ADATPLATFORM A FELHŐBEN MENNI VAGY NEM MENNI?

Gáspár Balázs | Solutions Engineer

2021 június 23.

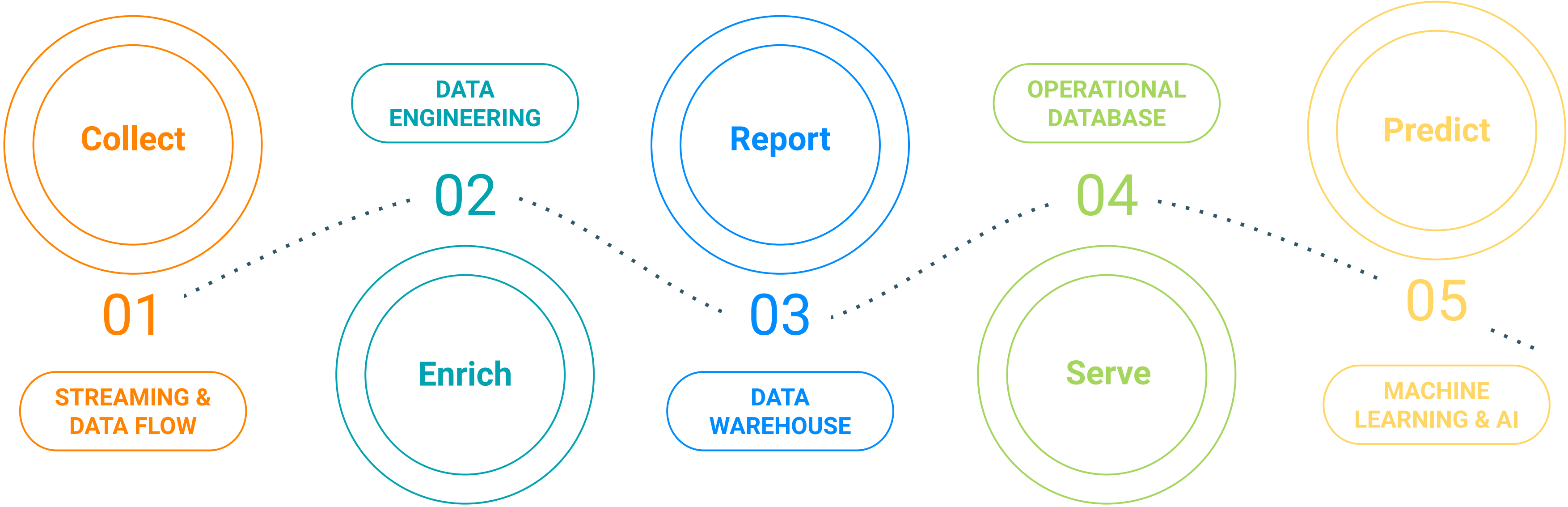
AGENDA

Mitől modern egy adatplatform?

Adatmenedzsment a felhőben

Az út a felhőbe

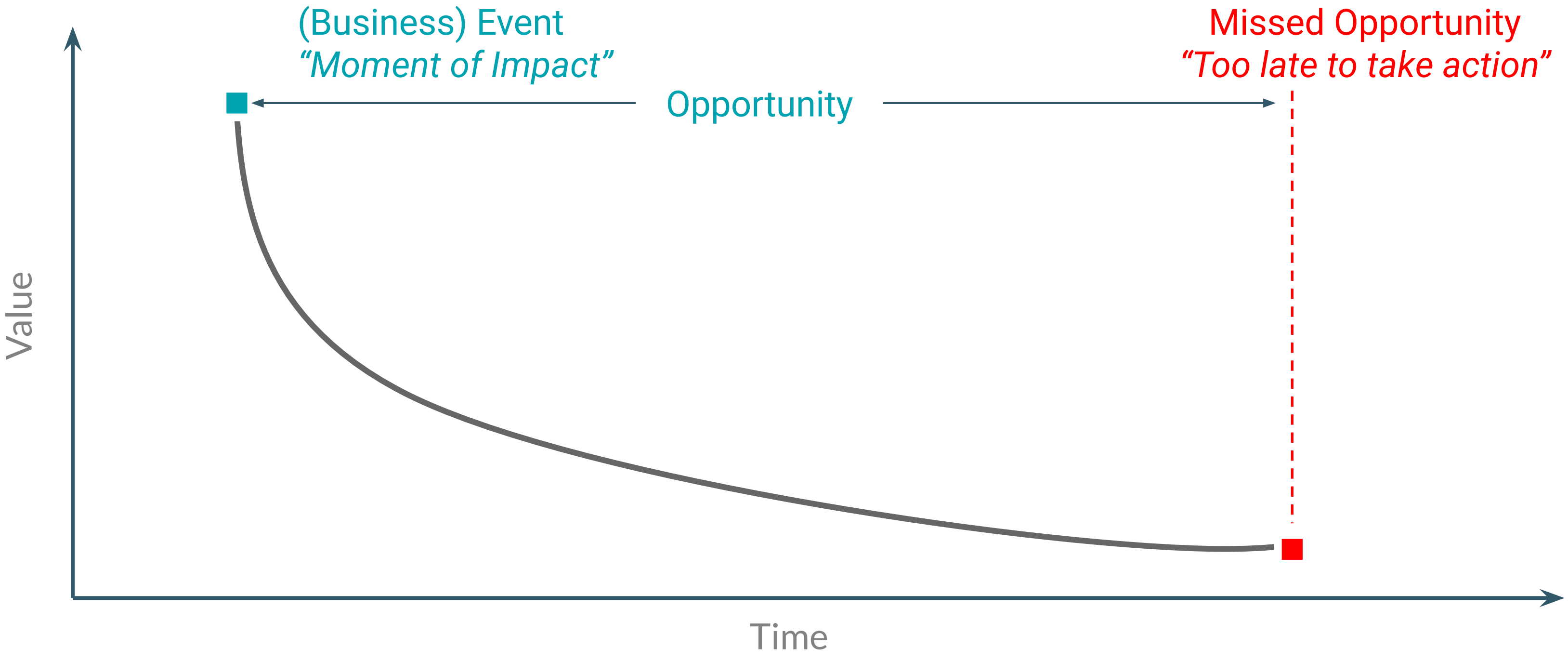
MANAGING THE DATA LIFECYCLE - BATCH AND STREAMING



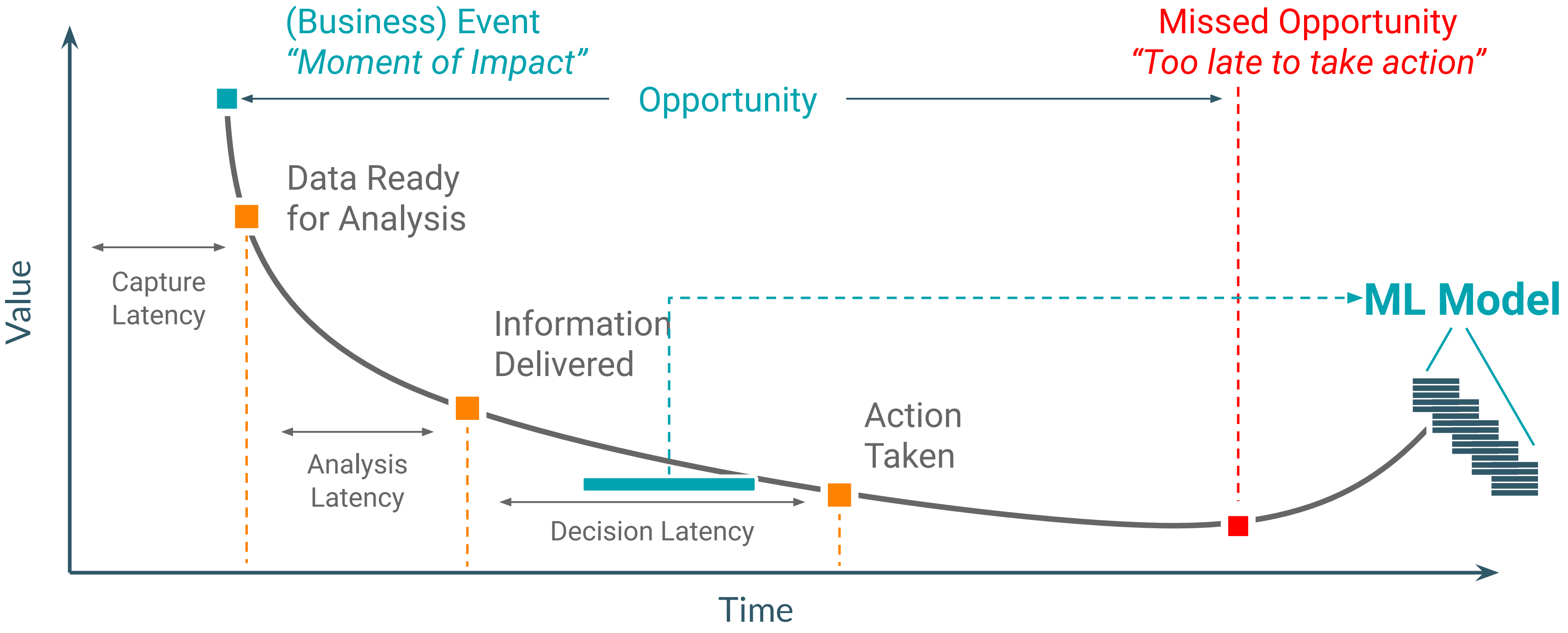
CLOUDERA SDX SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

Streaming and machine learning are platform problems

CREATING VALUE FROM THE REAL-TIME FLOW OF BIG DATA

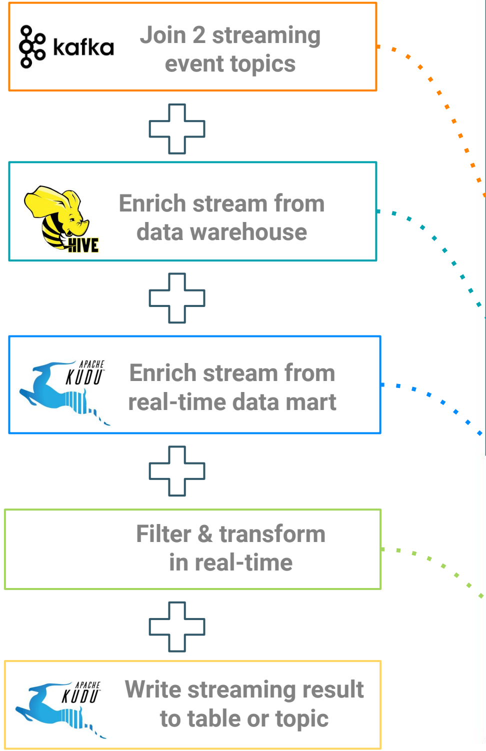


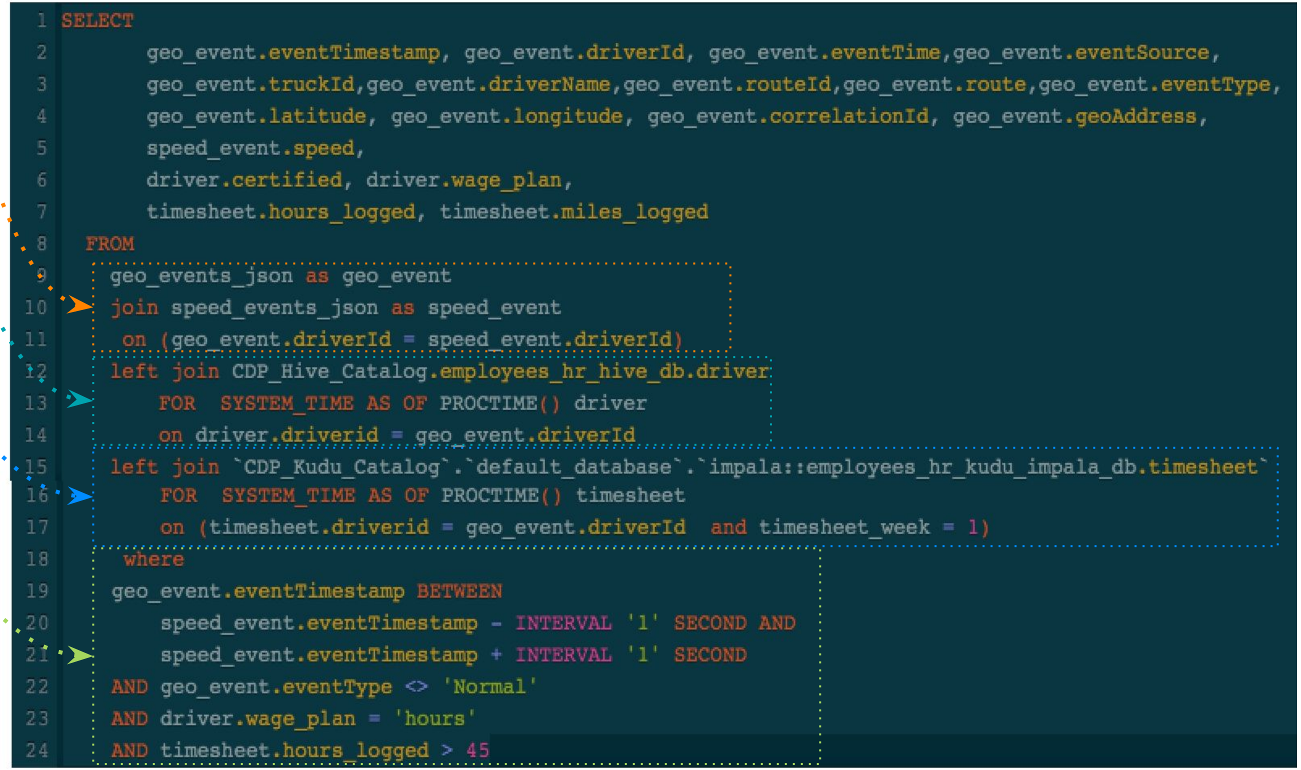
CREATING VALUE FROM THE REAL-TIME FLOW OF BIG DATA



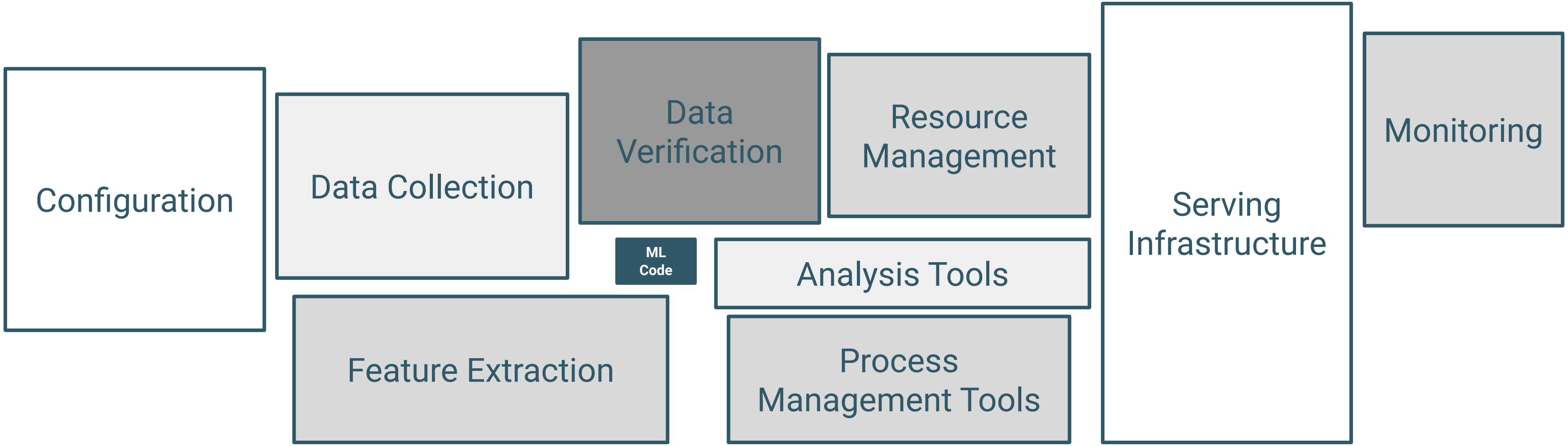
CLOUDERA

IMAGINE A WORLD WITH SQL ANALYTICS ON *ALL AND ANY DATA*


kafka
Join 2 streaming event topics
HIVE
Enrich stream from data warehouse
APACHE KUDU
Enrich stream from real-time data mart
Filter & transform in real-time
APACHE KUDU
Write streaming result to table or topic


```
SELECT
       geo_event.eventTimestamp, geo_event.driverId, geo_event.eventTime,geo_event.eventSource,
       geo_event.truckId,geo_event.driverName,geo_event.routeId,geo_event.route,geo_event.eventType,
       geo_event.latitude, geo_event.longitude, geo_event.correlationId, geo_event.geoAddress,
       speed_event.speed,
       driver.certified, driver.wage_plan,
       timesheet.hours_logged, timesheet.miles_logged
  FROM
    geo_events_json as geo_event
    join speed_events_json as speed_event
     on (geo_event.driverId = speed_event.driverId)
    left join CDP_Hive_Catalog.employees_hr_hive_db.driver
        FOR  SYSTEM_TIME AS OF PROCTIME() driver
        on driver.driverid = geo_event.driverId
    left join `CDP_Kudu_Catalog`.`default_database`.`impala::employees_hr_kudu_impala_db.timesheet`
        FOR  SYSTEM_TIME AS OF PROCTIME() timesheet
        on (timesheet.driverid = geo_event.driverId  and timesheet_week = 1)
     where
    geo_event.eventTimestamp BETWEEN
        speed_event.eventTimestamp - INTERVAL '1' SECOND AND
        speed_event.eventTimestamp + INTERVAL '1' SECOND
    AND geo_event.eventType <> 'Normal'
    AND driver.wage_plan = 'hours'
    AND timesheet.hours_logged > 45
```

HIDDEN TECHNICAL DEBT IN MACHINE LEARNING SYSTEMS



Source: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

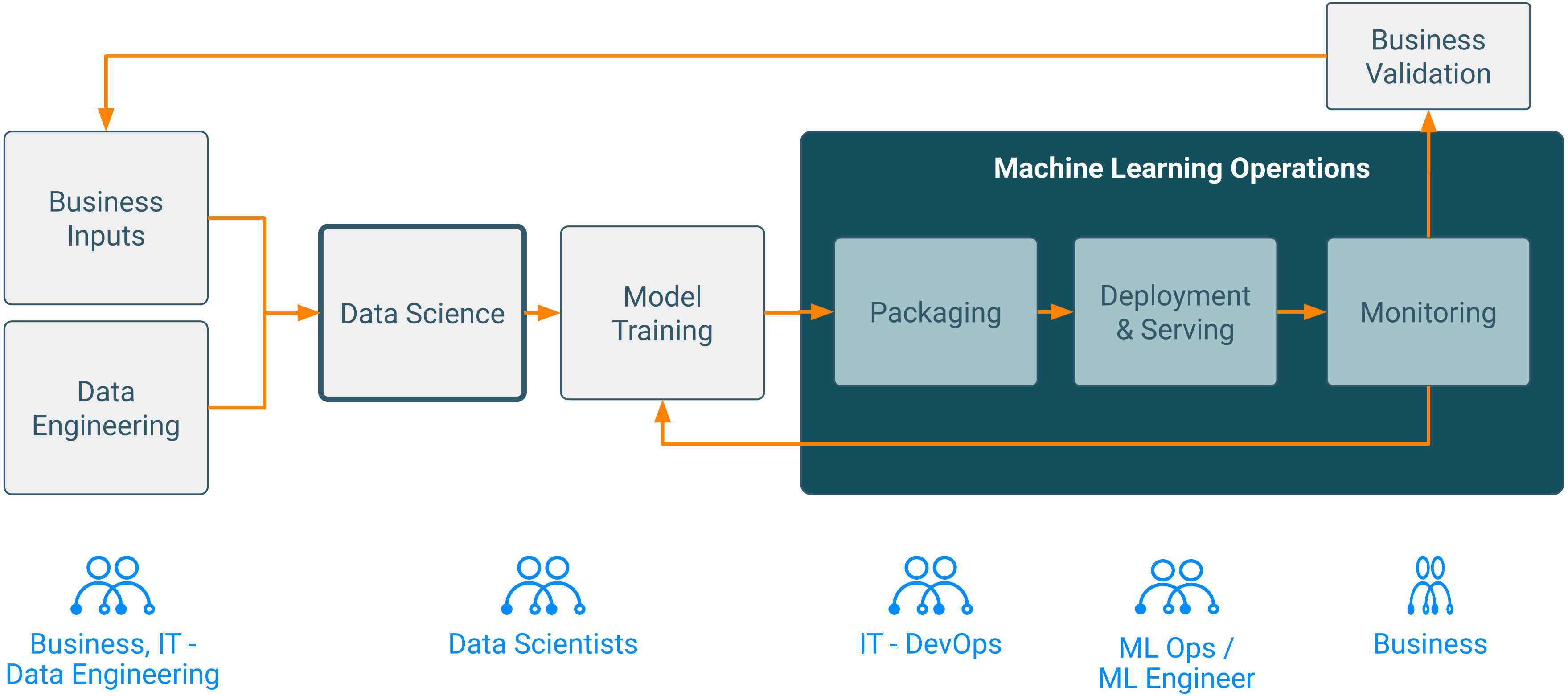
35%

Making it to production

Currently only 35% of organizations indicate that analytical models are fully deployed in production and are often challenged in the “last mile” of the complex and iterative ML workflow

** IDC’s Advanced and Predictive Analytics survey and interviews, n = 400, 2017 – 2019*

MACHINE LEARNING IN PRODUCTION



Business Validation

Business Inputs

Data Engineering

Data Science

Model Training

Machine Learning Operations

Packaging

Deployment & Serving

Monitoring

Business, IT - Data Engineering

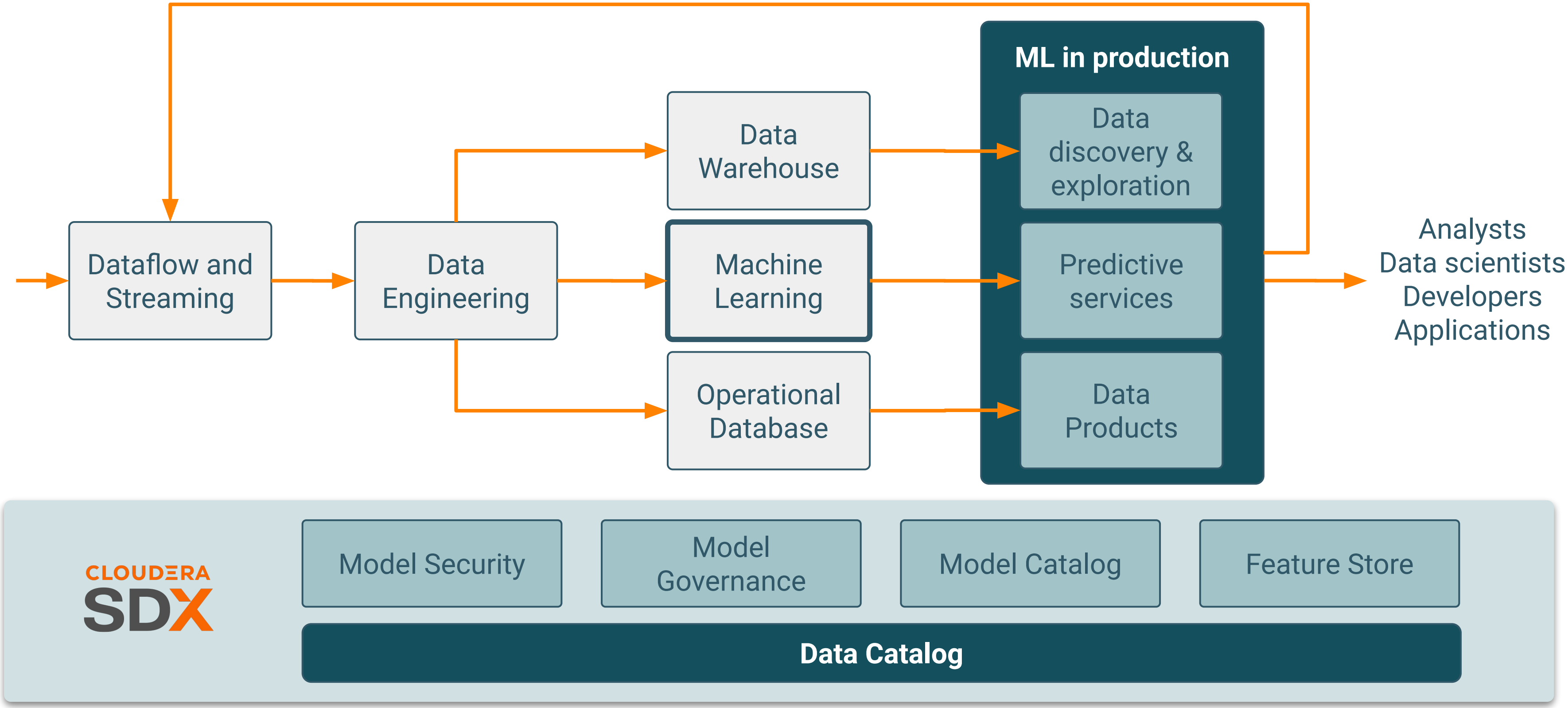
Data Scientists

IT - DevOps

ML Ops / ML Engineer

Business

MACHINE LEARNING IN PRODUCTION



AGENDA

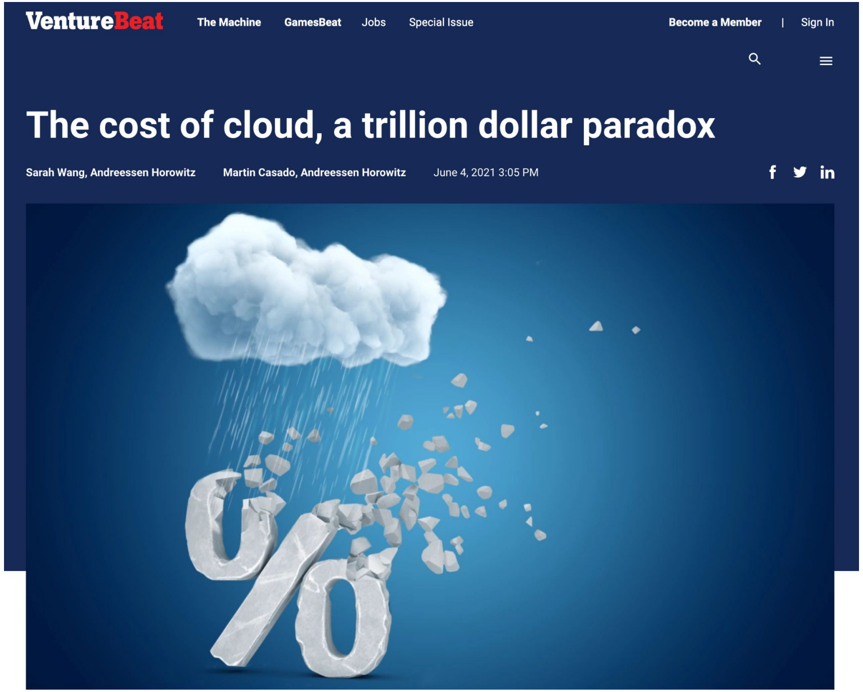
Mitől modern egy adatplatform?

■ **Adatmenedzsment a felhőben**

Az út a felhőbe

[...] on the other hand, we have the phenomenon we've outlined in this post, where the cost of cloud "takes over" at some point, locking up hundreds of billions of market cap that are now stuck in this paradox:

You're crazy if you don't start in the cloud; you're crazy if you stay on it.

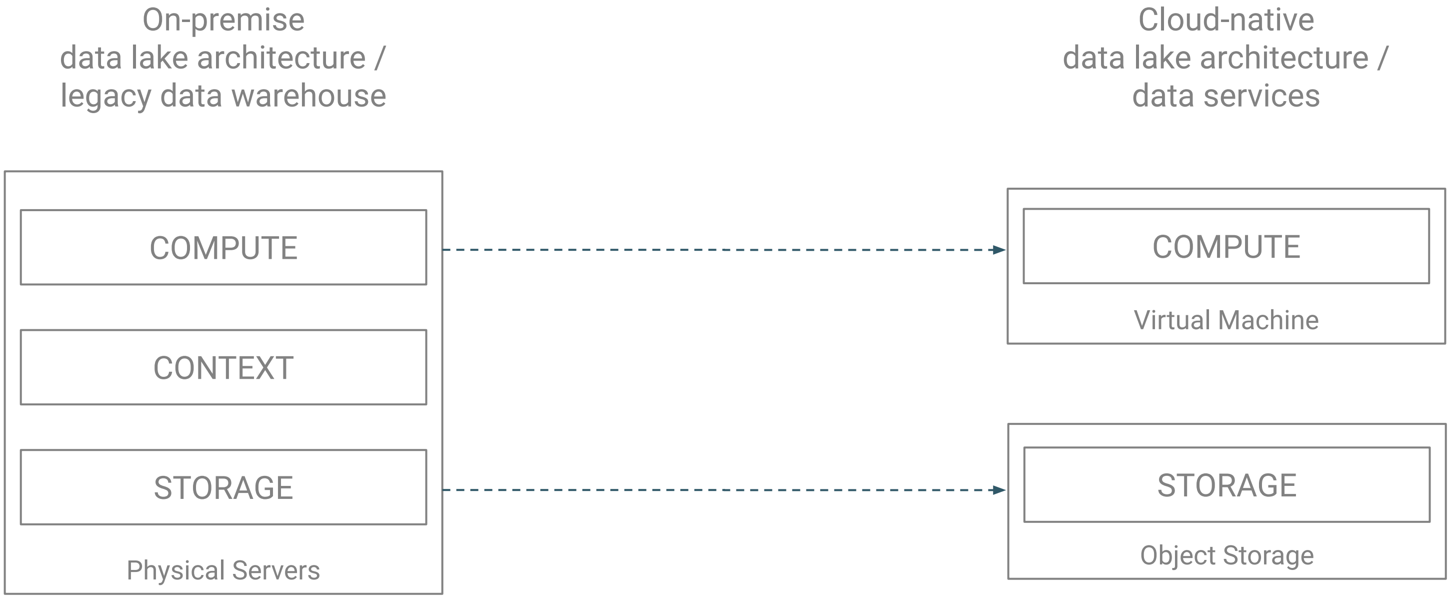


The cost of cloud, a trillion dollar paradox

Sarah Wang, Andreessen Horowitz   Martin Casado, Andreessen Horowitz   June 4, 2021 3:05 PM

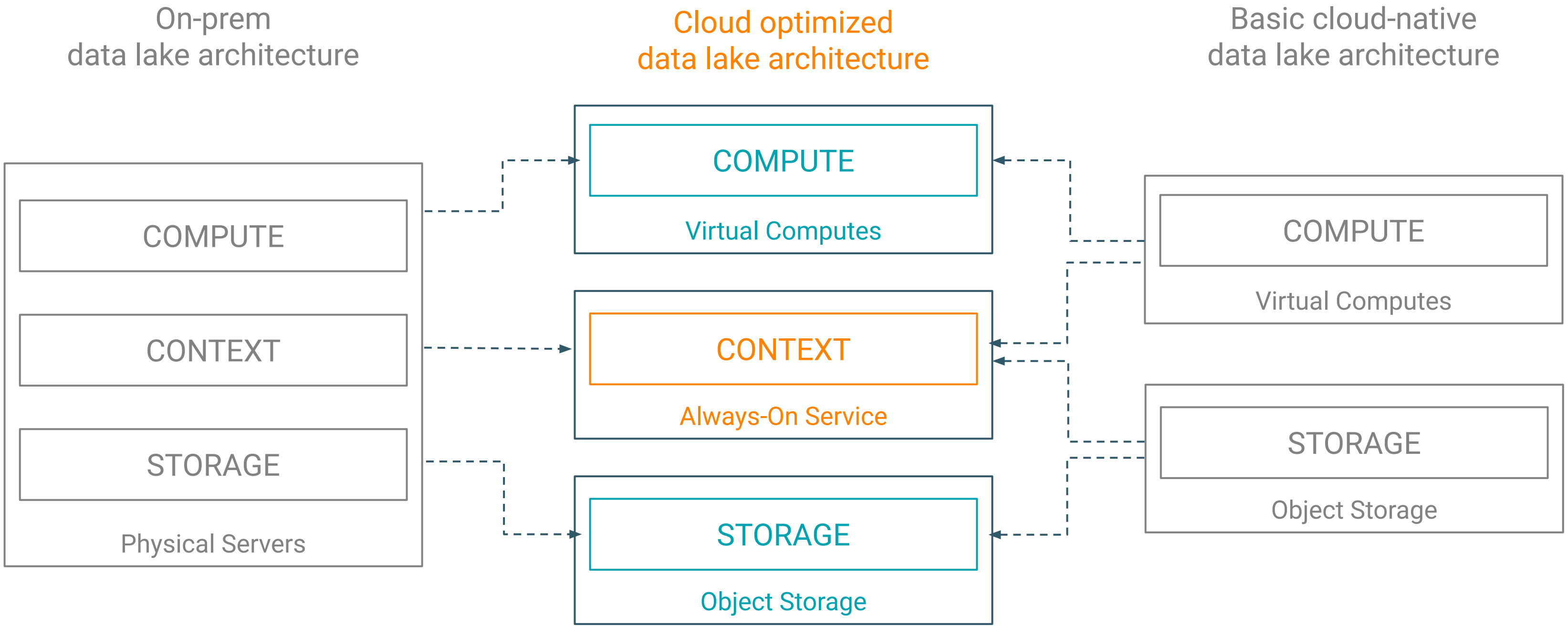
https://venturebeat.com/2021/06/04/the-cost-of-cloud-a-trillion-dollar-paradox

CLOUD 101: "SEPARATION OF COMPUTE AND STORAGE"



CLOUDERA

A DATA CLOUD SEPARATES STORAGE & COMPUTE & CONTEXT



CLOUD DATA LAKES REQUIRE …

Identities

How do we easily connect to **corporate** identity?

Schema

How do multiple personas **find** and **share** datasets across different compute engines?

Policy

Is there a central place to protect sensitive data at a **field** and **row** level?

Audits

Can I comply to the regulatory requirements with comprehensive **audits** and **lineage**?

← Privacy, compliance & regulation →

… AND *ACTIVE* MANAGEMENT

Data Profiling & Stewardship

Can I detect **sensitive** data automatically, **tag** and **curate** for easy discovery and security controls?

Replication

What happens to metadata and lineage, when I move data **across environments**?

Workload Management

How do I track workloads for **troubleshooting** and optimization with elastic computes?

Encryption

Is the data **protected** at rest and in-transit?

Context is required for all but the simplest use cases

- Multi-tenant
- Multi-workload

- Sensitive data
- Regulated use case

- Long Running Apps

AGENDA

Mitől modern egy adatplatform?

Adatmenedzsment a felhőben

■ **Az út a felhőbe**

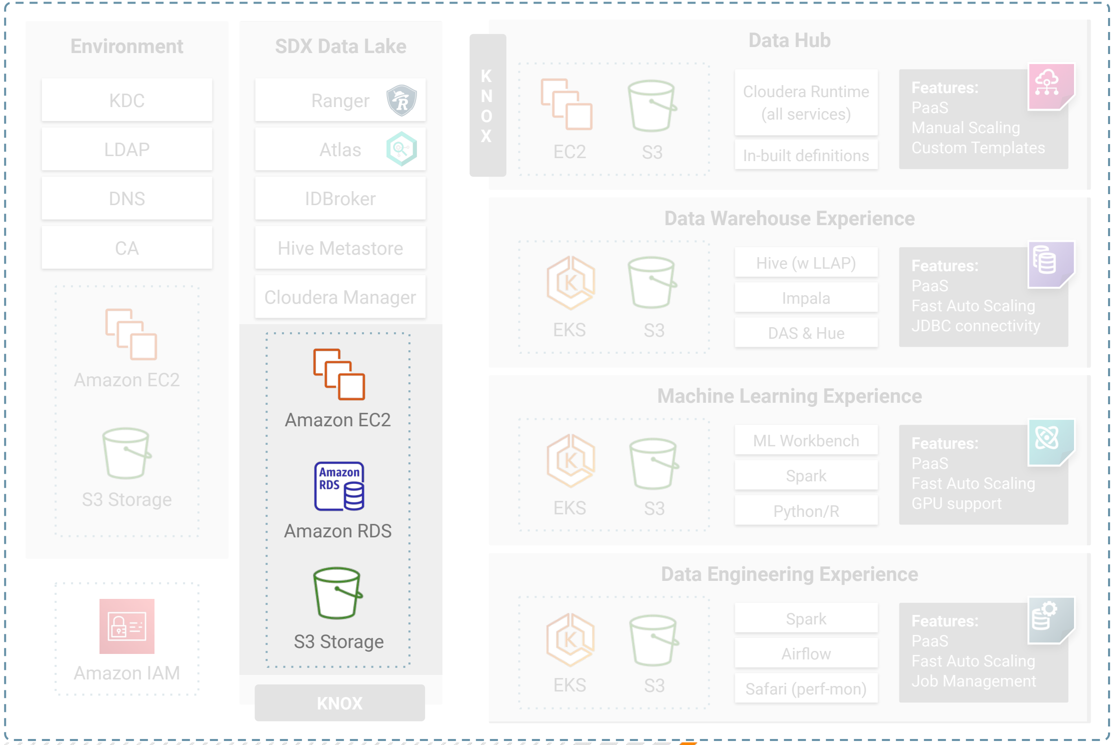
CDP PUBLIC CLOUD PHILOSOPHY

Cloudera Data Platform Public Cloud enables you to:

- **leverage existing public cloud** environment including networking and data
- **migrate existing workloads** with minimal transformation effort
- **use existing user identities** to access cloud data and CDP services
- **modernize data applications** with CDP experiences

It provides a quick way to public cloud and integrates with current cloud infrastructure blueprints.

Public cloud environment - customer VPC



Environment
KDC
LDAP
DNS
CA
Amazon EC2
S3 Storage
Amazon IAM

SDX Data Lake
Ranger
Atlas
IDBroker
Hive Metastore
Cloudera Manager
Amazon EC2
Amazon RDS
S3 Storage
KNOX

KNOX

Data Hub
EC2
S3
Cloudera Runtime (all services)
In-built definitions
Features:
PaaS
Manual Scaling
Custom Templates

Data Warehouse Experience
EKS
S3
Hive (w LLAP)
Impala
DAS & Hue
Features:
PaaS
Fast Auto Scaling
JDBC connectivity

Machine Learning Experience
EKS
S3
ML Workbench
Spark
Python/R
Features:
PaaS
Fast Auto Scaling
GPU support

Data Engineering Experience
EKS
S3
Spark
Airflow
Safari (perf-mon)
Features:
PaaS
Fast Auto Scaling
Job Management

CLOUDERA

1) LEVERAGE AN EXISTING PUBLIC CLOUD ENVIRONMENT

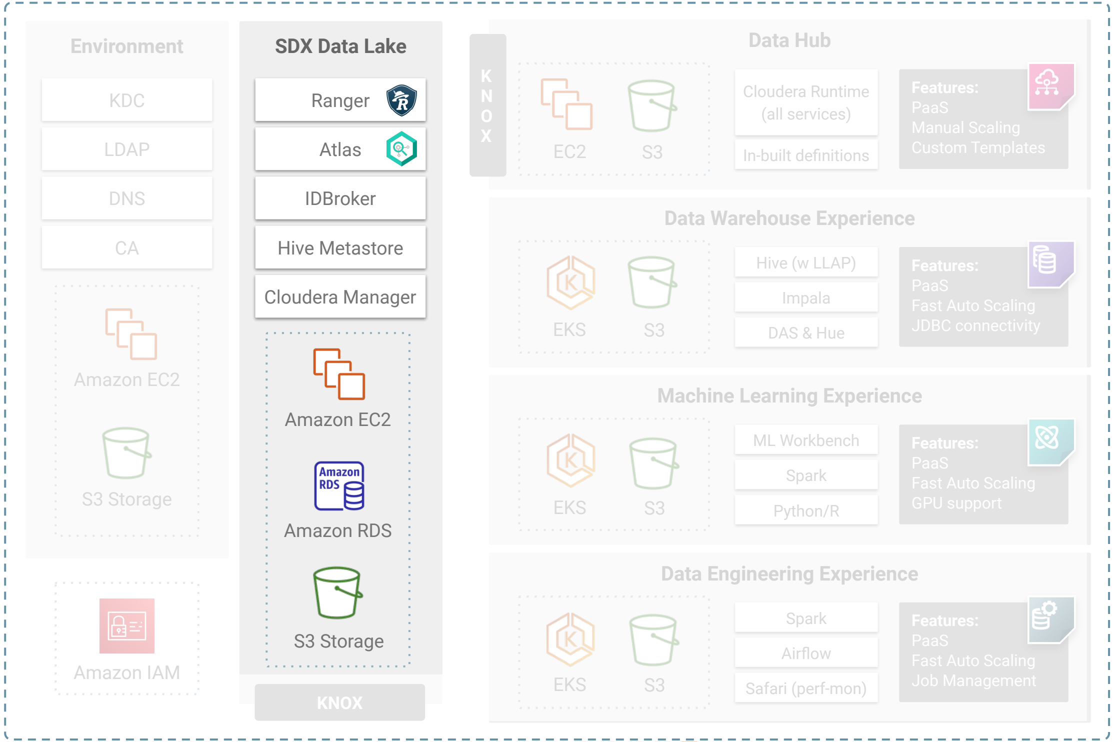
Create a **centralized data lake** that provides security and governance services for all **data and services**.

The **SDX** (Shared Data Experience) layer provides unified services:

- schema / metastore
- authorization (fine-grain access to S3 as well as structured data)
- data catalog / metadata management with visual lineage
- audit service to track data access

All data and SDX metadata lives within a customer environment.

Public cloud environment - customer VPC



Environment: KDC, LDAP, DNS, CA, Amazon EC2, S3 Storage, Amazon IAM

SDX Data Lake: Ranger, Atlas, IDBroker, Hive Metastore, Cloudera Manager, Amazon EC2, Amazon RDS, S3 Storage, KNOX

KNOX

Data Hub: EC2, S3; Cloudera Runtime (all services); In-built definitions; Features: PaaS, Manual Scaling, Custom Templates

Data Warehouse Experience: EKS, S3; Hive (w LLAP), Impala, DAS & Hue; Features: PaaS, Fast Auto Scaling, JDBC connectivity

Machine Learning Experience: EKS, S3; ML Workbench, Spark, Python/R; Features: PaaS, Fast Auto Scaling, GPU support

Data Engineering Experience: EKS, S3; Spark, Airflow, Safari (perf-mon); Features: PaaS, Fast Auto Scaling, Job Management

2) MIGRATE EXISTING WORKLOADS WITH MINIMAL EFFORT

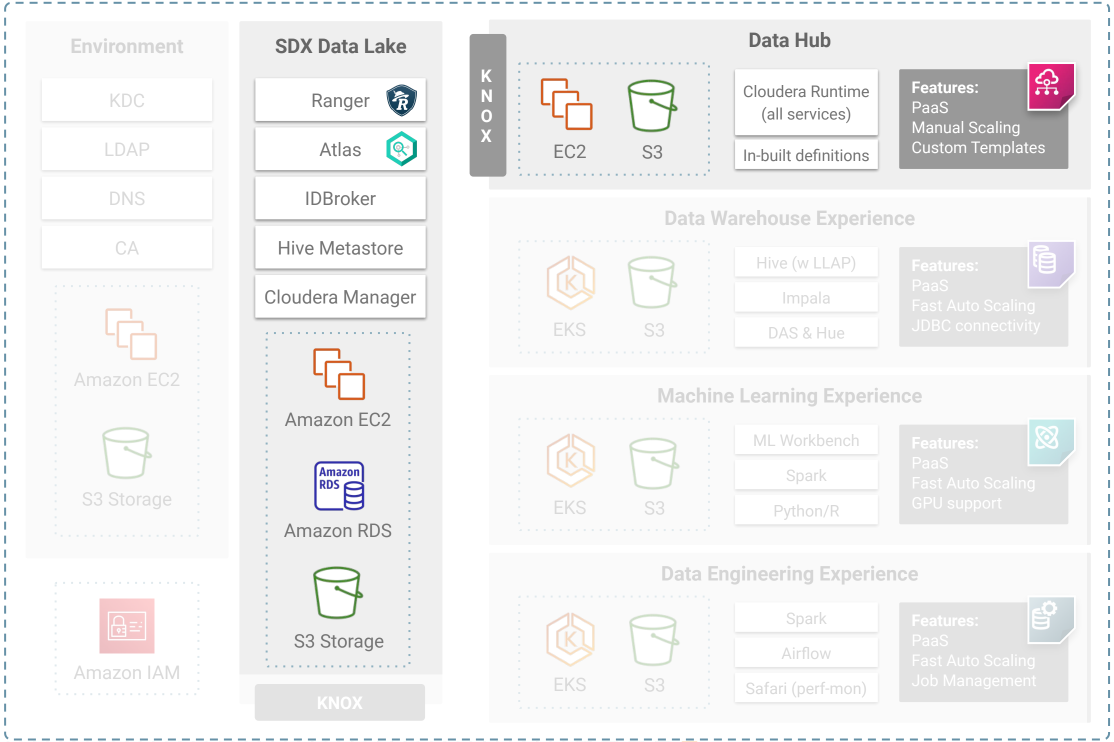
Create fully secured and integrated **clusters-as-a-service** within minutes

This is best suited for **lift & shift** of existing on-premise Cloudera workloads or complex applications.

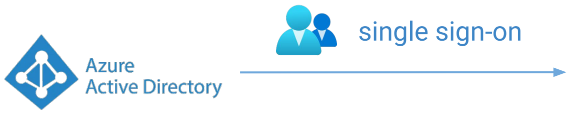
Data Hub uses **Cloudera Runtime 7.x** in the public cloud, the same platform codebase as CDP on-prem.

All data is **covered by SDX and is persisted on S3**, enabling seamless access from non-Cloudera apps.

Public cloud environment - customer VPC



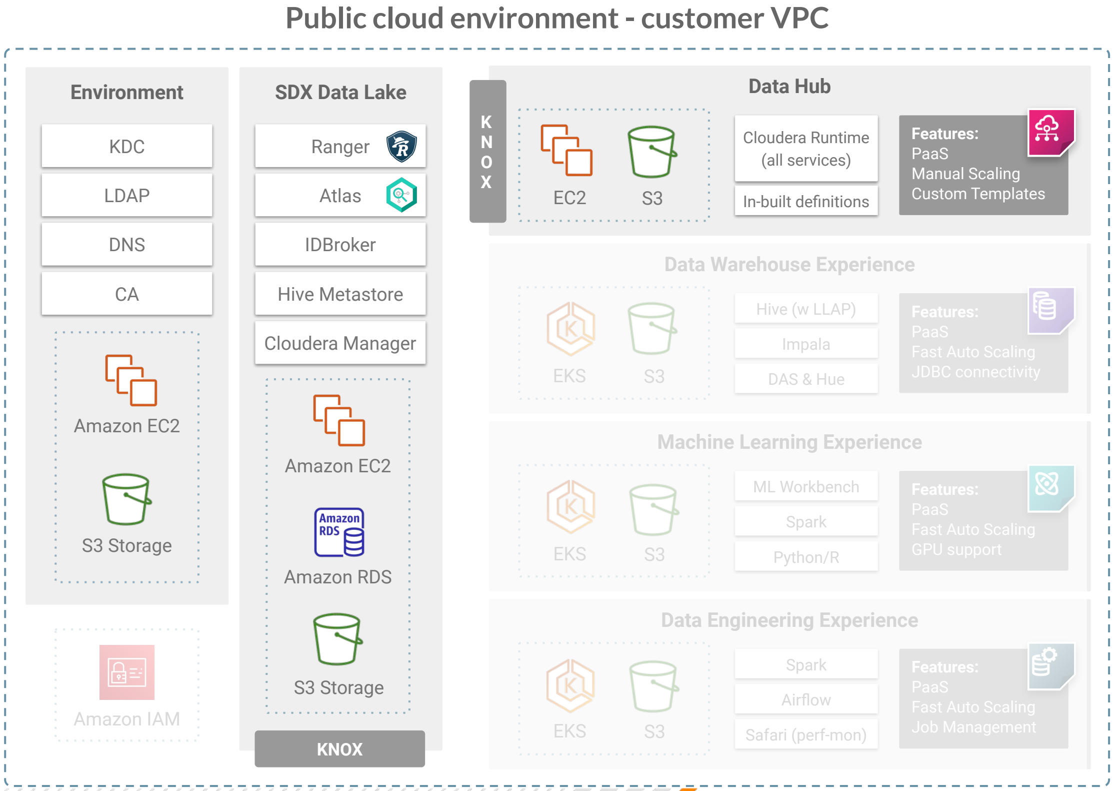
3) USE EXISTING USER IDENTITIES TO ACCESS CDP SERVICES



CDP provides centralized backend and management services (KDC, LDAP, DNS, certificates) for all services within a CDP **environment**.

Customers **bring existing user identities via single sign-on** (SAML).

CDP creates a unified identity for each user and automatically provides secure access to data, UIs, API endpoints and SQL interfaces.



4) MODERNIZE DATA APPLICATIONS WITH CDP EXPERIENCES

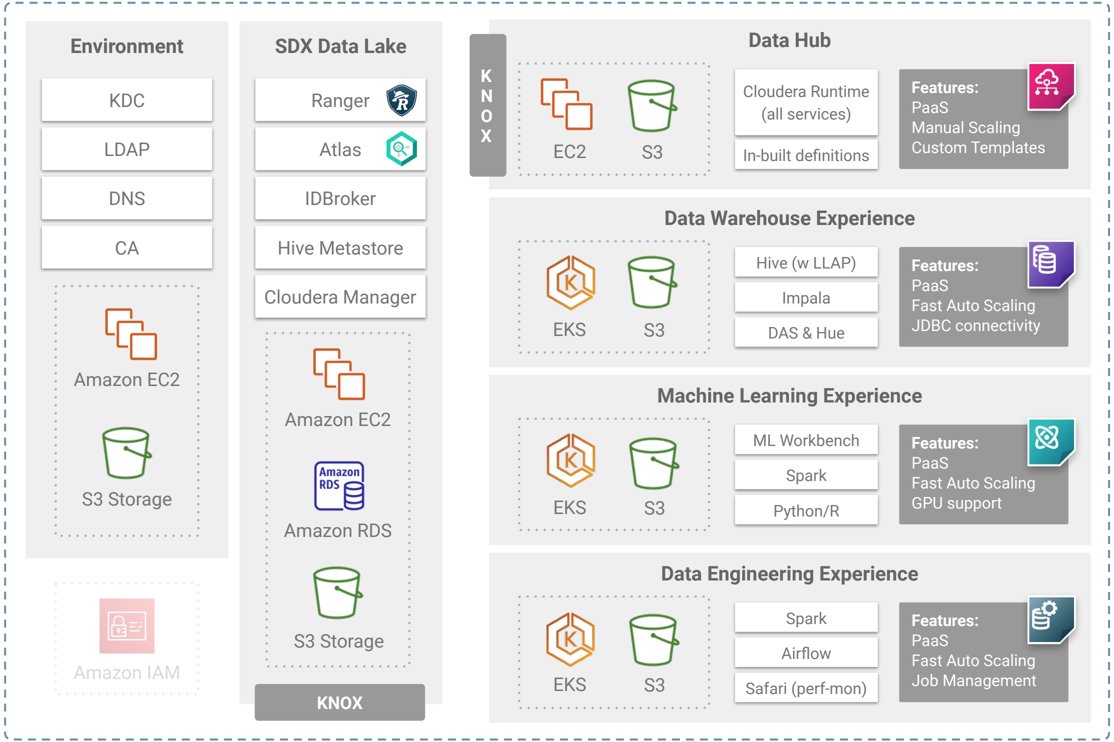
Move beyond clusters and adopt CDP experiences, cloud services that provide **on-demand autoscaling**, **workload isolation** and a **redefined self-service user experience**:

Independent data warehouses and data marts that autoscale to meet workload demands with **CDW**.

Unified self-service data science and data engineering in a single, portable service with **CML**.

Hive and Spark jobs on an auto-scaling cluster scheduled with Apache Airflow with **CDE**.

Public cloud environment - customer VPC



Environment

KDC

LDAP

DNS

CA

Amazon EC2

S3 Storage

Amazon IAM

SDX Data Lake

Ranger

Atlas

IDBroker

Hive Metastore

Cloudera Manager

Amazon EC2

Amazon RDS

S3 Storage

KNOX

K N O X

Data Hub

EC2 S3

Cloudera Runtime (all services)

In-built definitions

Features: PaaS, Manual Scaling, Custom Templates

Data Warehouse Experience

EKS S3

Hive (w LLAP)

Impala

DAS & Hue

Features: PaaS, Fast Auto Scaling, JDBC connectivity

Machine Learning Experience

EKS S3

ML Workbench

Spark

Python/R

Features: PaaS, Fast Auto Scaling, GPU support

Data Engineering Experience

EKS S3

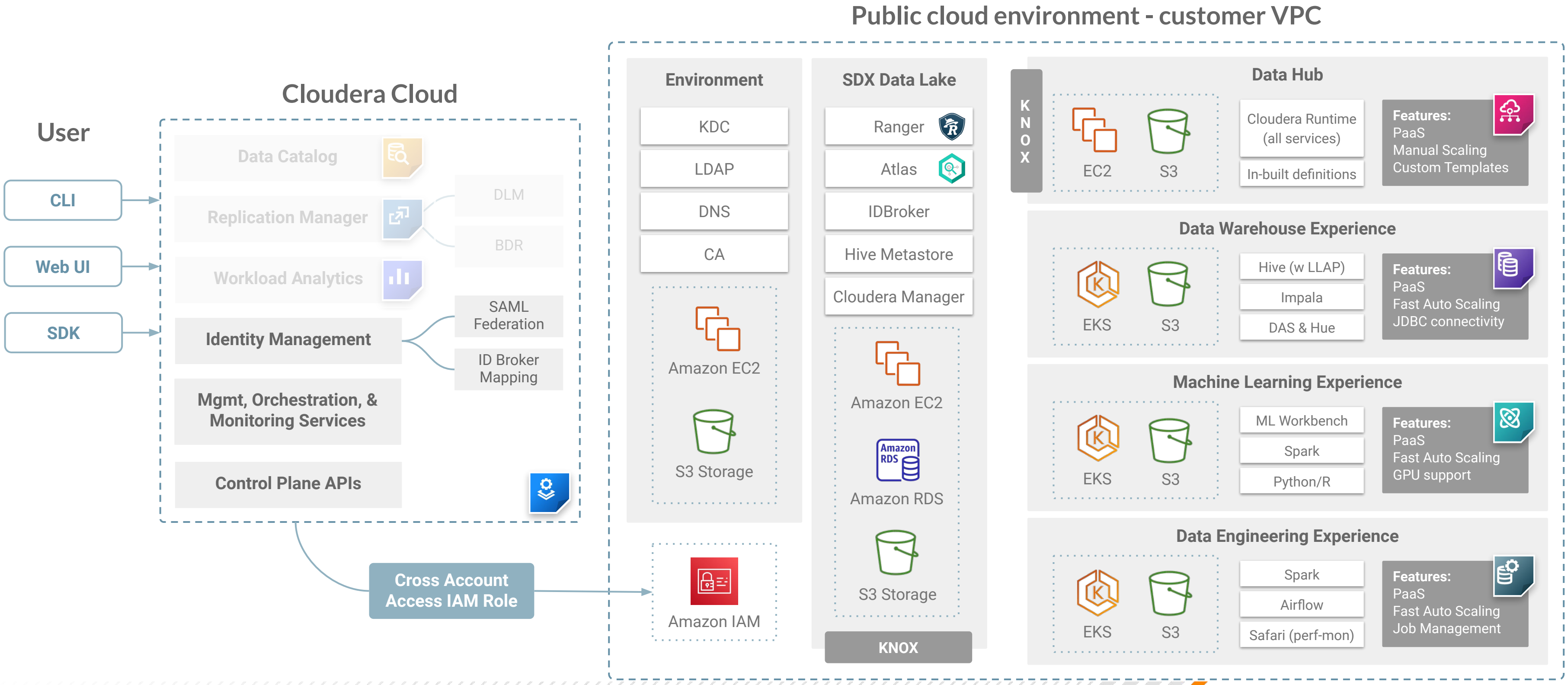
Spark

Airflow

Safari (perf-mon)

Features: PaaS, Fast Auto Scaling, Job Management

… ALL MANAGED FROM A UNIFIED SINGLE PANE OF GLASS



Public cloud environment - customer VPC

User

CLI

Web UI

SDK

Cloudera Cloud

Data Catalog

Replication Manager

DLM

BDR

Workload Analytics

SAML Federation

Identity Management

ID Broker Mapping

Mgmt, Orchestration, & Monitoring Services

Control Plane APIs

Cross Account Access IAM Role

Environment

KDC

LDAP

DNS

CA

Amazon EC2

S3 Storage

Amazon IAM

SDX Data Lake

Ranger

Atlas

IDBroker

Hive Metastore

Cloudera Manager

Amazon EC2

Amazon RDS

S3 Storage

KNOX

KNOX

Data Hub

EC2

S3

Cloudera Runtime (all services)

In-built definitions

Features:
PaaS
Manual Scaling
Custom Templates

Data Warehouse Experience

EKS

S3

Hive (w LLAP)

Impala

DAS & Hue

Features:
PaaS
Fast Auto Scaling
JDBC connectivity

Machine Learning Experience

EKS

S3

ML Workbench

Spark

Python/R

Features:
PaaS
Fast Auto Scaling
GPU support

Data Engineering Experience

EKS

S3

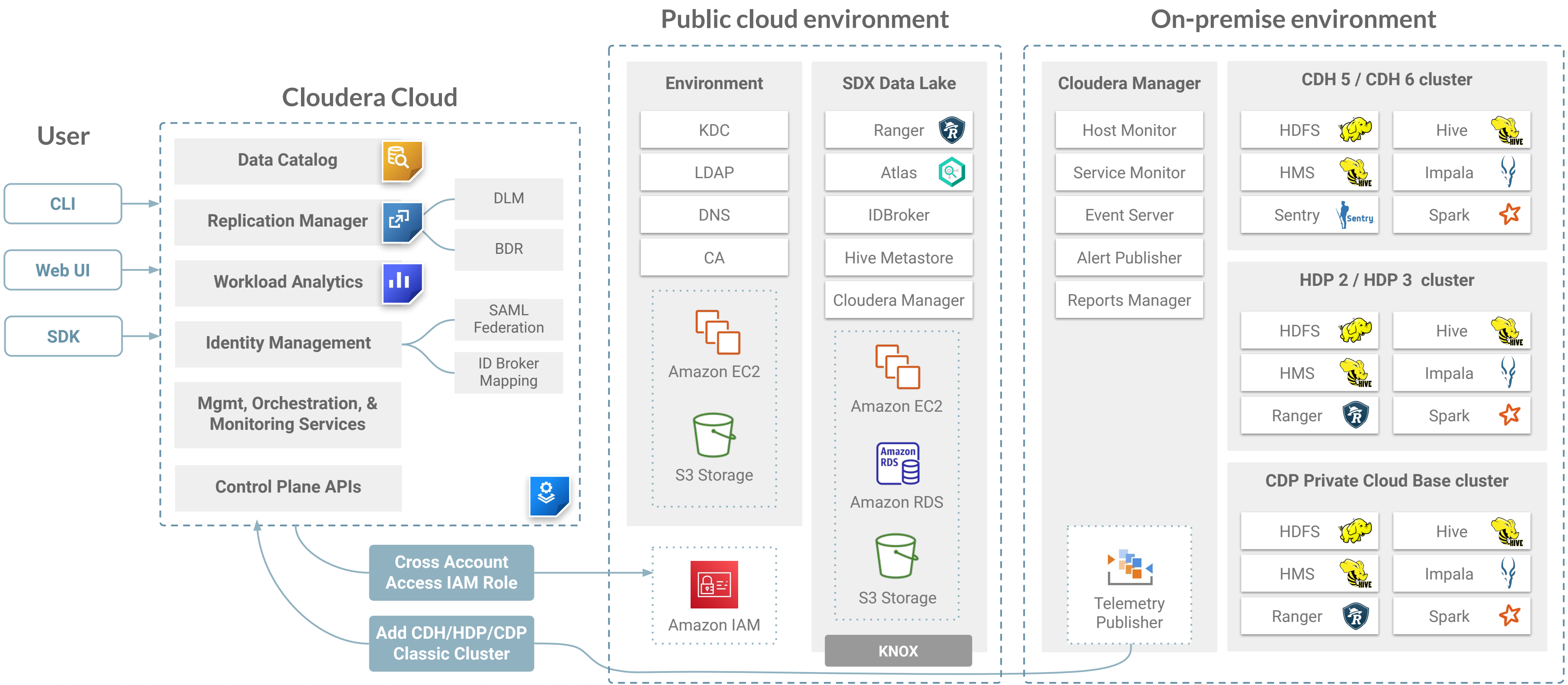
Spark

Airflow

Safari (perf-mon)

Features:
PaaS
Fast Auto Scaling
Job Management

… SUPPORTING MIGRATIONS AND A TRUE HYBRID EXPERIENCE



User

- CLI
- Web UI
- SDK

Cloudera Cloud

- Data Catalog
- Replication Manager
  - DLM
  - BDR
- Workload Analytics
- Identity Management
  - SAML Federation
  - ID Broker Mapping
- Mgmt, Orchestration, & Monitoring Services
- Control Plane APIs

Cross Account Access IAM Role

Add CDH/HDP/CDP Classic Cluster

Public cloud environment

Environment

- KDC
- LDAP
- DNS
- CA
- Amazon EC2
- S3 Storage

Amazon IAM

SDX Data Lake

- Ranger
- Atlas
- IDBroker
- Hive Metastore
- Cloudera Manager
- Amazon EC2
- Amazon RDS
- S3 Storage

KNOX

On-premise environment

Cloudera Manager

- Host Monitor
- Service Monitor
- Event Server
- Alert Publisher
- Reports Manager

Telemetry Publisher

CDH 5 / CDH 6 cluster

- HDFS
- Hive
- HMS
- Impala
- Sentry
- Spark

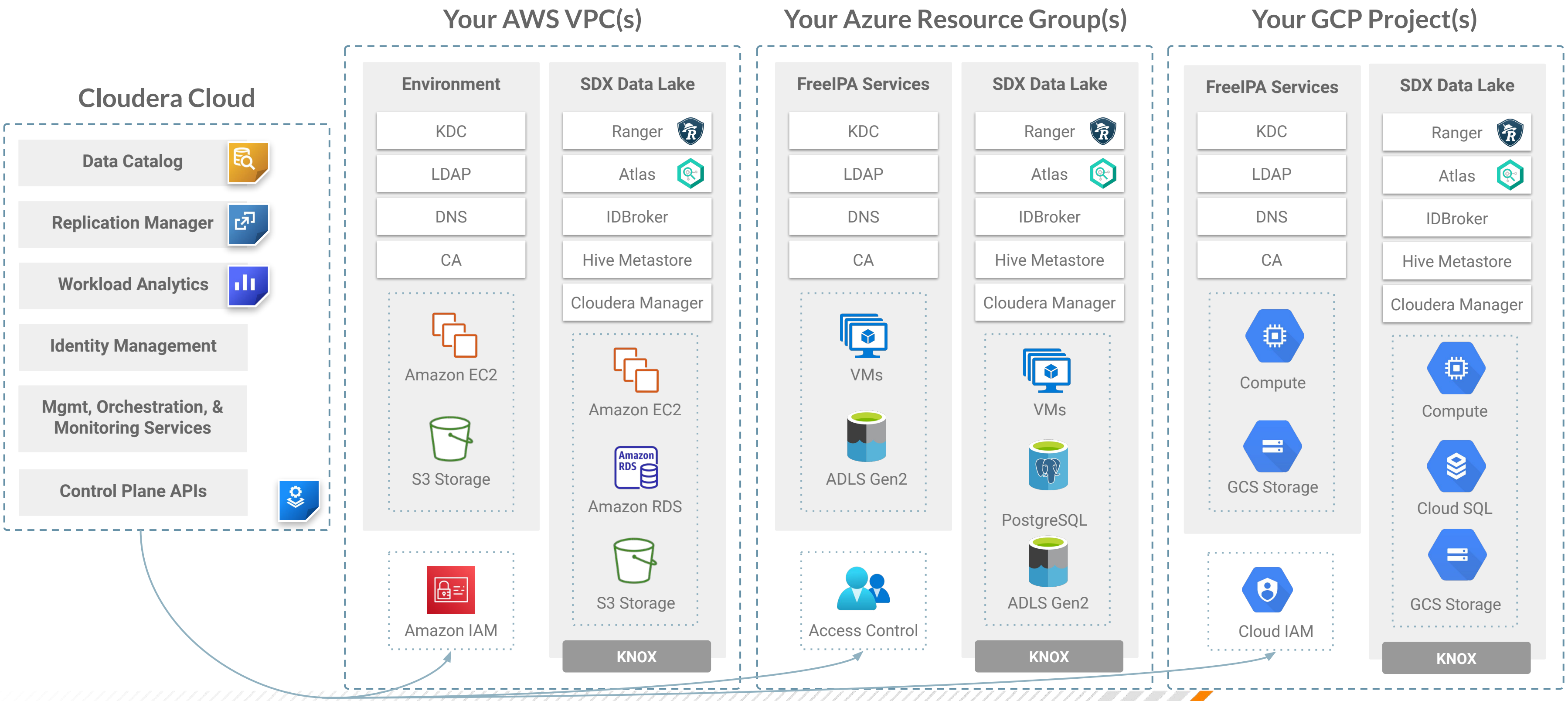
HDP 2 / HDP 3 cluster

- HDFS
- Hive
- HMS
- Impala
- Ranger
- Spark

CDP Private Cloud Base cluster

- HDFS
- Hive
- HMS
- Impala
- Ranger
- Spark

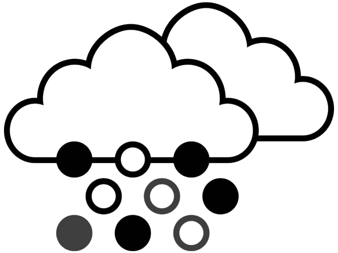
… ACROSS ANY CLOUDS



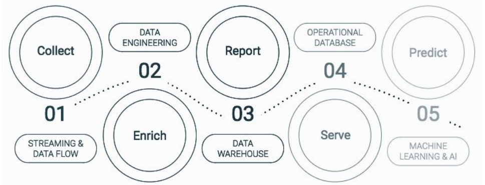
DEMO

CLOUDERA

THE ENTERPRISE DATA CLOUD COMPANY



Any Cloud



01 Collect — STREAMING & DATA FLOW

02 Enrich — DATA ENGINEERING

03 Report — DATA WAREHOUSE

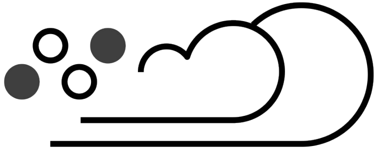
04 Serve — OPERATIONAL DATABASE

05 Predict — MACHINE LEARNING & AI

Data Lifecycle

CLOUDERA SDX

Secure & Governed



Open

THANK YOU

CLOUDERA