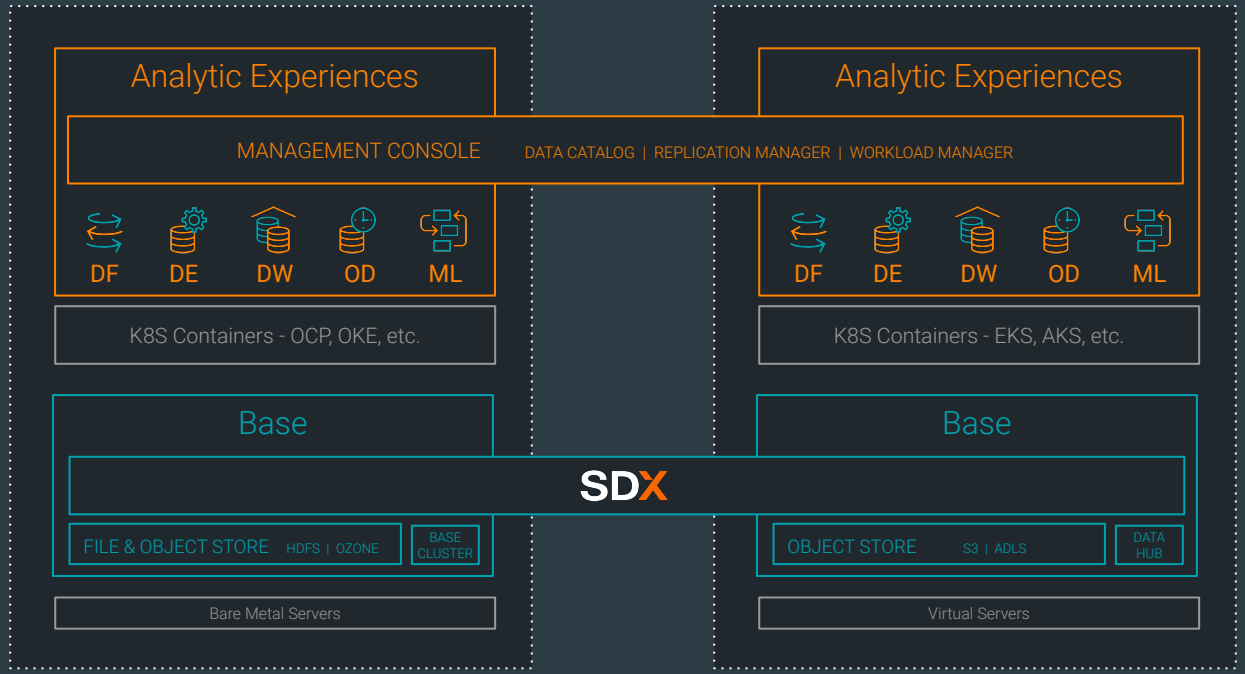
CDP - ONE PLATFORM, THREE FORM FACTORS



CDP HYBRID CLOUD

Consistent operations and analytics experiences across private and public clouds

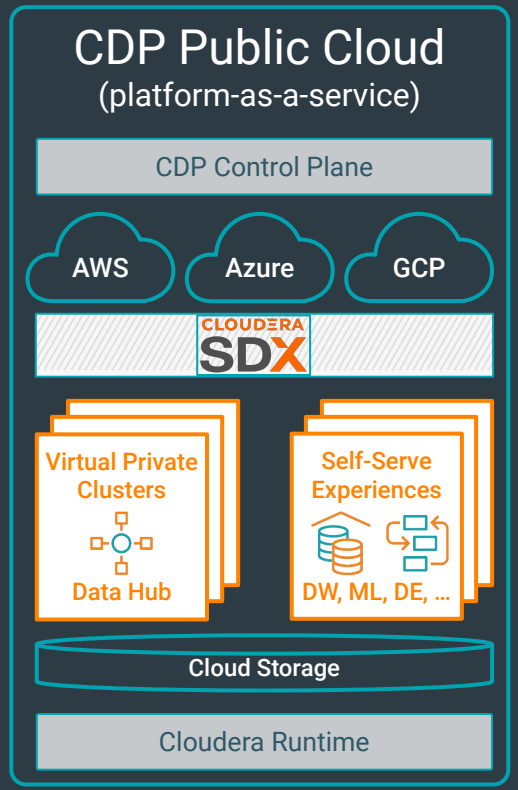
CDP Private Cloud



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

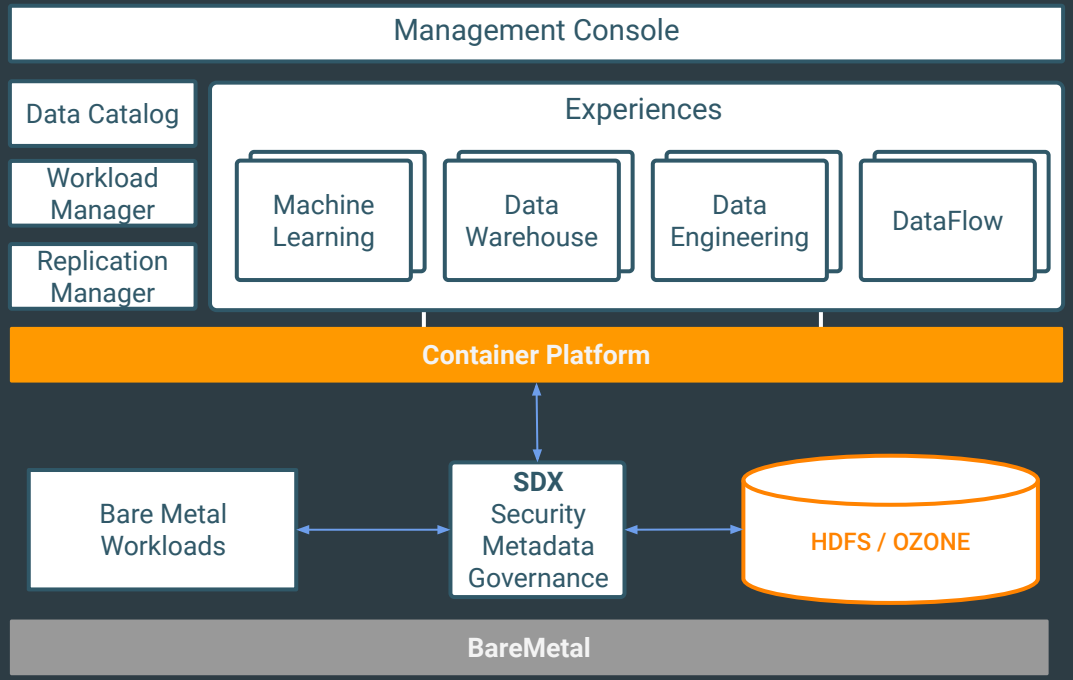
Burst to Cloud

- 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 DATA SERVICES

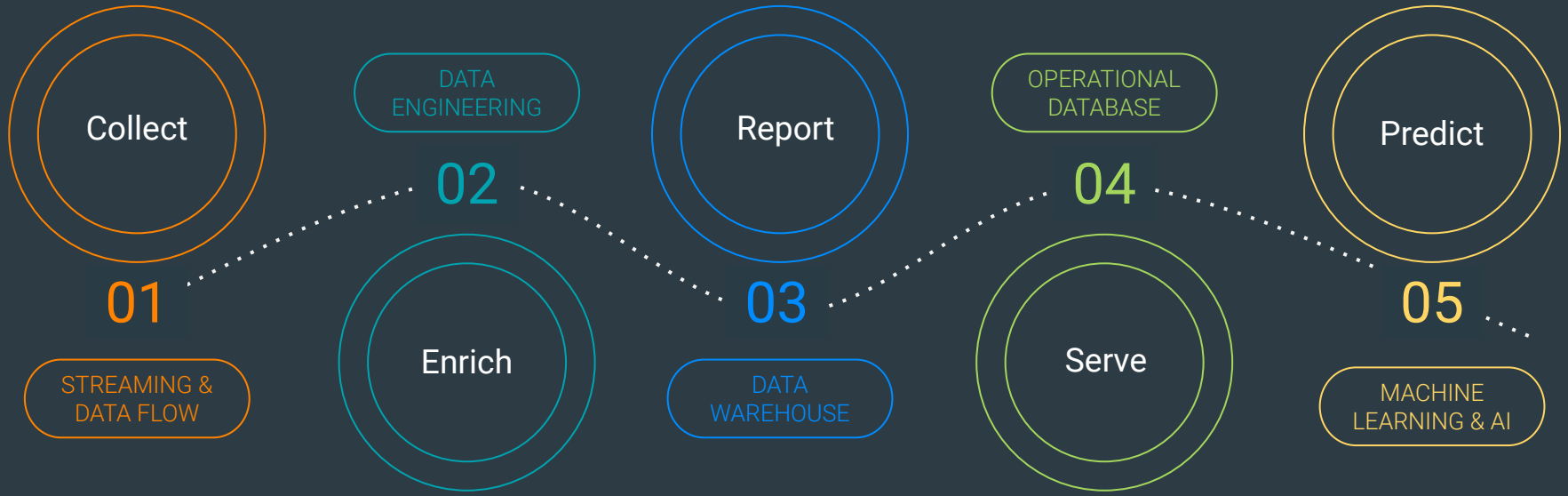
What are the consumption options?

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

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

CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter

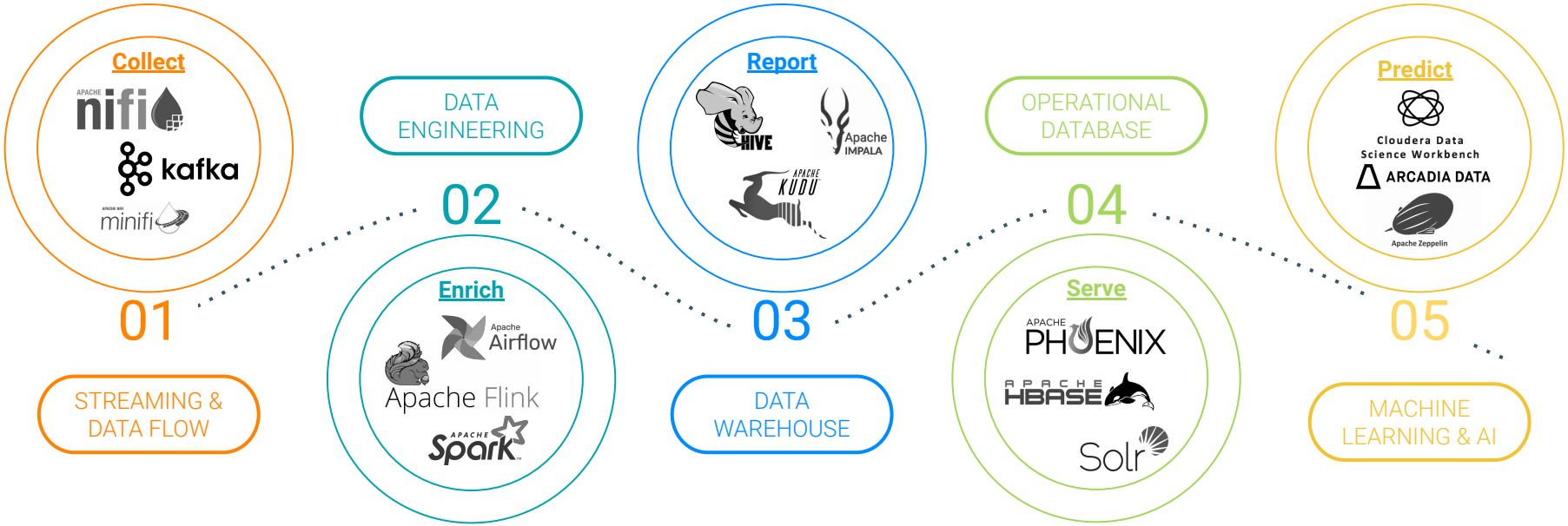


CLOUDERA SDX

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

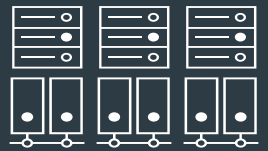
CLOUDERA DATAFLOW SERVICE “EXPERIENCE” (CDF)

A complete SDLC experience for data pipelines from dev through prod

Cloudera DataFlow

Servers/Clusters/Runtimes

(NiFi, Kafka, Flink)



Cloudera DataFlow Experience

Flows (H1) — Schema Aware

Topics (H2) — Schema Aware

Stream Apps (H2) — Schema Aware

Data Pipelines

Simplifies app development by enabling developers to focus on business logic

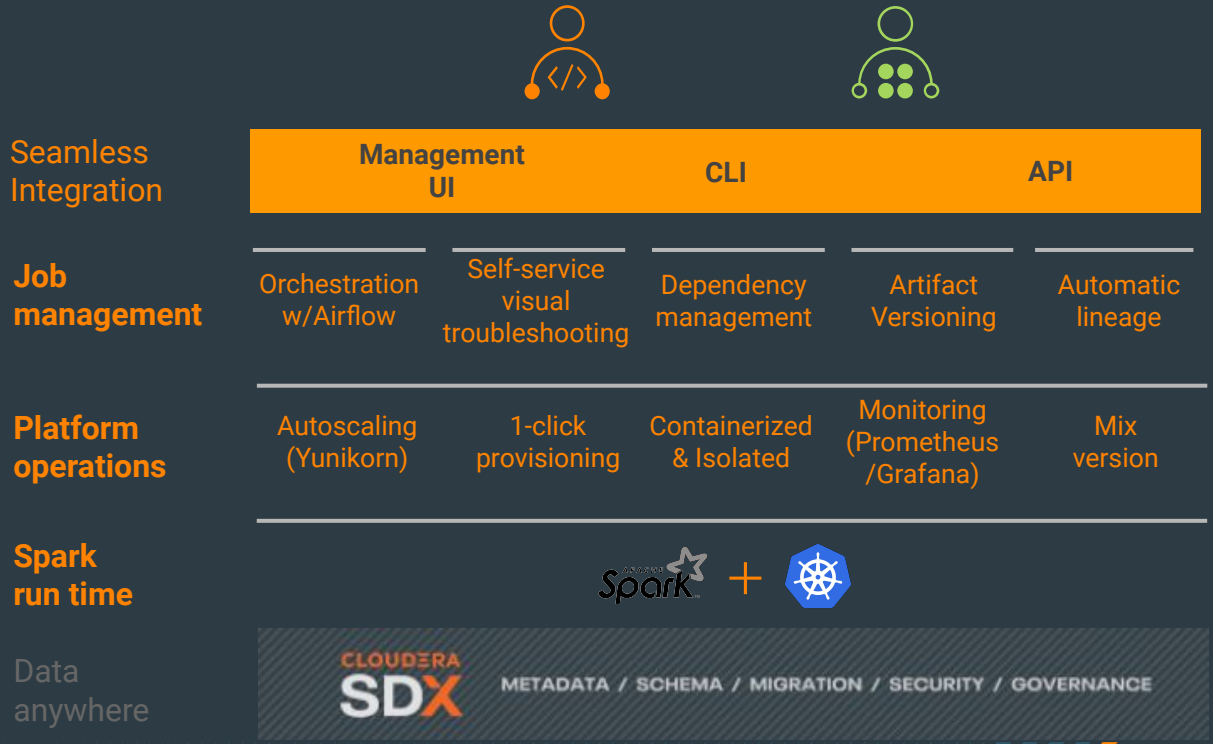
CONNECTING THE DATA LIFECYCLE

Enriching the data lifecycle journey

CLOUDERA SDX

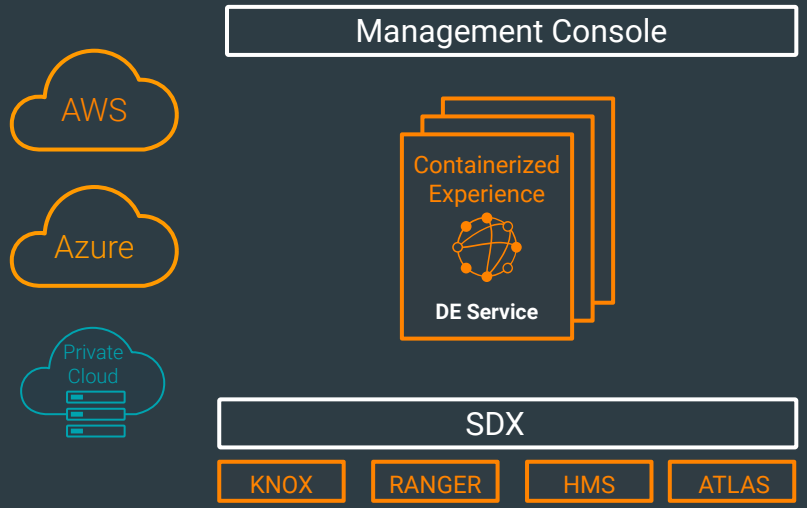
SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

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

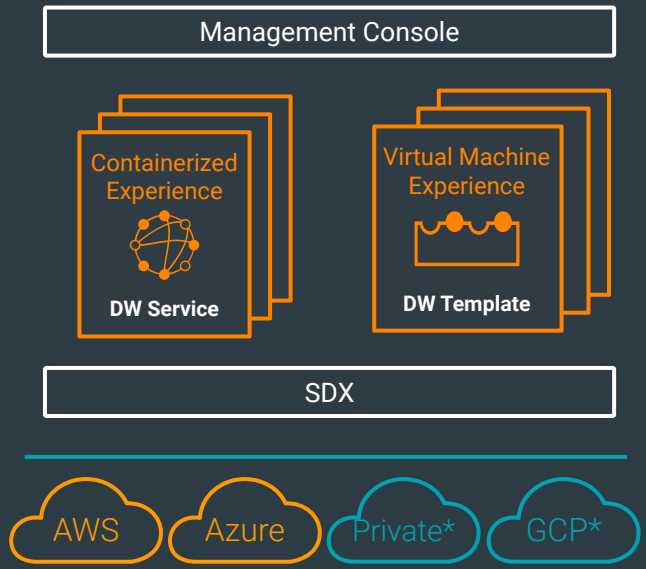
CLOUDERA SDX SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

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

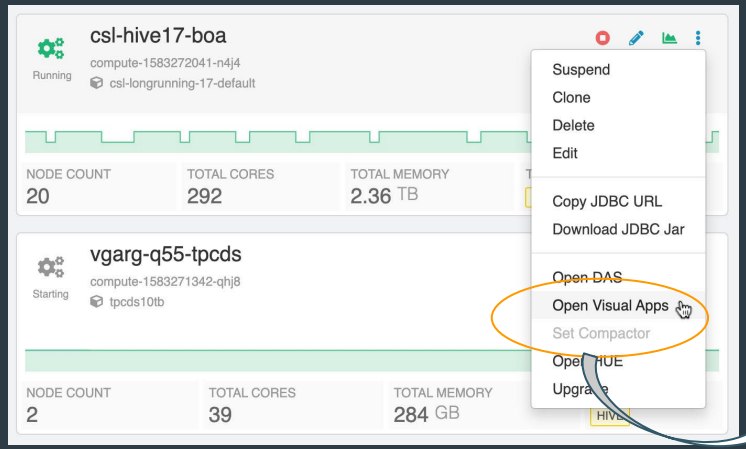


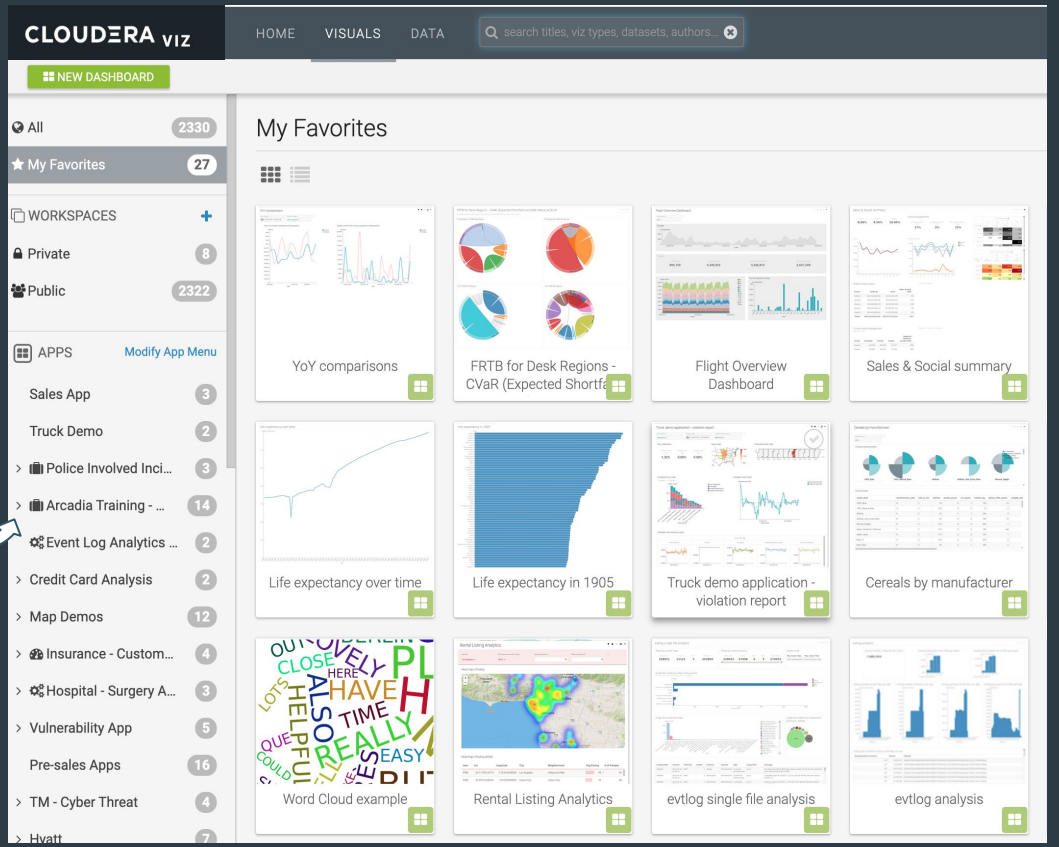
DW Template

- 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





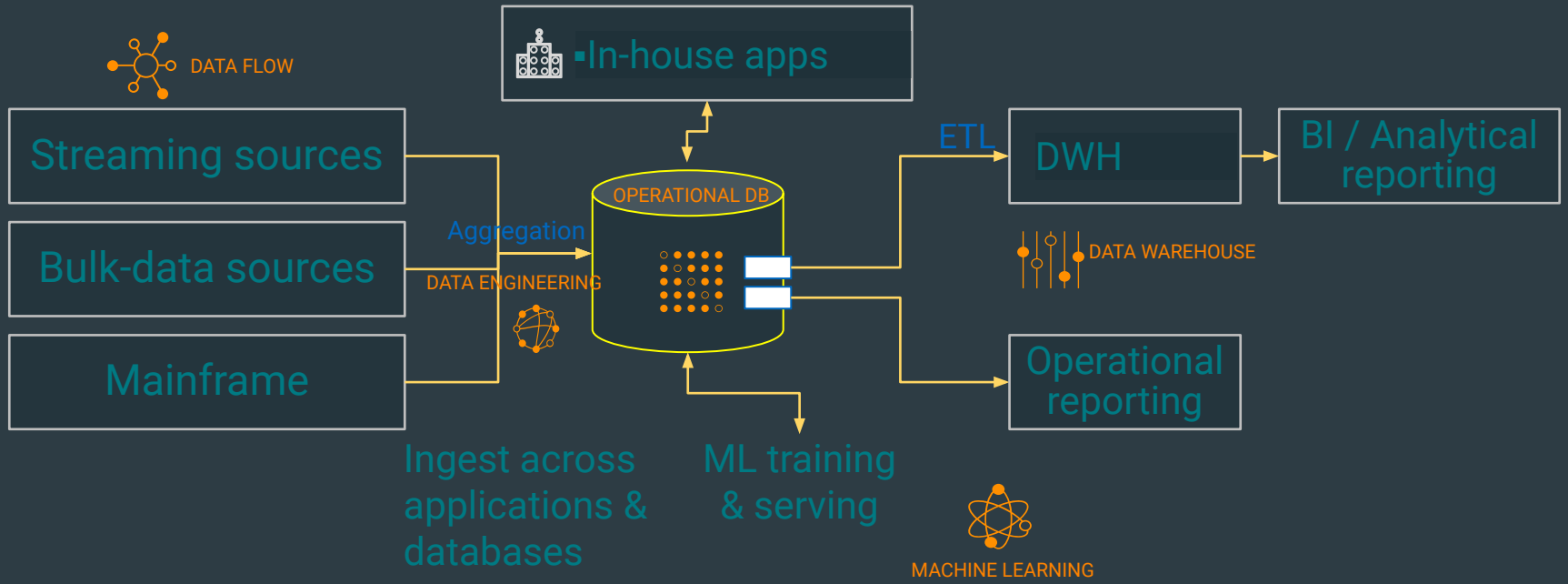
CONNECTING THE DATA LIFECYCLE

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

CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

HOW DOES OPDB FIT IN YOUR ENVIRONMENT



IMPROVES OPERATIONAL AGILITY

Auto-configuration

- Initial config (e.g., kerberos, cache)
- Resiliency (e.g., replication)

Eliminate configuration

- Auto-setup of kerberos, caching, etc
- HA (3 AZs, instance placement groups)
- Replication manager enables replication across regions, clouds

Auto-scaling

- Performance optimization for peak needs vs avg needs

Eliminate sizing

- Automatically scales up based on application needs
- Automatically scales down during periods of low workloads

Auto-tuning

- Hot spotting
- Space management

Eliminate tuning

- Detects hotspotting and alleviates it
- Eliminates need for region management and rebalancing as data grows

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

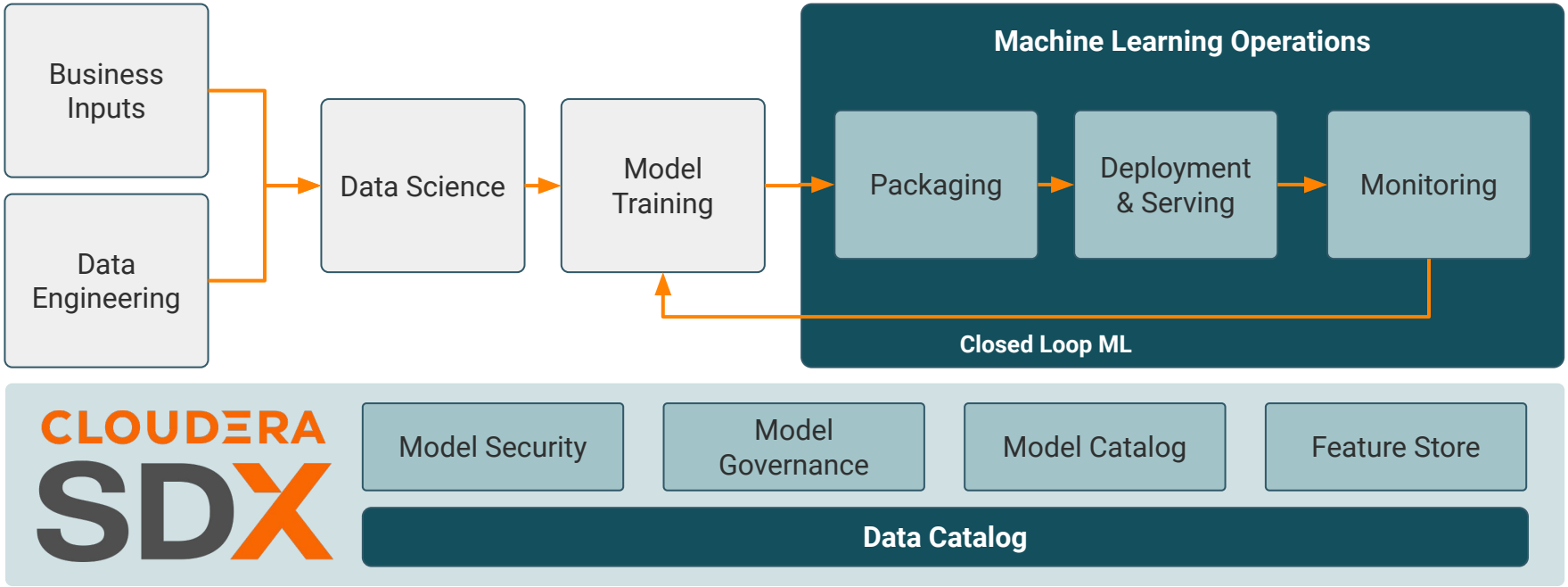
Eliminates operational management

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

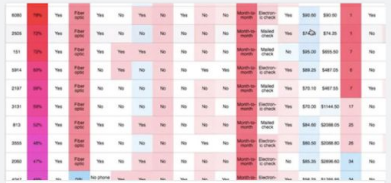


Applied ML Prototypes Catalog

Projects
Sessions
Experiments
Models
Jobs
Applications
Settings
Prototype Catalog
Admin

Search prototypes

Tags: all



Churn Modeling with XGBoost

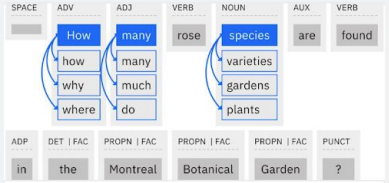
EXPLAINABILITY, XGBOOST

Airline Delay Prediction

XGBOOST

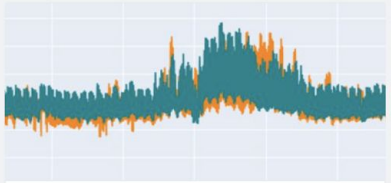
Deep Learning for Image Analysis

COMPUTER VISION IMAGES



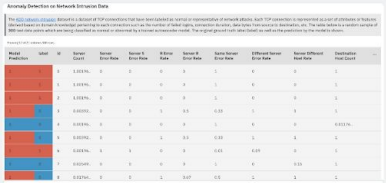
Neural Question Answering

NLP



Structural Time Series

TIME SERIES

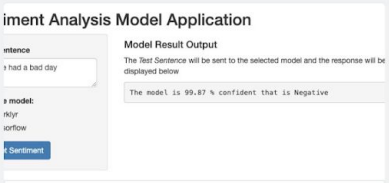


Deep Learning for Anomaly Detection

ANOMALY DETECTION DEEP NEURAL NETS

Fraud Detection

FRAUD DETECTION ANOMALY DETECTION



Sentiment Analysis with Sparklyr and Tensorflow

SENTIMENT ANALYSIS SPARK

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

SECURITY

CHALLENGES:

Security & Governance

Sharing data across workloads

- Requires multiple copies of data need to be created
- Each with its own set of data context

Burdensome admin effort

- Multiple clusters = multiple places to administer

One missing permission in one copy of the data can lead to significant financial and reputation risk

Difficult to share data safely for new analyses

Heavy new regulation such as GDPR makes the challenges even greater

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)

PERIMETER CONTROL : KNOX

Identity & Perimeter

Validate users in enterprise directory

Technical Concepts:
Authentication
User/group mapping

Kerberos,
Apache Knox

Provide a consistent user experience, on all infrastructures and leverage existing directory structure and protocols

- **Minimize** host and port exposure
- Strong **authentication** between users and services automated with Kerberos
- **Centralized** single sign on across all interfaces, which connects to your Active Directory on prem or in the cloud.
- **Complement** Kerberos - hides away complexities of Kerberos from end user applications

ACCESS CONTROL : RANGER & ATLAS

Access

Defining what users and applications can do with data

Technical Concepts:
Permissions
Authorization

Apache Ranger,
Apache Atlas

Maintain one set of data, control access centrally with fine grained policies down to the column and the row level.

- **Anonymize** PII with Dynamic column masking
- **Customize** views for users with Dynamic row filtering
- **Manage** user access with Role-based Access Control
- **Unify** policies across many data sets with Attribute-based Access Control

VISIBILITY CONTROL: ATLAS & RANGER

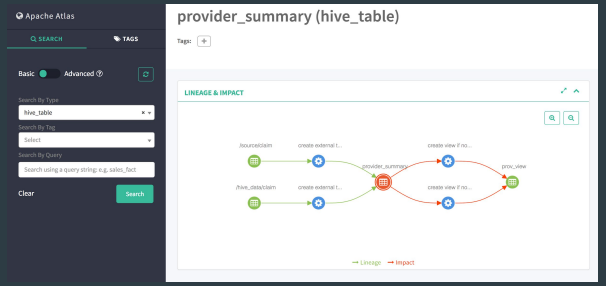
Visibility

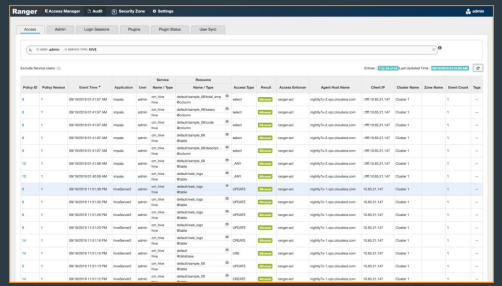
Reporting on where data came from and how it's being used

Technical Concepts:
Auditing
Lineage

Apache Atlas,
Apache Ranger

- Lineage
  - What data do I consume?
  - What consumes my Data?
  - Who uses my data?
  - What data was used to train my model that resulted in X prediction?
-
- Audit who accessed what
  - Track access events from Apache Ranger
  - Metadata audit and versioning from Apache Atlas





DATA PROTECTION: ENCRYPTION

Data Protection

Shielding data in the cluster from unauthorized visibility

Technical Concepts:
Encryption
Key Management

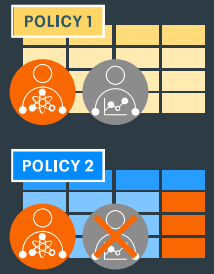
SSL/TLS
HDFS TDE

- Encryption of Data at Rest: Transparent Data Encryption (TDE)
  - Selective encryption of relevant files/folders
  - Prevent rogue admin access to sensitive data
  - Fine grained access controls
  - Transparent to client applications. (no changes required)
  - Enterprise-grade keystore with KTS
  - Easy key-management with Ranger
  - Support HSM integration
-
- Encryption of Data in Motion: SSL/TLS
  - Easily enable and manage encryption over-the-wire with AutoTLS

GOVERNANCE

Governance Can Be Complex

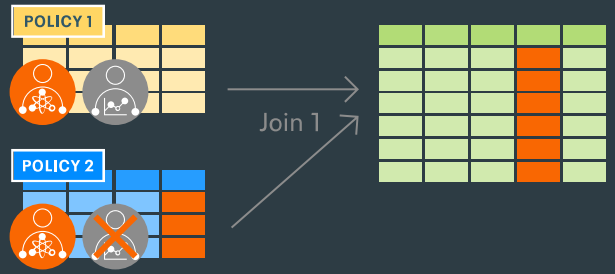
It’s an operational tax that adds complexity and slows everything down



Governance Can Be Complex

It's an operational tax that adds complexity and slows everything down

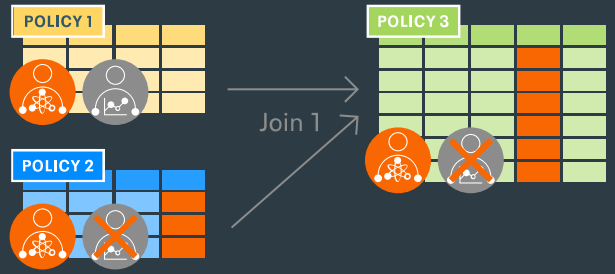
- Sensitive data can propagate!



Governance Can Be Complex

It's an operational tax that adds complexity and slows everything down

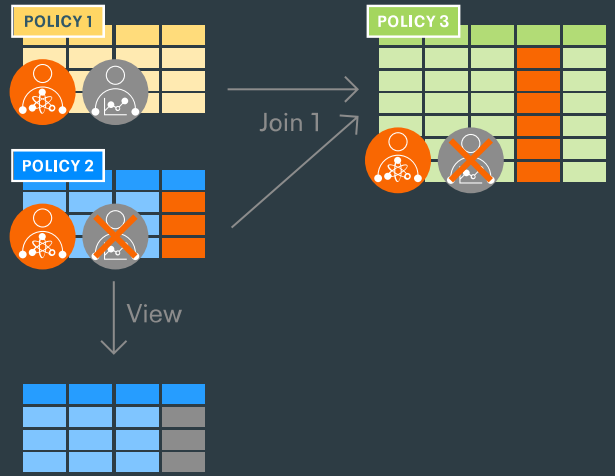
- Sensitive data can propagate!
- Admins need to set policies on every newly created table or view



Governance Can Be Complex

It's an operational tax that adds complexity and slows everything down

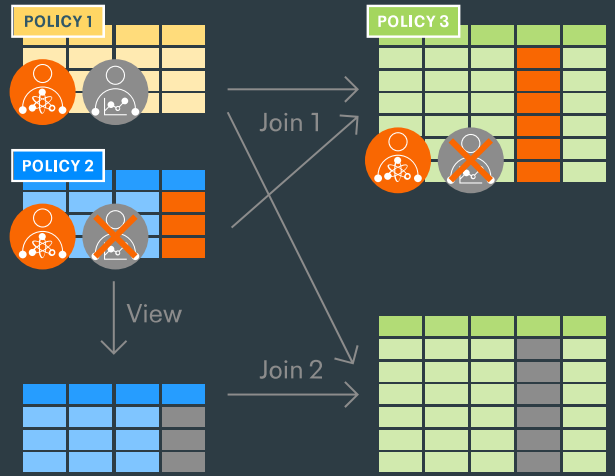
- Sensitive data can propagate!
- Admins need to set policies on every newly created table or view
- This can be managed with views



Governance Can Be Complex

It's an operational tax that adds complexity and slows everything down

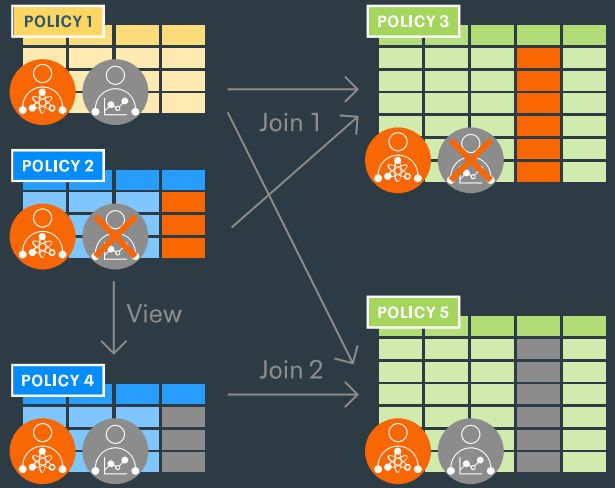
- Sensitive data can propagate!
- Admins need to set policies on every newly created table or view
- This can be managed with views



Governance Can Be Complex

It's an operational tax that adds complexity and slows everything down

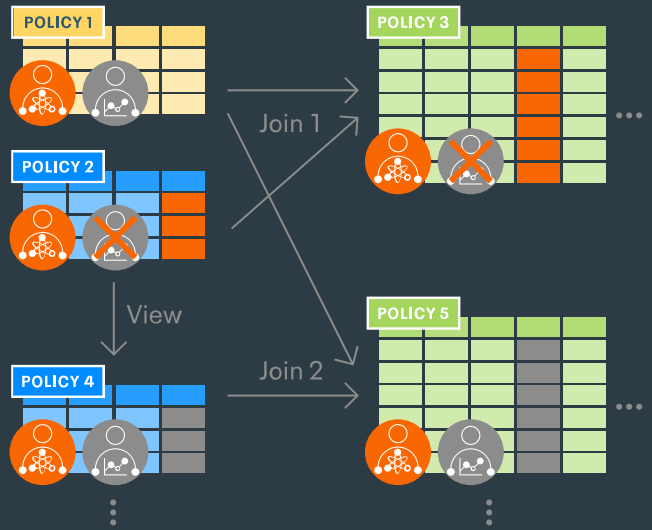
- Sensitive data can propagate!
- Admins need to set policies on every newly created table or view
- This can be managed with views
- Admins need to set perms on every new flavor of table or view created



Governance Can Be Complex

It's an operational tax that adds complexity and slows everything down

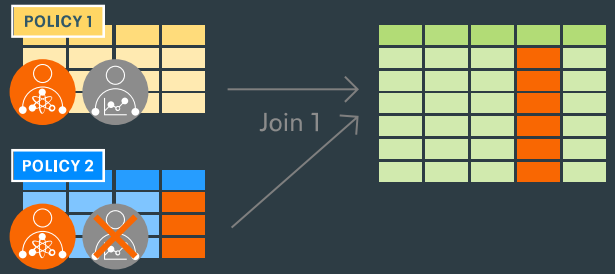
- Sensitive data can propagate!
- Admins need to set policies on every newly created table or view
- This can be managed with views
- Admins need to set perms on every new flavor of table or view created
- Each interaction takes time and the number of policies can scale exponentially!



Governance in Cloudera SDX is Simplified

Elegant Auto-propagating Attribute-Based Access Controls and Masking Policies

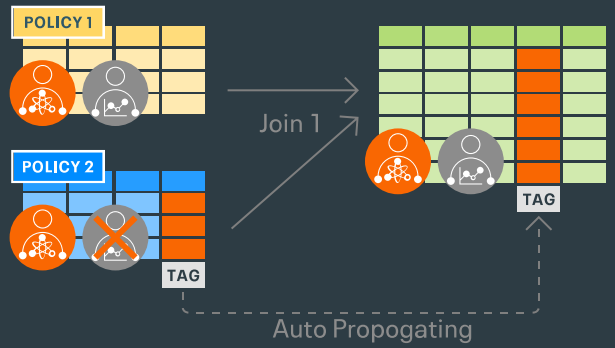
- Sensitive data can propagate!



Governance in Cloudera SDX is Simplified

Elegant Auto-propagating Attribute-Based Access Controls and Masking Policies

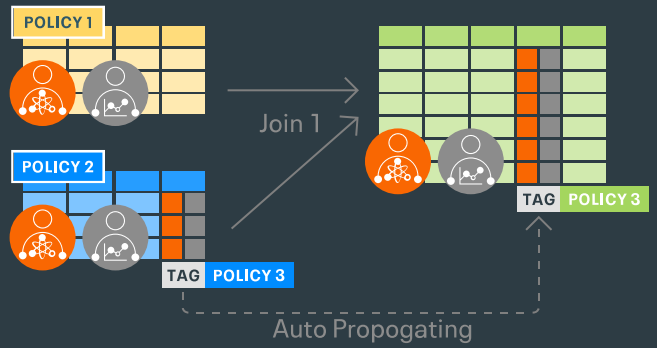
- Sensitive data can propagate!
- Admins apply **Tags** to columns and that propagate through lineage.



Governance in Cloudera SDX is Simplified

Elegant Auto-propagating Attribute-Based Access Controls and Masking Policies

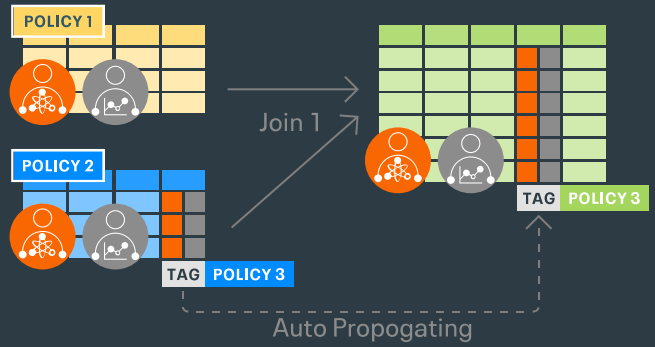
- Sensitive data can propagate!
- Admins apply **Tags** to columns and that propagate through lineage.
- Admins apply **Policies** to tags that propagate their effects through lineage.



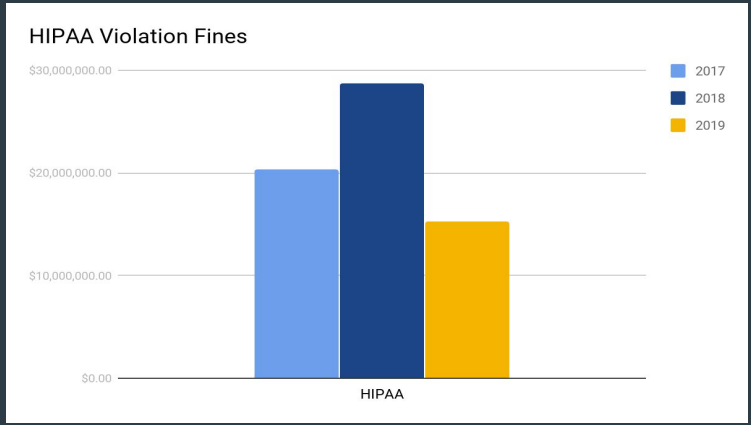
Governance in Cloudera SDX is Simplified

Elegant Auto-propagating Attribute-Based Access Controls and Masking Policies

- Sensitive data can propagate!
- Admins apply **Tags** to columns and that propagate through lineage.
- Admins apply **Policies** to tags that propagate their effects through lineage.
- Just define one tag and one policy and it gets automatically applied to all derived tables!



Data Governance is mandatory



Source : HIPAA Fines Listed by Year

Violation	Sum of Fines
Insufficient technical and organisational measures to ensure information security	**€ 332,967,397** (at 63 fines)
Insufficient legal basis for data processing	**€ 110,989,368** (at 104 fines)
Non-compliance with general data processing principles	**€ 16,070,665** (at 40 fines)
Insufficient fulfilment of data subjects rights	**€ 7,864,397** (at 25 fines)
Insufficient fulfilment of information obligations	**€ 557,265** (at 15 fines)
Insufficient fulfilment of data breach notification obligations	**€ 177,125** (at 7 fines)
Lack of appointment of data protection officer	**€ 111,000** (at 3 fines)
Insufficient cooperation with supervisory authority	**€ 55,511** (at 9 fines)
Insufficient data processing agreement	**€ 14,380** (at 2 fines)
Insufficient cooperation with supervisory authority	**€ 4,400** (at 1 fines)
Unknown	**€ 500** (at 1 fines)
Insufficient fulfilment of data breach obligations	**€ 286** (at 1 fines)

Source : GDPR Enforcement Tracker 2018 -2020

UNDERSTAND YOUR DATA

Search

User-driven Text based search

Technical Concepts:
Faceted search
Business Glossary

Cloudera **Data Catalog**
Apache **Atlas**

Navigate

Browse to find related data sets.

Technical Concepts:
Navigation links
Breadcrumbs

Cloudera **Data Catalog**
Apache **Atlas**

Discover

Automated insights about your data

Technical Concepts:
Profilers

Cloudera **Data Catalog**

Crowdsource

Use wisdom of the crowds to better understand data

Technical Concepts:
Curation
Collections
User ratings

Cloudera **Data Catalog**

Focus on your data and your people

UNDERSTAND YOUR DATA : SEARCH

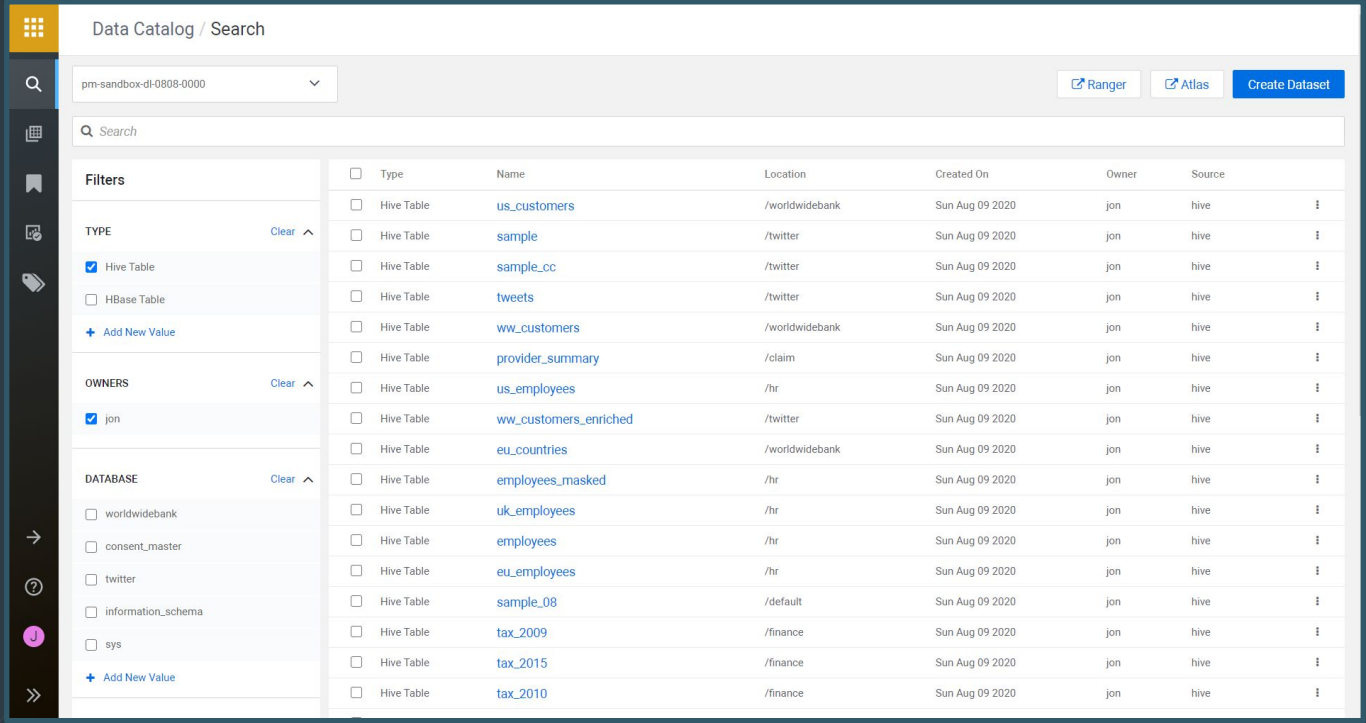
via Cloudera Data Catalog and Apache Atlas

Search

User-driven Text based search

Technical Concepts:
Faceted search
Business Glossary

Cloudera Data Catalog
Apache Atlas



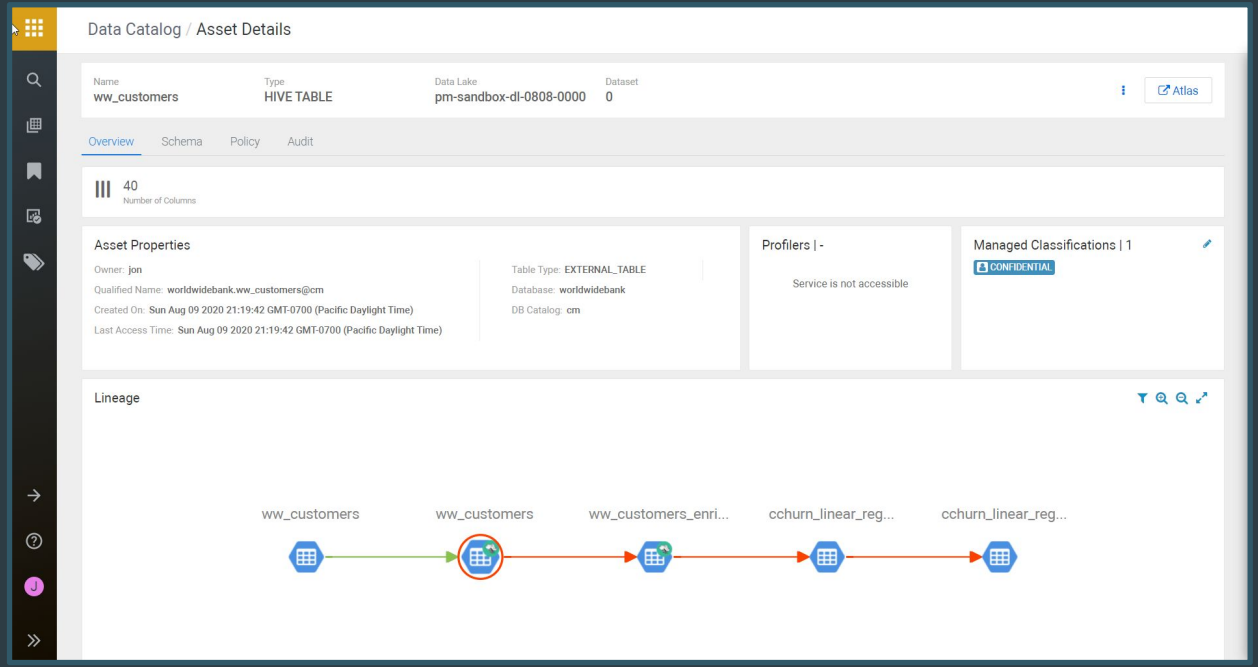
UNDERSTAND YOUR DATA : NAVIGATE

Navigate

Browse to find related data sets.

Technical Concepts:
Navigation links
Breadcrumbs

Cloudera Data Catalog
Apache Atlas



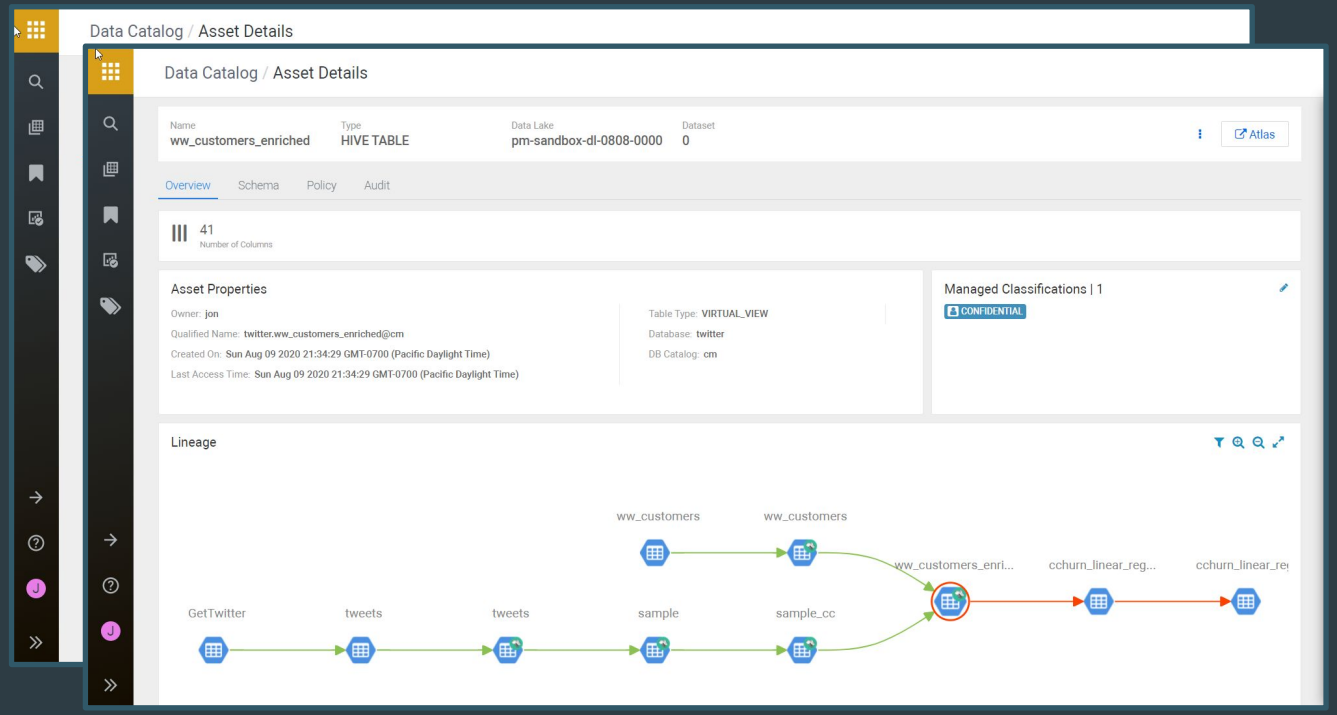
UNDERSTAND YOUR DATA : NAVIGATE

Navigate

Browse to find related data sets.

Technical Concepts:
Navigation links
Breadcrumbs

Cloudera Data Catalog
Apache Atlas



UNDERSTAND YOUR DATA : DISCOVERY

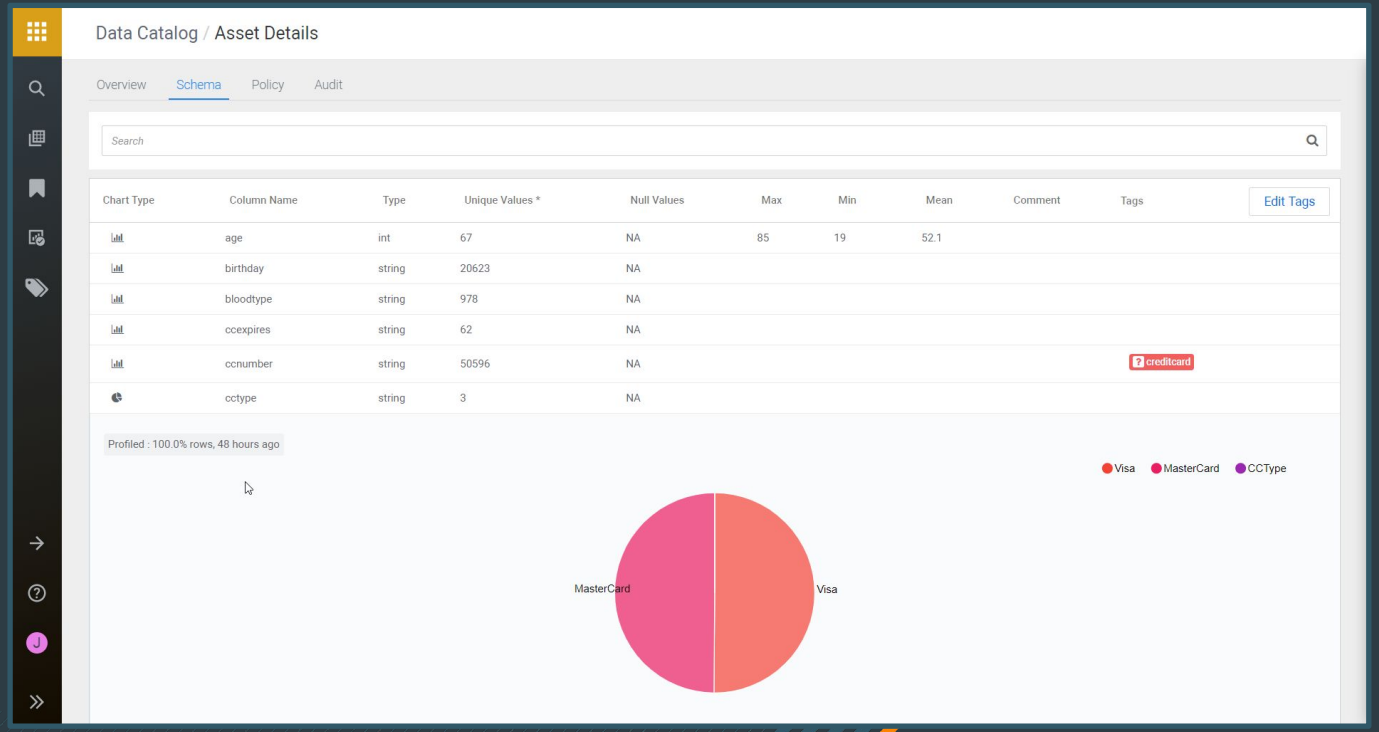
via Cloudera Data Catalog's Profilers

Discover

Automated insights about your data

Technical Concepts:
Profilers

Cloudera Data Catalog



UNDERSTAND YOUR DATA : CROWDSOURCE

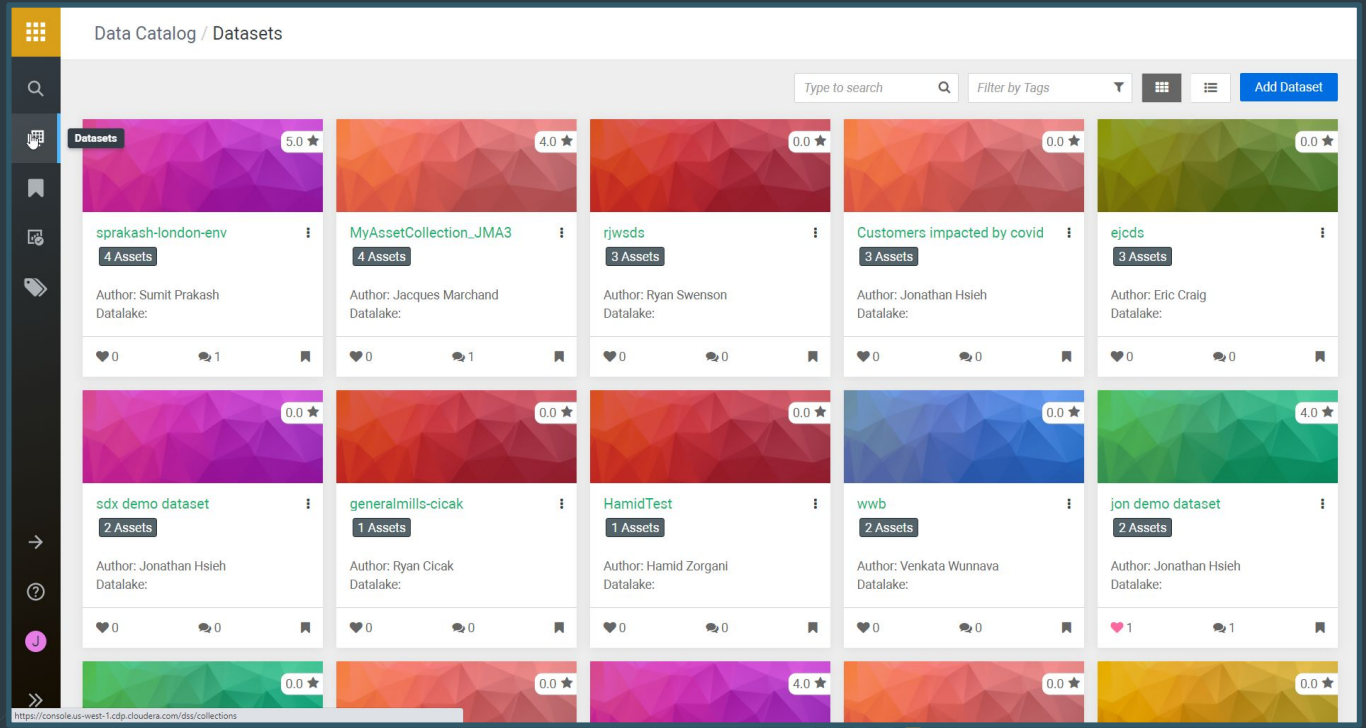
via Data Catalog Asset Collections

Crowdsource

Use wisdom of the crowds to better understand data

Technical Concepts:
Curation
Collections
User ratings

Cloudera Data Catalog



WHAT DOES ATLAS DO?

METADATA

- A **catalog** for metadata of assets in an enterprise
  - Hive (HiveServer2 and HMS) , Impala, Spark, HBase, Sqoop, NiFi
- Dynamically create asset types with complex attributes and relationships
- Extensible **type system** with
  - Custom Business Metadata
  - Labels, properties, etc

LINEAGE

- Answers questions of
  - Where did data come from?
  - Where has it been used?
- Is automatically maintained for most asset types
- Lineage graphs are basis for Classification propagation

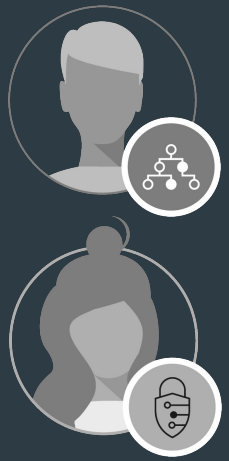
CLASSIFICATION

- Classifications (**Tags**) with custom attributes
  - Automatic propagation through lineage graph
  - Use with Classification-based access control (**ABAC**)
- Describe assets with Business **Glossary Terms**
  - With relationships between business terms
  - And associated classifications

CDP DATA CATALOG TARGET PERSONAS

Data Architects

- Data Architects
- Compliance Managers



Concerns: **Control**

- Compliance / Governance
- Data Lifecycle / Curation
- Multi-team integration

Data Users

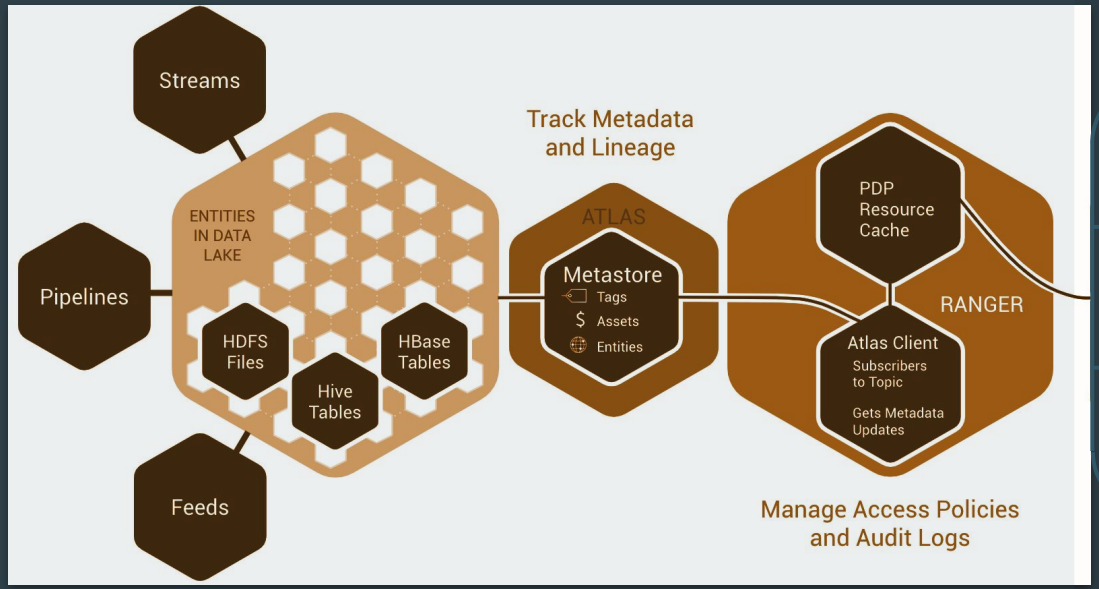
- Data Analysts
- Data Engineers
- Data Scientists
- App Developers

Concerns: **Understand**

- Domain Expertise
- Insight about Data
- Data Quality

CDP — GOVERNANCE & SECURITY

Attribute-Based Access Controls and more with Atlas & Ranger

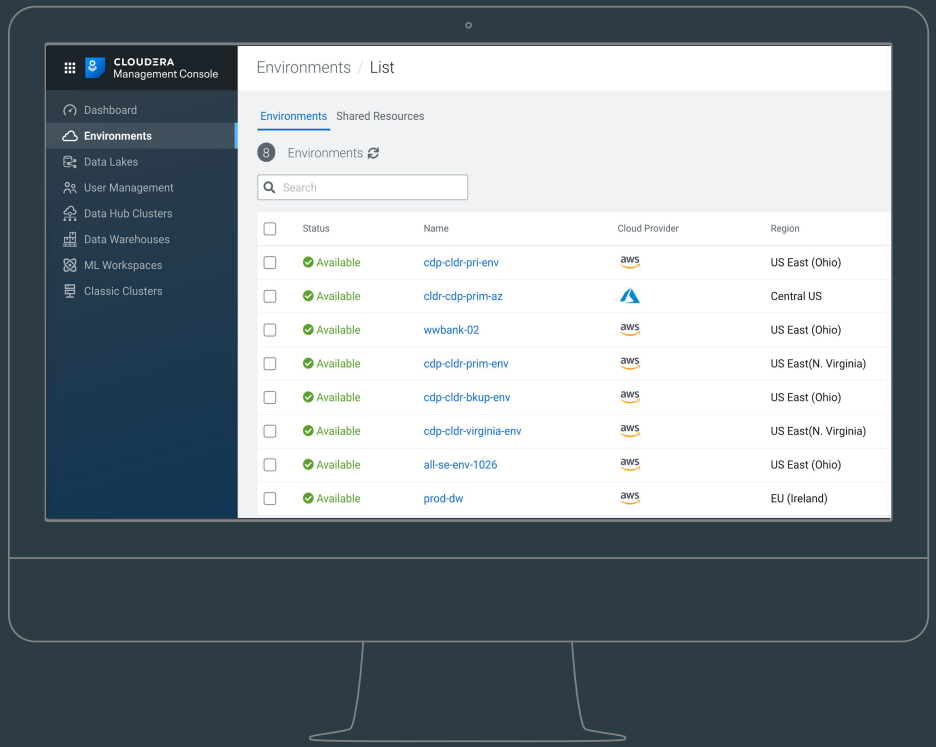


Policies

- Classification Based Access & Masking
- Prohibition
- Time Based, Schedulable Policies
- Location

Cloudera Data Platform Demo

KEY CONCEPTS & COMPONENTS



ENVIRONMENTS

Environment

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

1:1

Data Lake

- SDX: Security, Governance, & Lineage
- Associated with groups/users

1:N

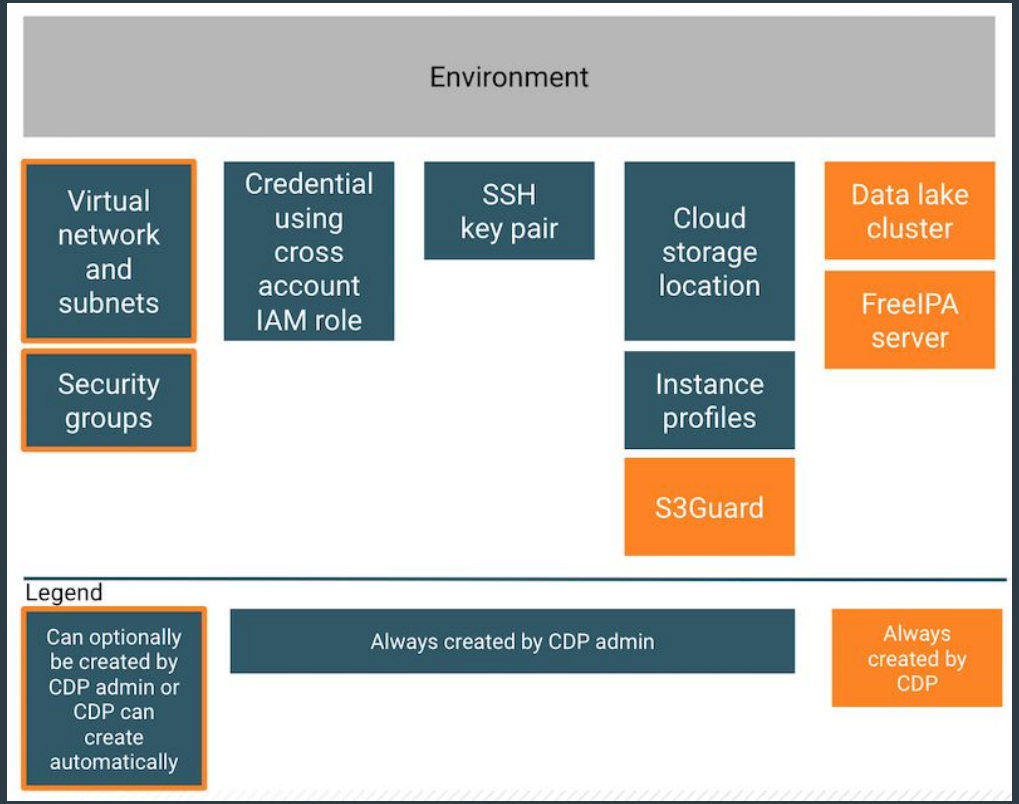
Data Hub Clusters / Experiences

- DH templates
- ML Env
- DW Database Catalogs/Virtual Compute

ENVIRONMENT

What is an environment?

Definition of where CDP creates resources in a customer environment.
A long running permanent cluster called a Data Lake gets created here.



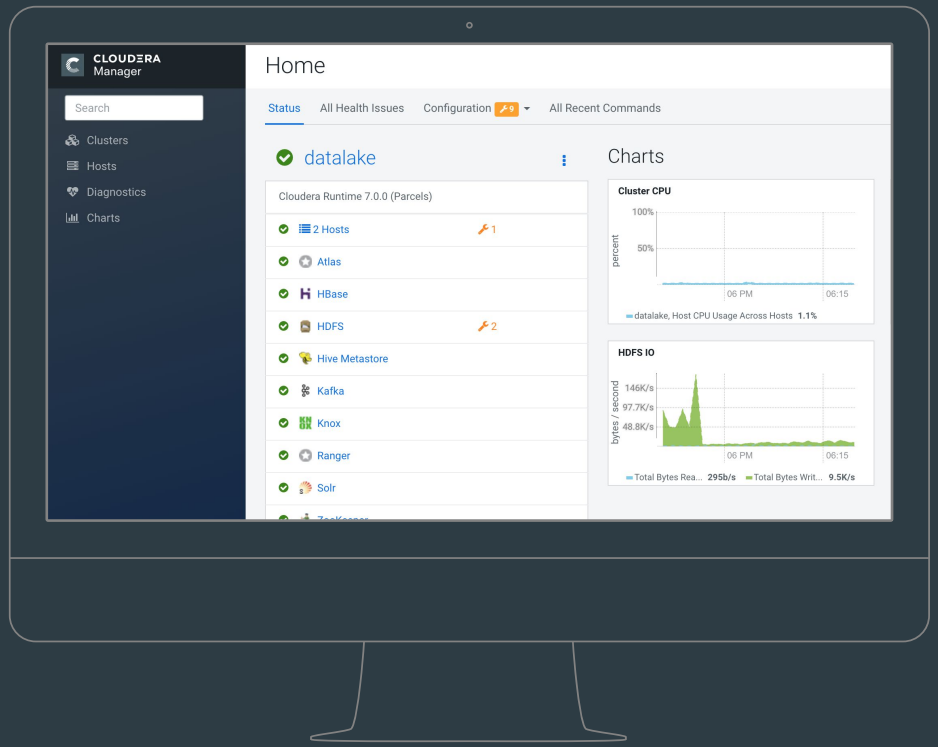
DATA LAKE

What is a Data Lake?

A common set of Services (SDX) within an Environment that are shared across multiple Clusters/Experiences.

These include Services for:

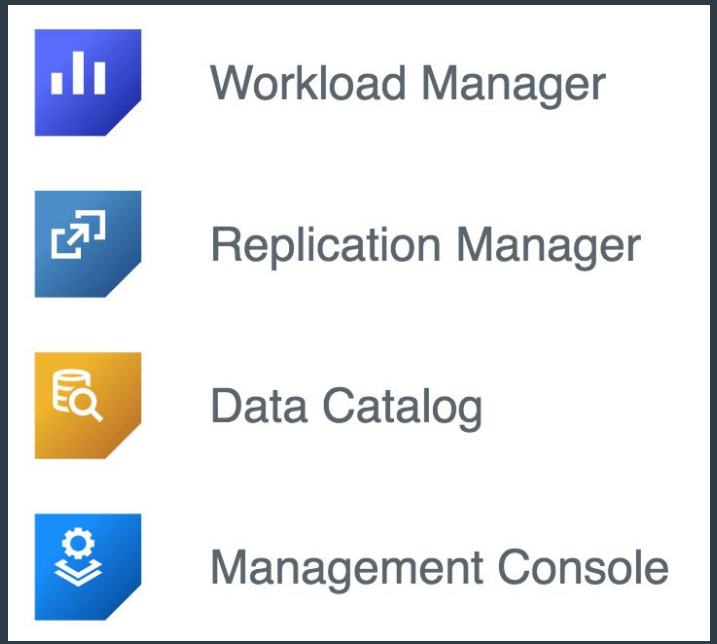
- Security
- Auditing
- Governance
- Data Discovery



CONTROL PLANE

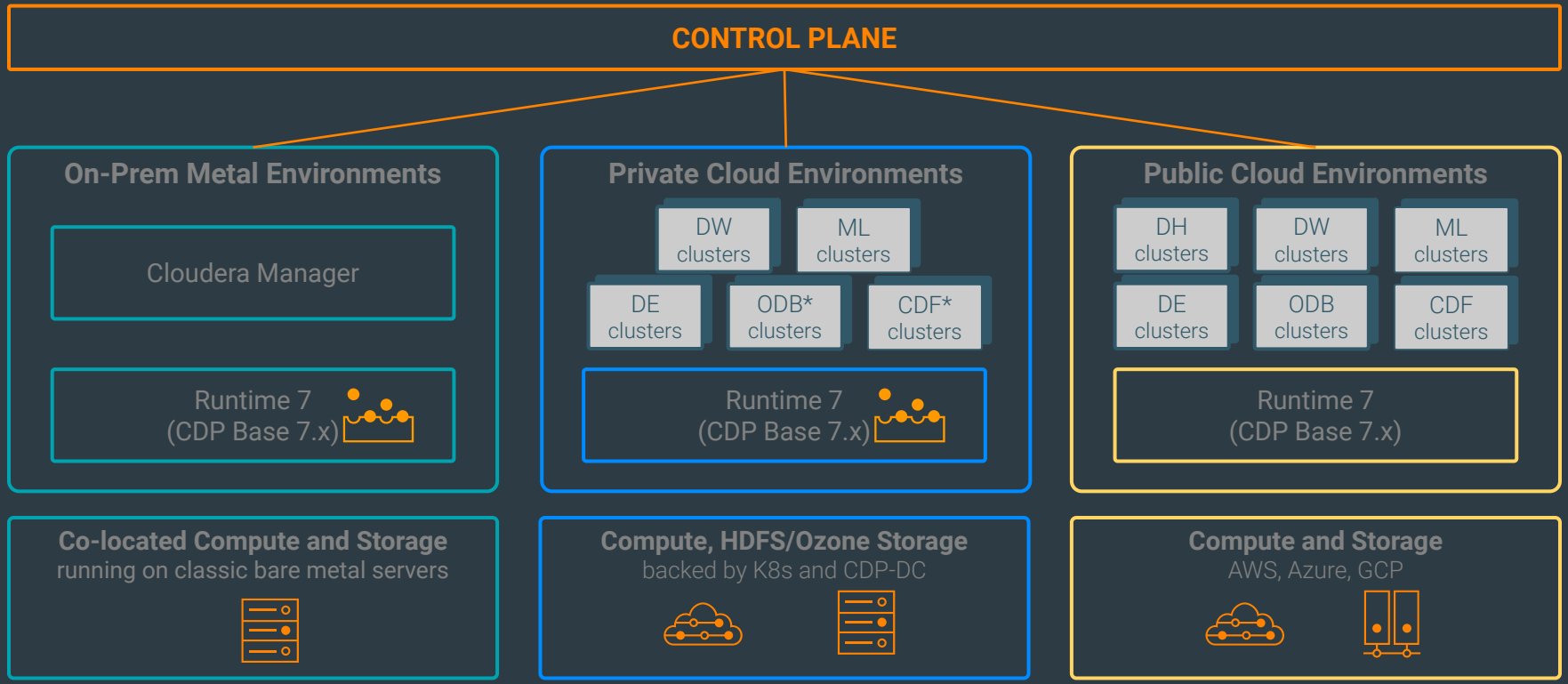
What is the Control Plane?

The Control Plane is the common set of tools for management, workload analysis, data movement and data discovery across multiple environments



More on CDP Technicals

3 FORM FACTORS

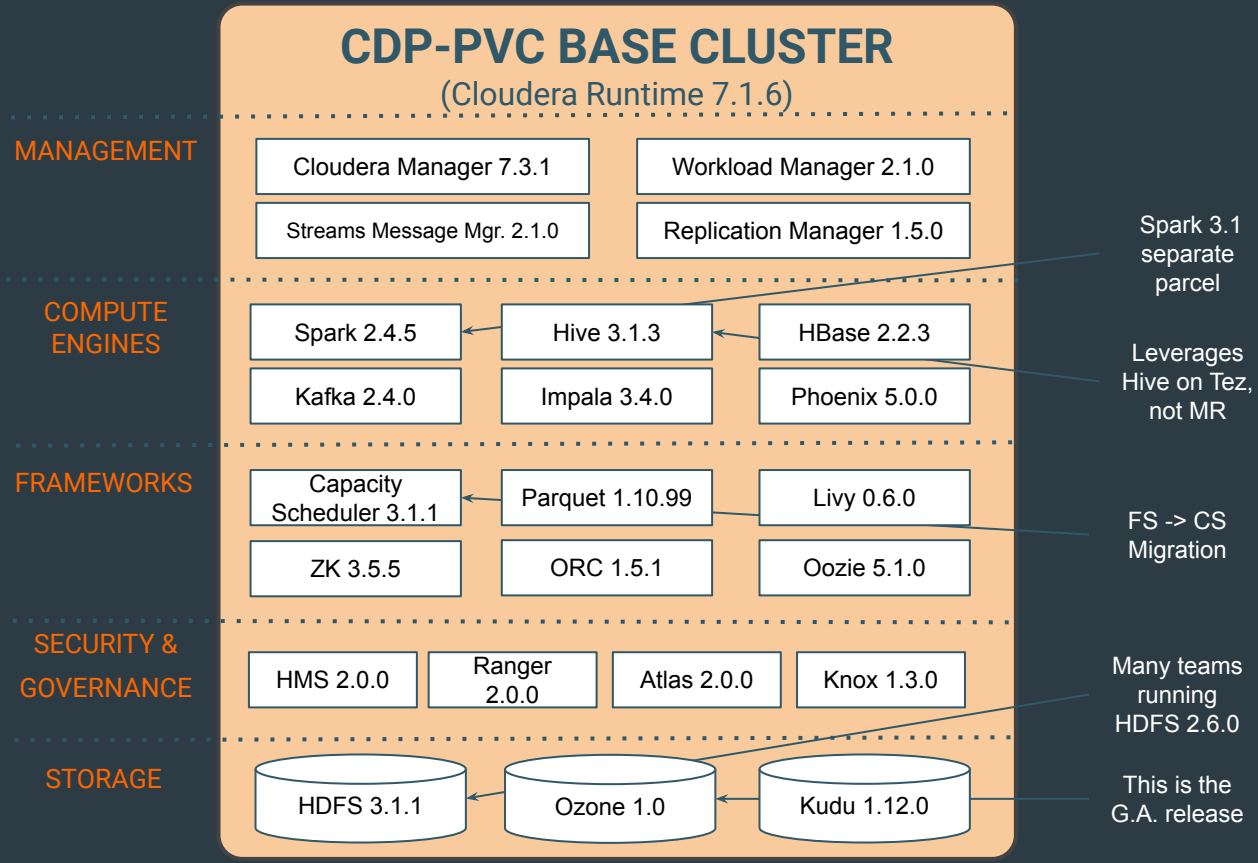


CDP RUNTIME 7.1

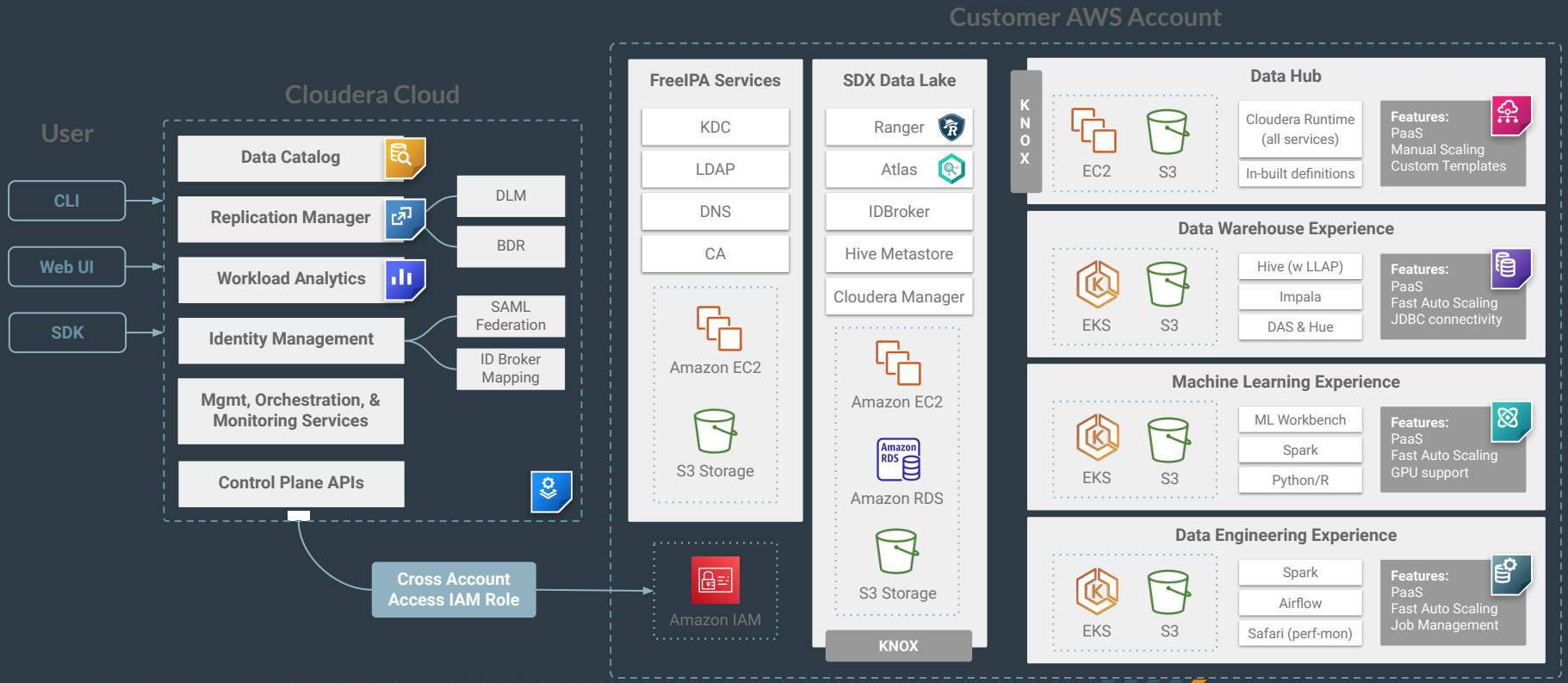
A CDP Private Cloud Base cluster (powered by Cloudera Runtime), can serve as a traditional "data lake" (storage & compute) cluster, or as a "base storage cluster" (storage only) serving compute workloads running on Kubernetes.

This image shows the component versions in Cloudera Runtime 7.1.6.

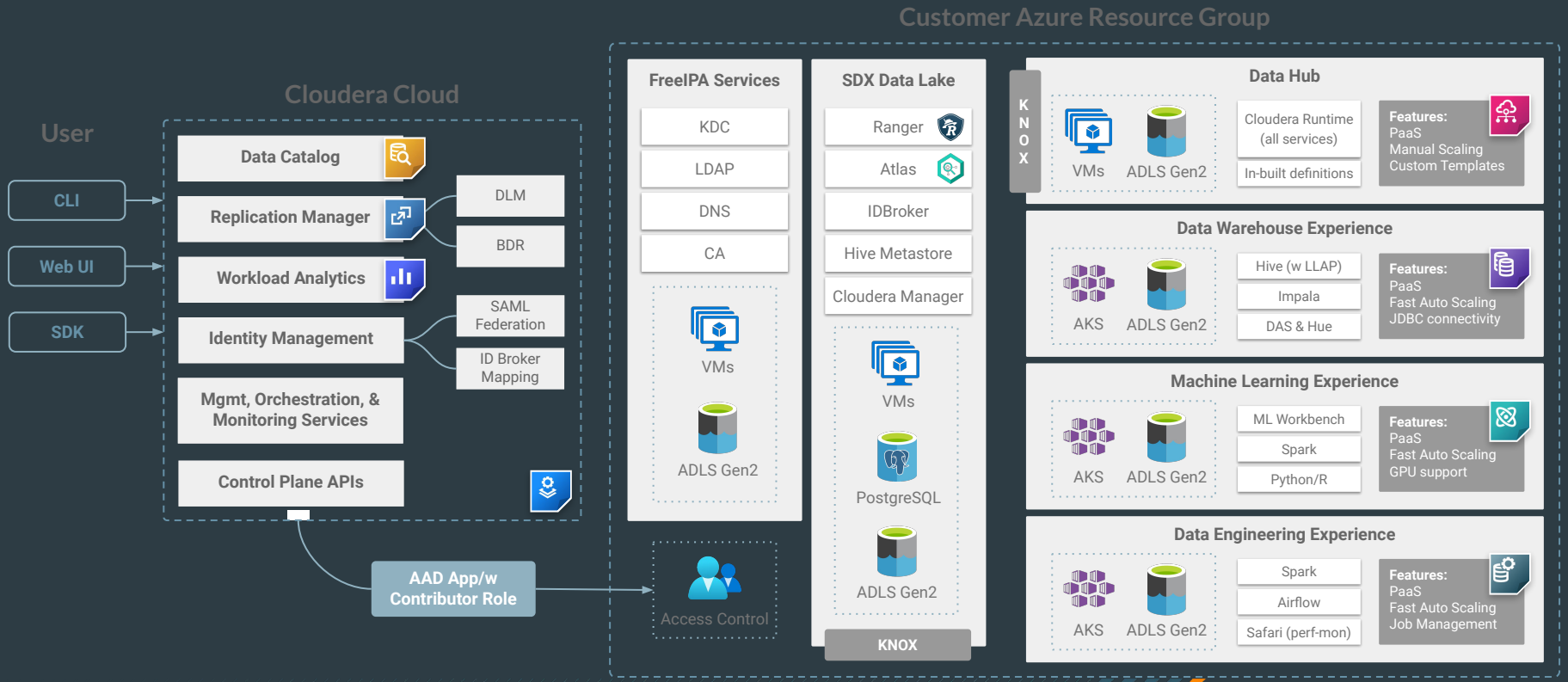
Click HERE for the complete list of supported components.



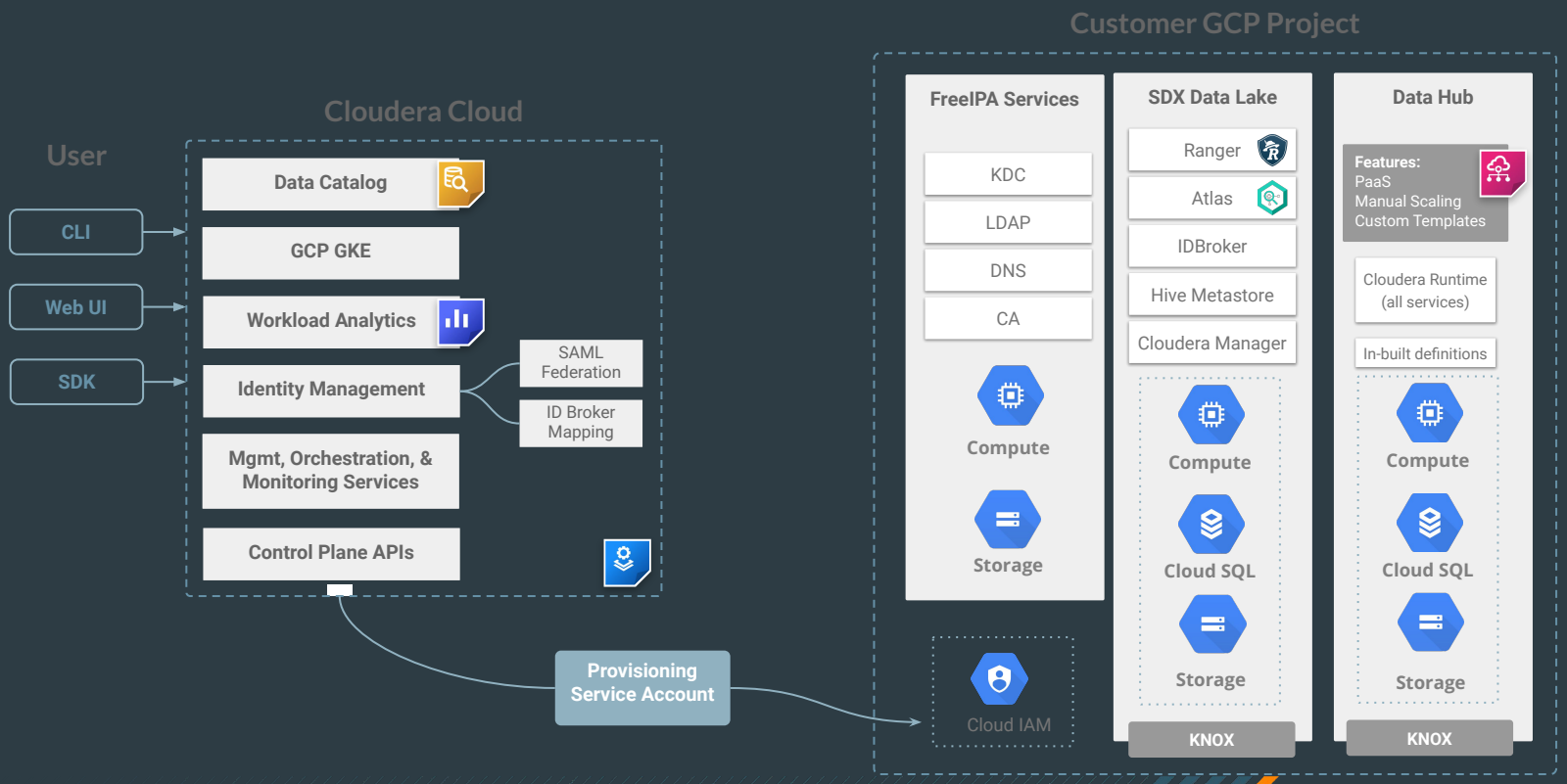
CDP - AWS HIGH LEVEL ARCHITECTURE



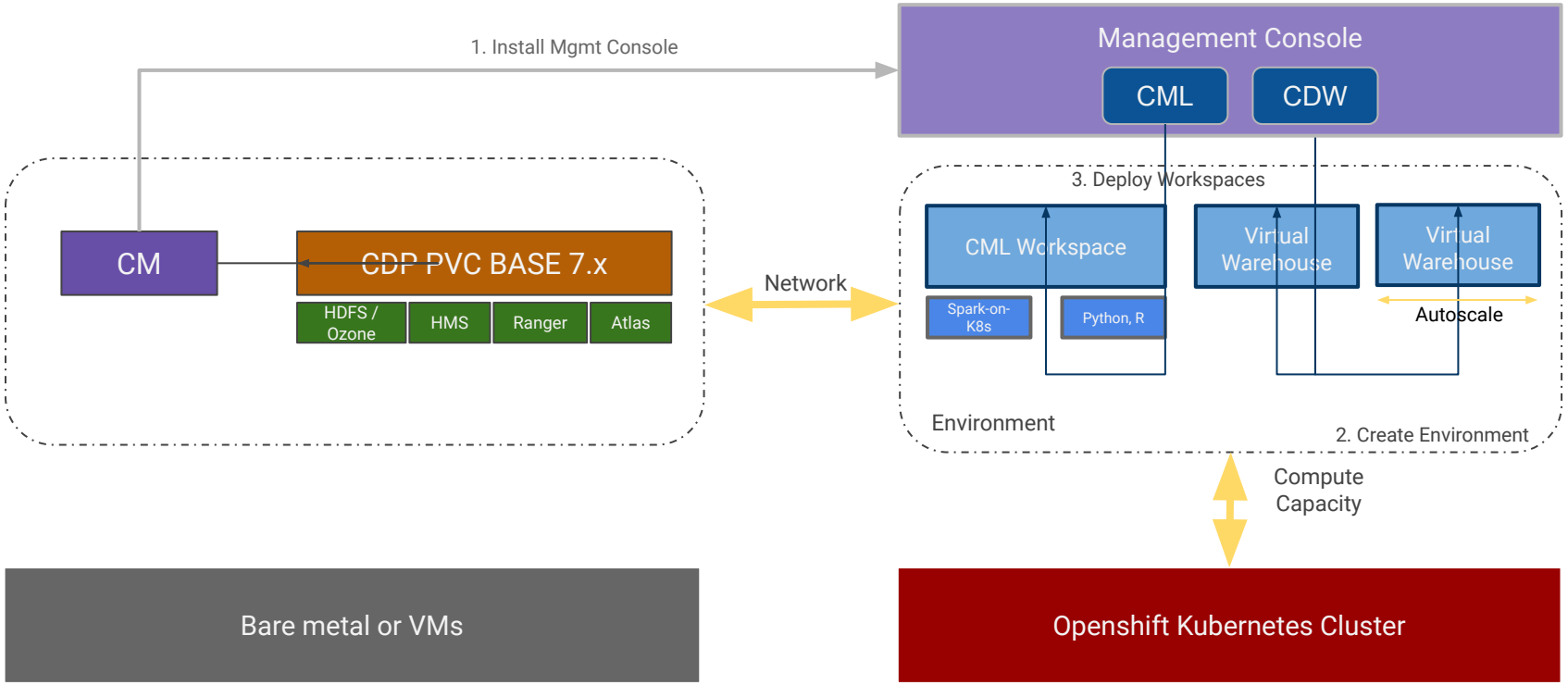
CDP - AZURE HIGH LEVEL ARCHITECTURE



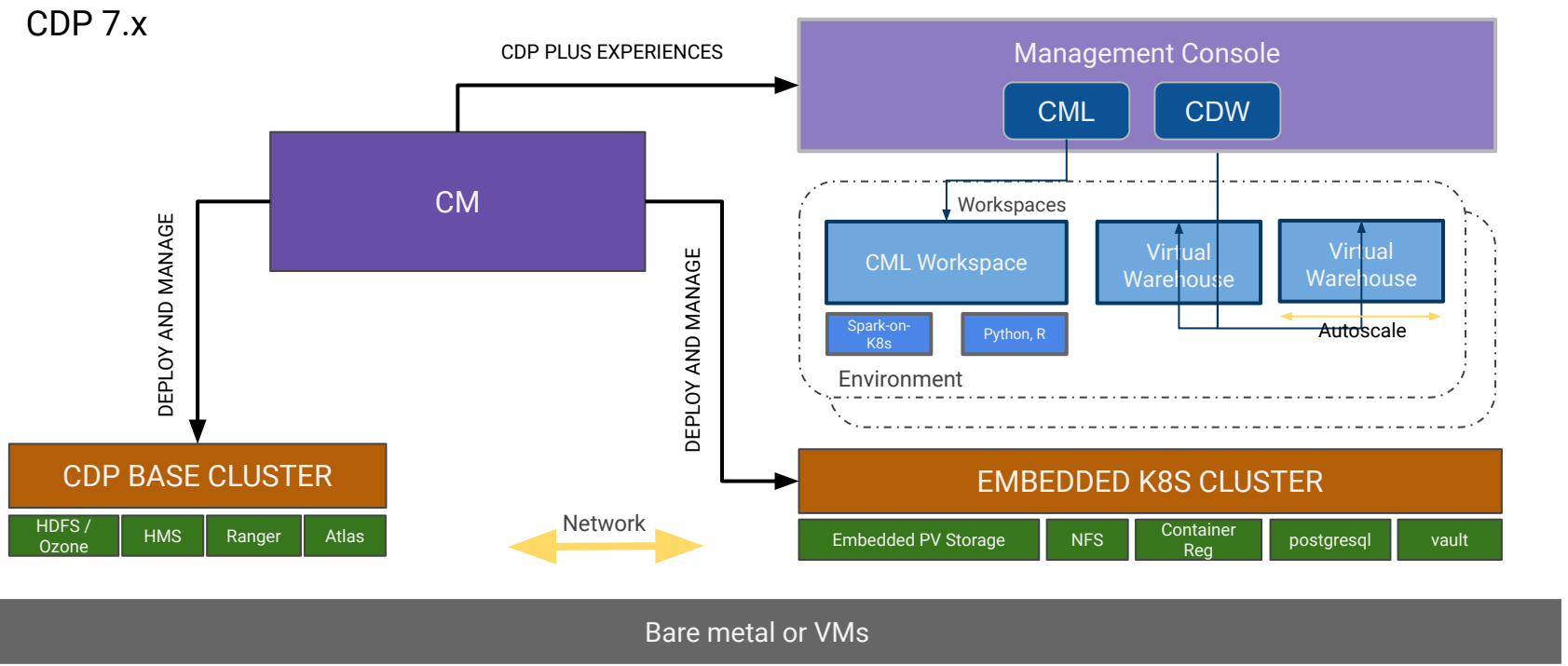
CDP - GCP HIGH LEVEL ARCHITECTURE



CDP Private Cloud – Dedicated Kubernetes

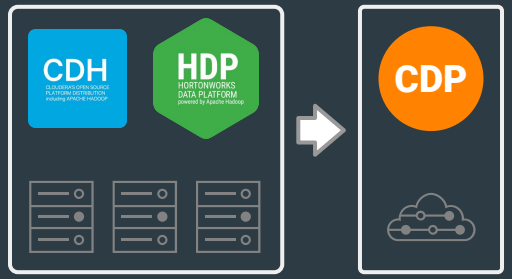


CDP Private Cloud – Embedded Version



THREE PATHS TO CDP

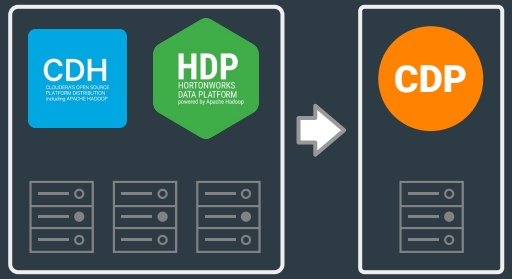
Migrate to Public Cloud



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.

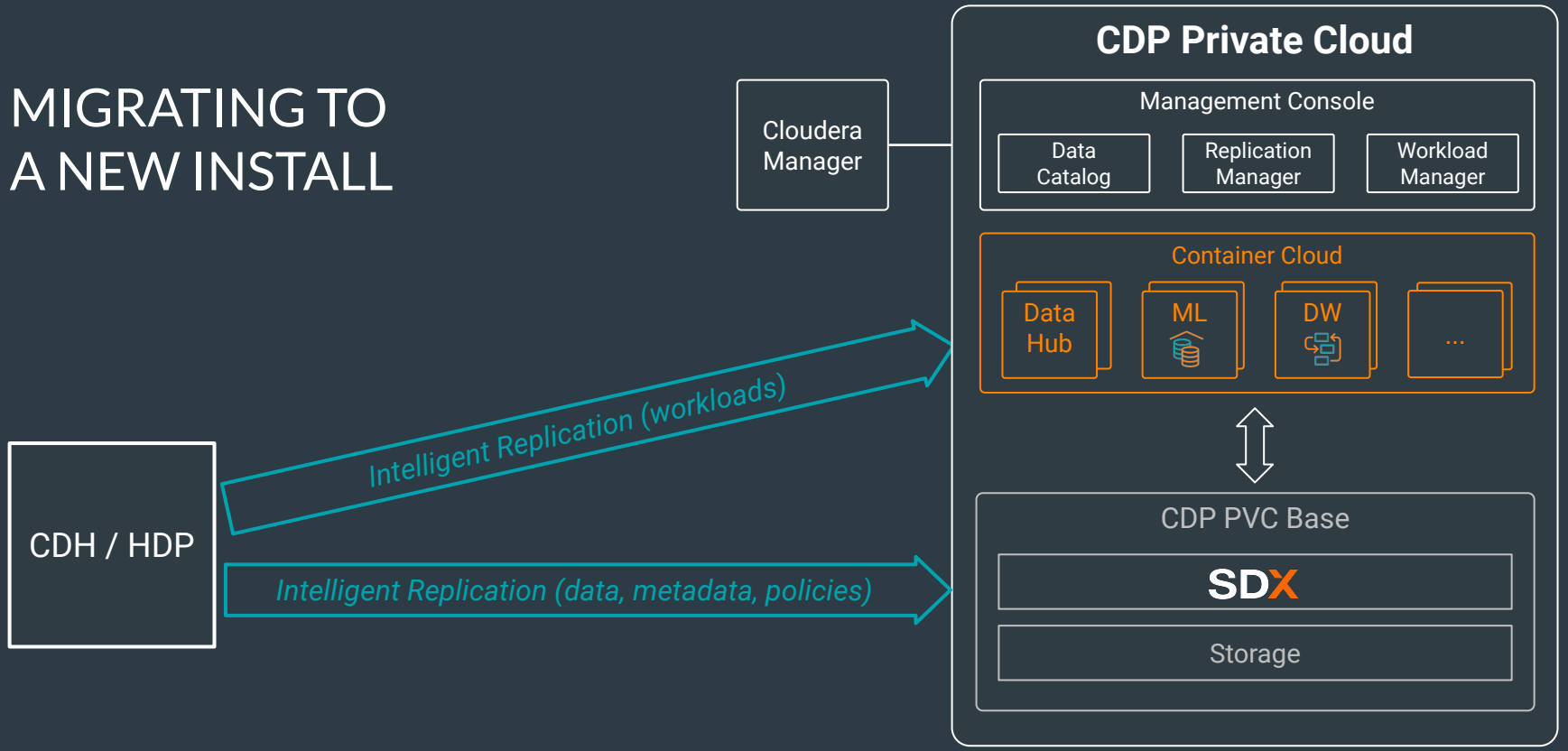
Higher initial investment

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

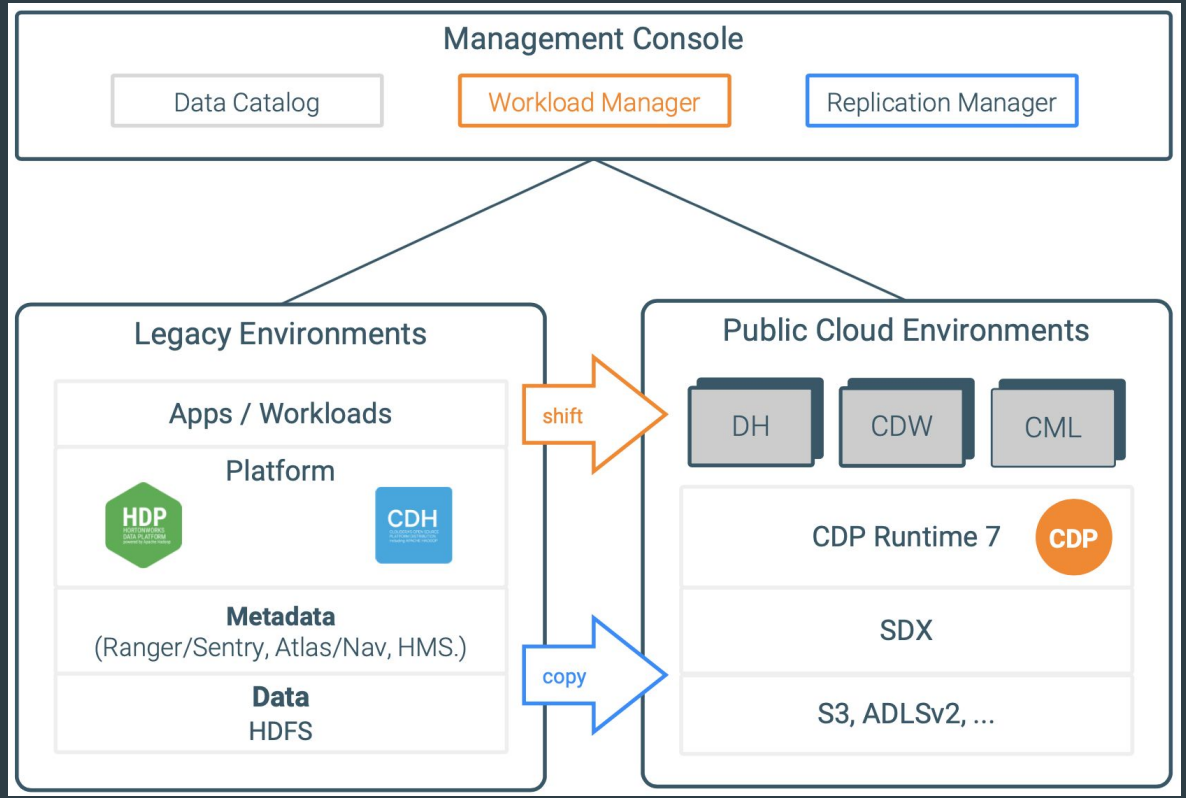
MIGRATING TO A NEW INSTALL



MIGRATE TO PUBLIC CLOUD

Process

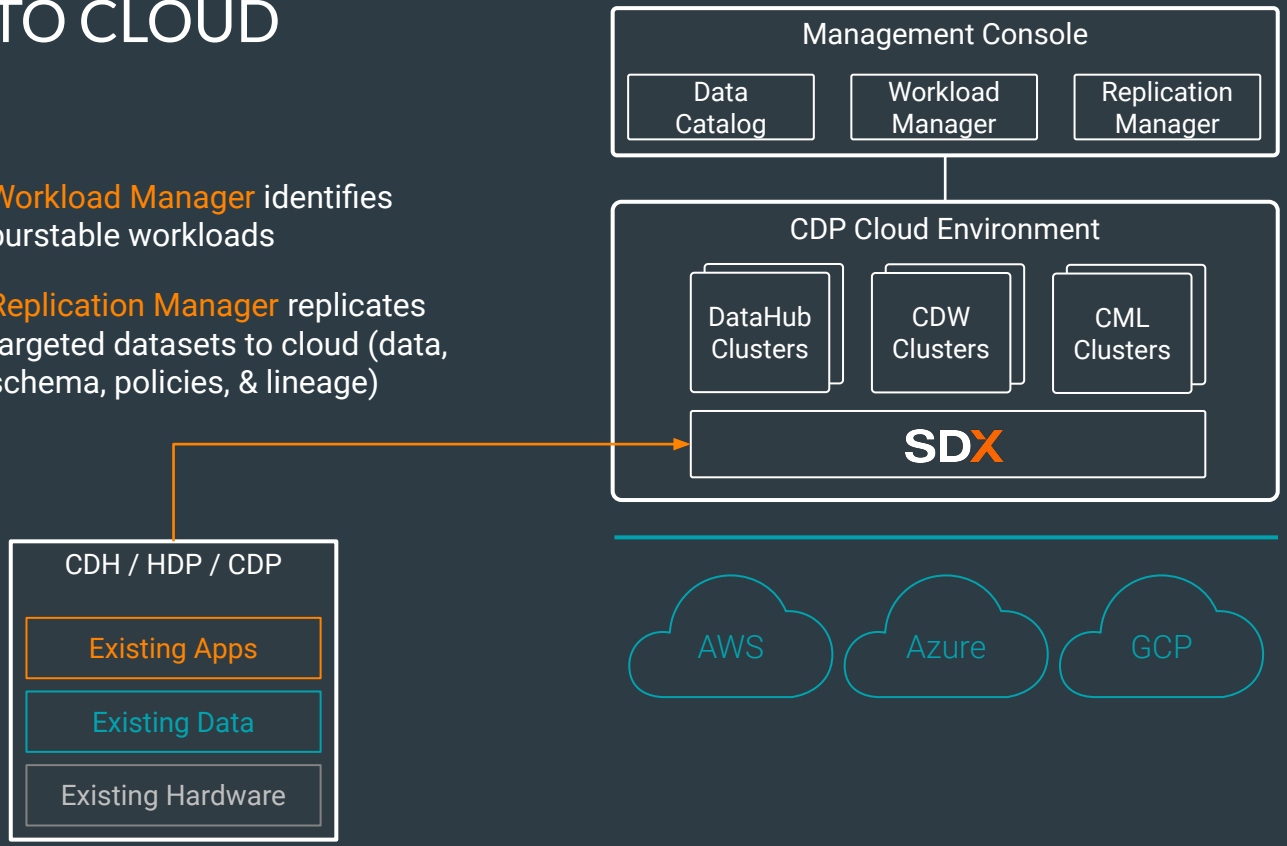
- Set up public cloud environments
- Register classic cluster(s)
- Identify candidate workloads
- Migrate workload data and metadata using Replication Manager (“Burst to Cloud”)
- Test and promote to production



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 Data Services (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 Data Services (New)** — Optimized for modernizing apps on prem to accelerate time to value and improving practitioner experience

CDP DATA SERVICES

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

PVC BASE & DATA HUB		CDP DATA SERVICES
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

Data Engineering

Schedule, monitor, and debug data pipelines to streamline ETL processes quickly and securely.

$0.07/CCU

Hourly rate

Data Warehouse

Deploy data warehouses with secure, self-service access to enterprise data.

$0.07/CCU

Hourly rate

Operational Database

Develop future-proof applications that deliver unparalleled scale, performance, and reliability.

$0.08/CCU

Hourly rate

Machine Learning

Provide collaborative ML workspaces with secure, self-service access to enterprise data.

$0.17/CCU

Hourly rate

Data Hub

Easily manage clusters running Apache Spark, Hive, Impala, HBase, Phoenix, Kafka, Flink, and more.

$0.04/CCU

Hourly rate

DataFlow

Catalog, deploy, manage and monitor Apache NiFi data flow deployments

$0.30 / CCU

Hourly rate

Flow Management on Data Hub

A premium Data Hub service to ingest, transform and manage streaming data, powered by Apache NiFi.

$0.15/CCU

Hourly rate

Data Warehouse Data Service on AWS - Minimum Hardware Requirement

Data Warehouse Data Service (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)

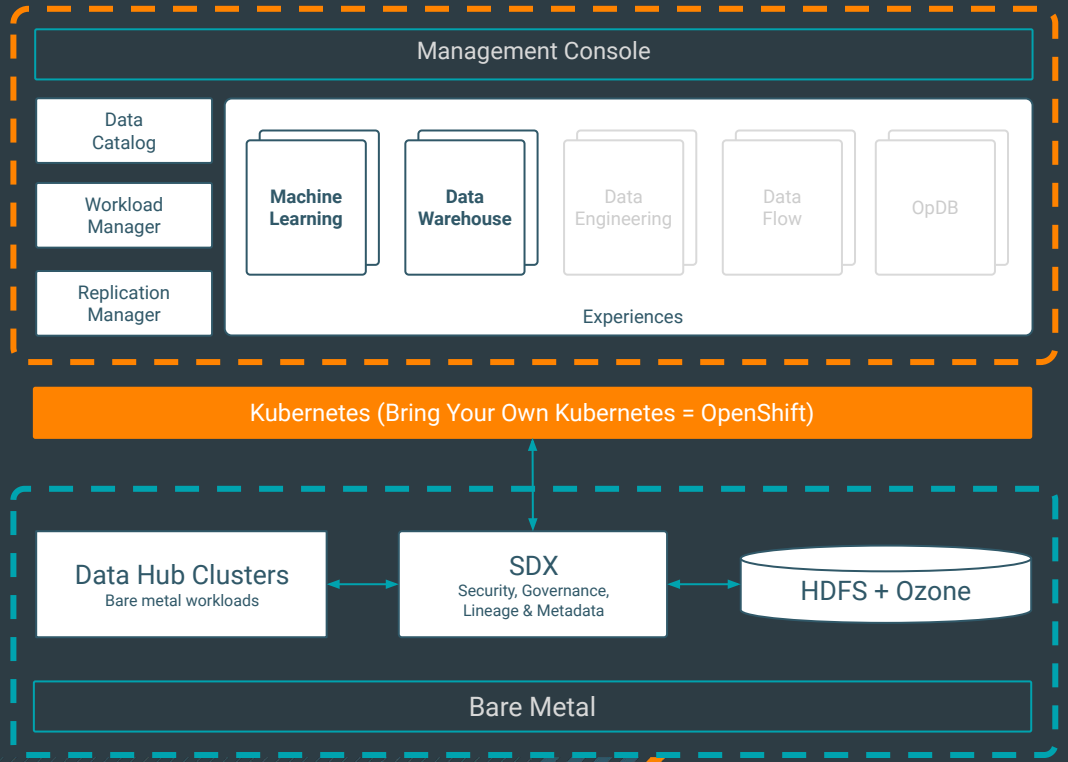
- PLUS +
- All Experiences
- Additional features (TBD)

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

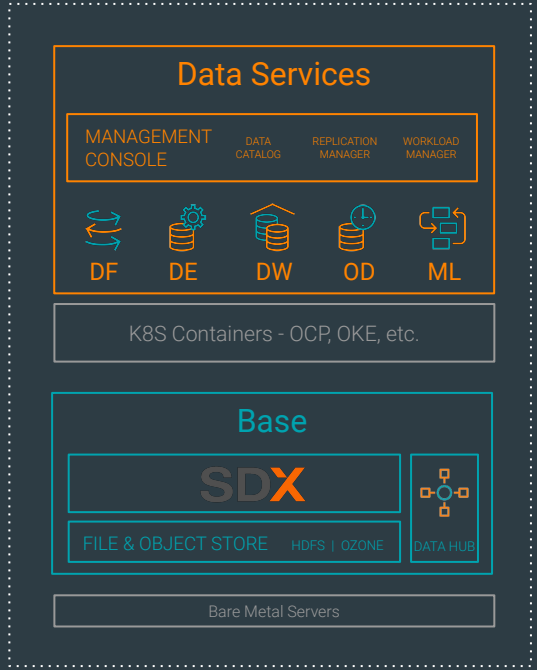


CDP PRIVATE CLOUD – HARDWARE REQUIREMENTS

	Bare metal or VMs		OpenShift/ECS 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?

Data Services	Data Engineering	$650/CCU
	Data Warehouse	
	Machine Learning	
Base	Ozone Object Store with SDX	$10,000/Node + Variable Compute & Storage
	Select 3rd Party Storage with SDX	
	HDFS with SDX	
	Traditional Data Clusters	



Cloudera Compute Unit (CCU) *- 1 Core and 8 GB RAM*

Variable compute price: *$75 per CCU over 16 Cores / 128GB RAM Node cap;*

Variable storage price: *HDFS: $25 per TB over 48TB Node cap or Ozone/Third Party Storage $100 per TB over 48TB Node cap.*

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

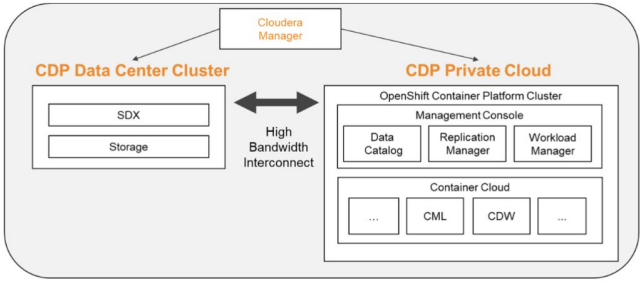


Figure 2 CDP Private Cloud components

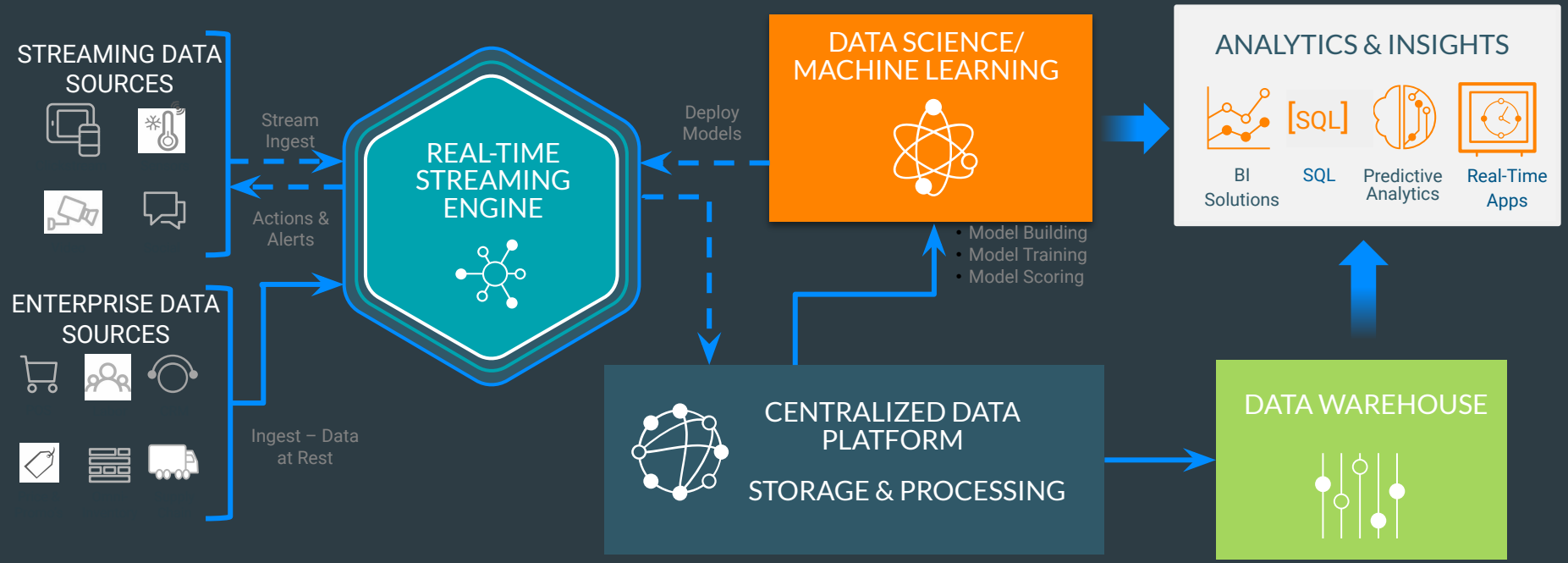
A full installation consists of two cooperating clusters – CDP Data Center, and a private cloud platform running OpenShift. The primary functions are:

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

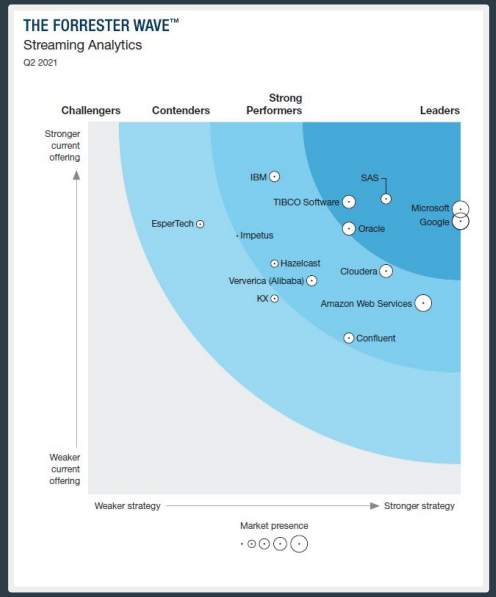
ENABLING ANALYTICS & INSIGHTS ANYWHERE

Driving Enterprise Business Value

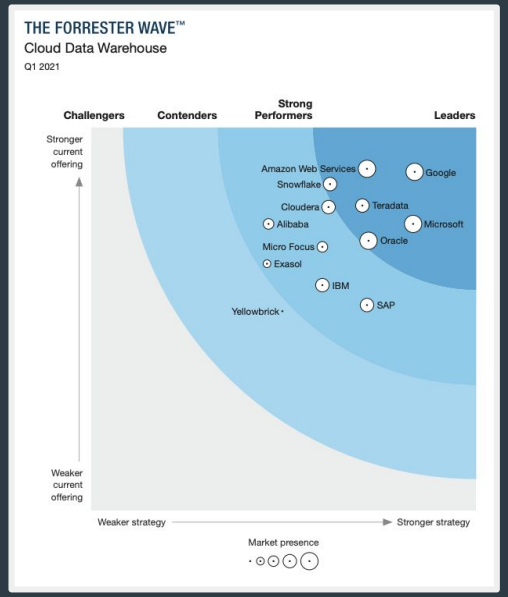


CLOUDERA RECOGNIZED IN FORRESTER WAVES

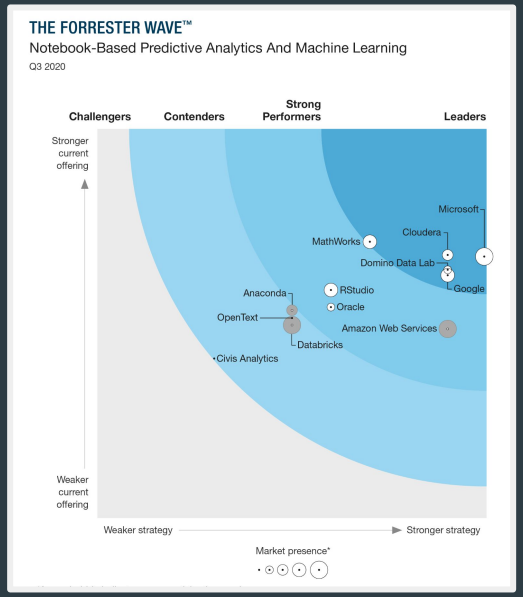
Streaming Analytics



Data Warehouse

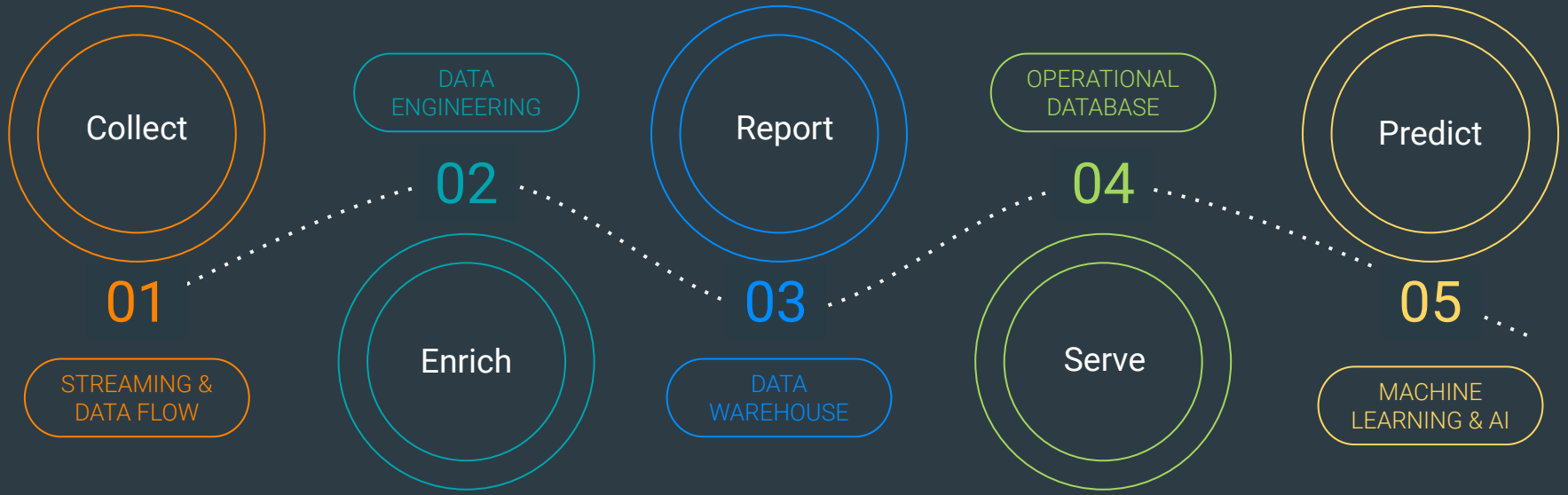


Machine Learning



CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter

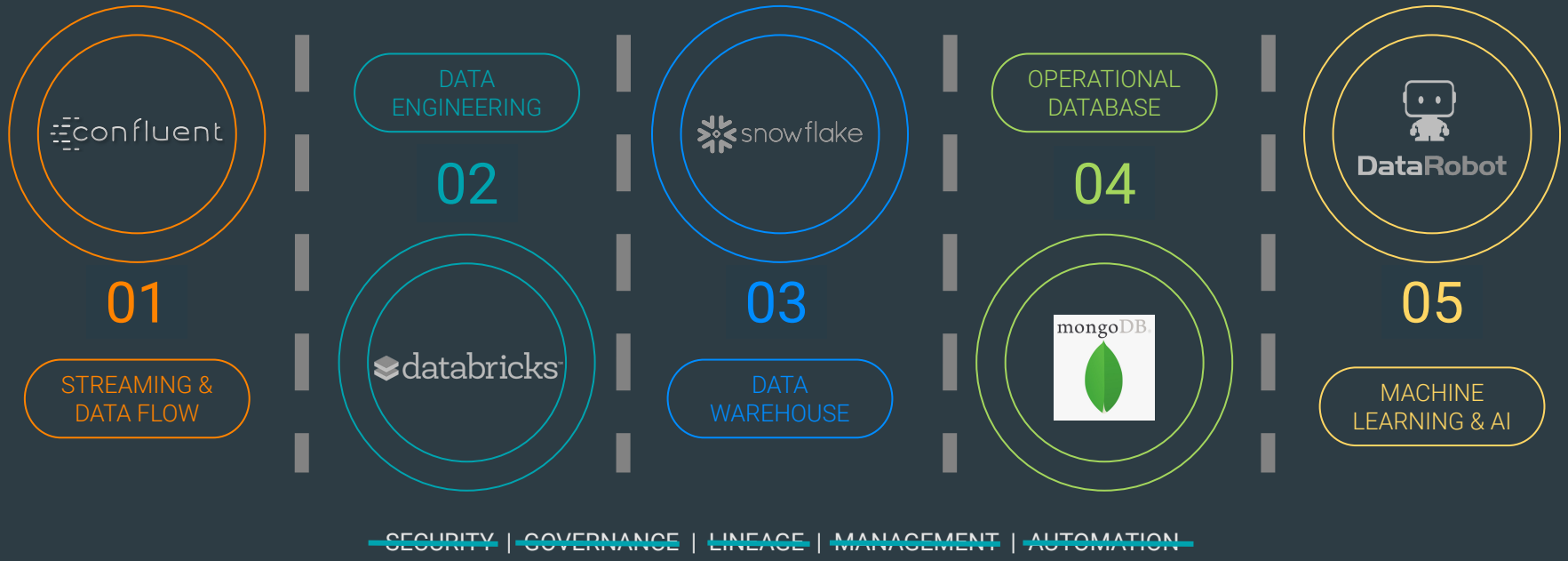


CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

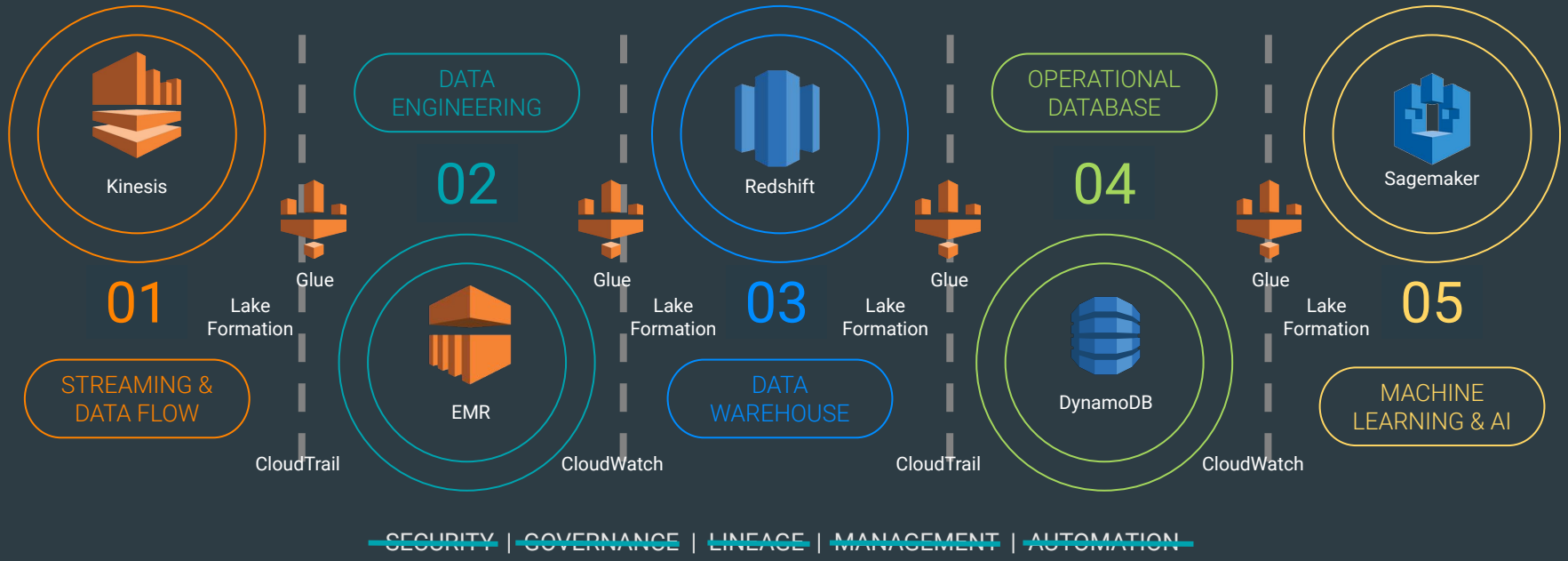
POINT SOLUTIONS HAVE AN INTEGRATION TAX

Security & governance is an afterthought



POINT SOLUTIONS FROM PUBLIC CLOUD PROVIDERS

Building blocks



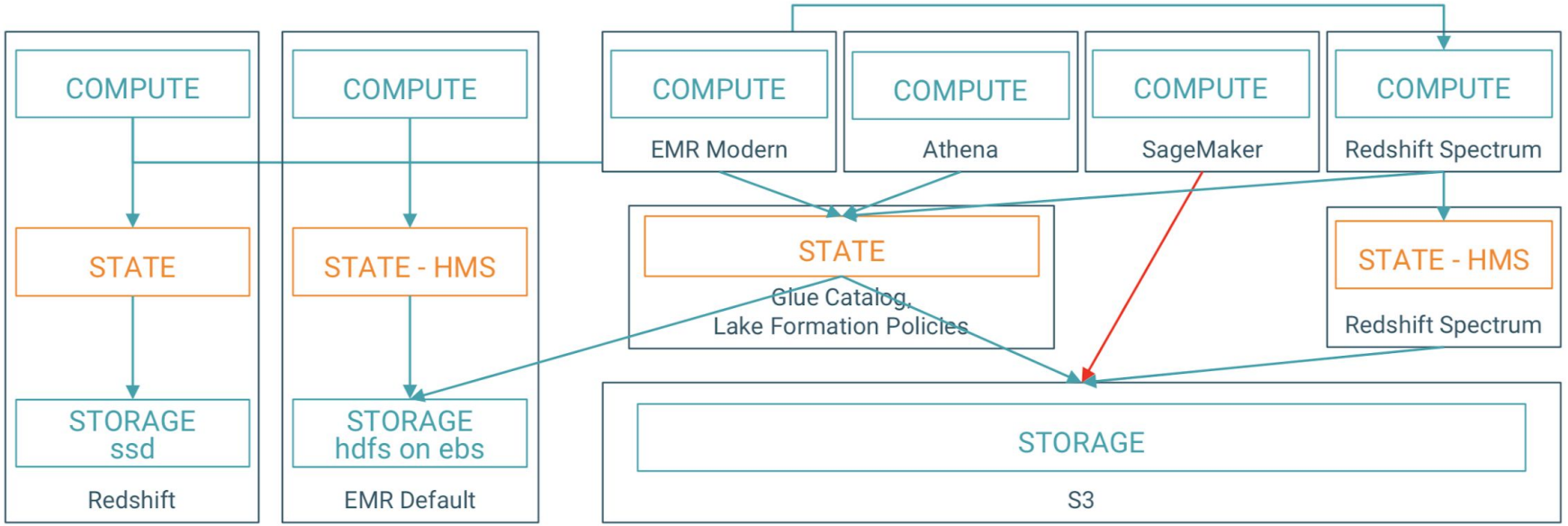
Key points to consider – CDP Vs AWS Native Offerings

- **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.

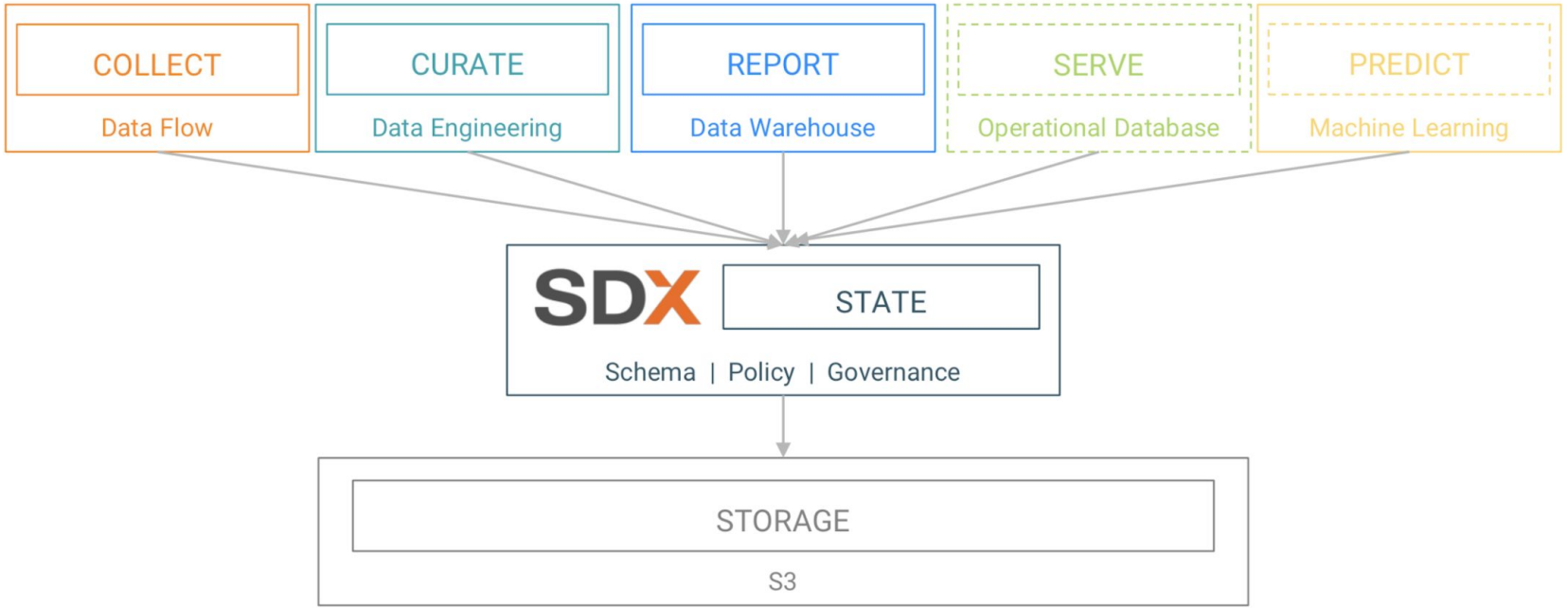
	Cloudera	AWS EMR
Unified Security / Governance	✔	Lake Formation*
Data Engineering	✔	EMR*
Machine Learning	✔	Sagemaker*
Data Warehouse	✔	***Redshift****
Ingest / Streaming	✔	Kinesis*
Unified Data Platform	✔	
Hybrid Cloud	✔	Outpost*
Open Source	✔	
Long Running Cost Optimized	✔	

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

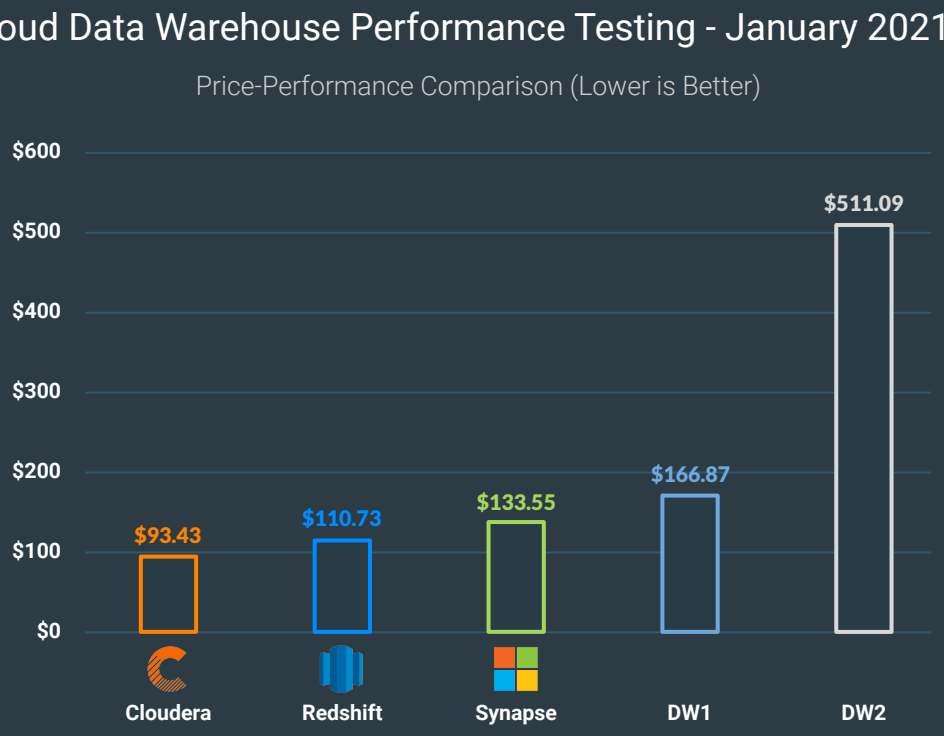
Industry standard TPC Benchmark

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