

CLOUDERA

WELCOME

CLOUD NATIVE  
DATA SERVICES

27<sup>th</sup> October 2021, 11 am IST

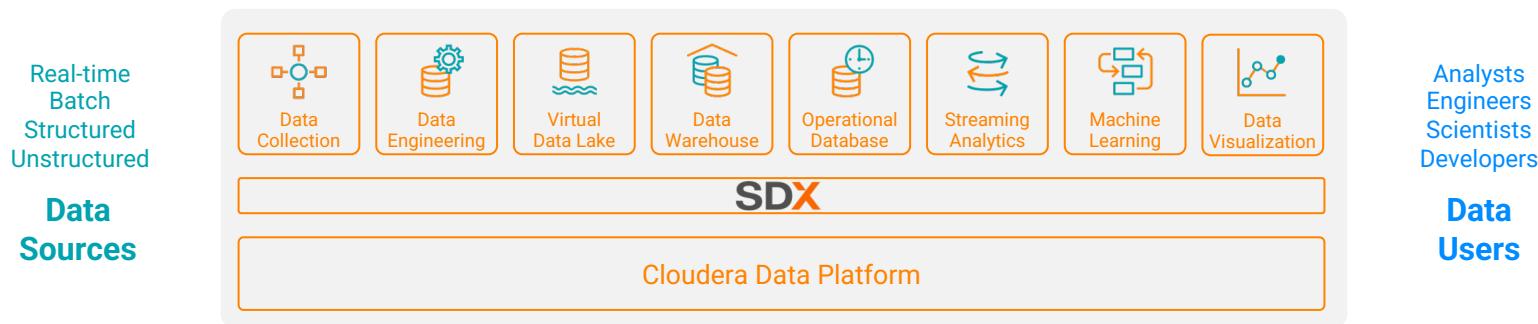
The Webinar will begin shortly



# CLOUD NATIVE DATA SERVICES

Vinay Rayker | Partner Technology Lead

# A HYBRID / MULTI-CLOUD DATA PLATFORM AND AN INTEGRATED SUITE OF SECURED DATA SERVICES



**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

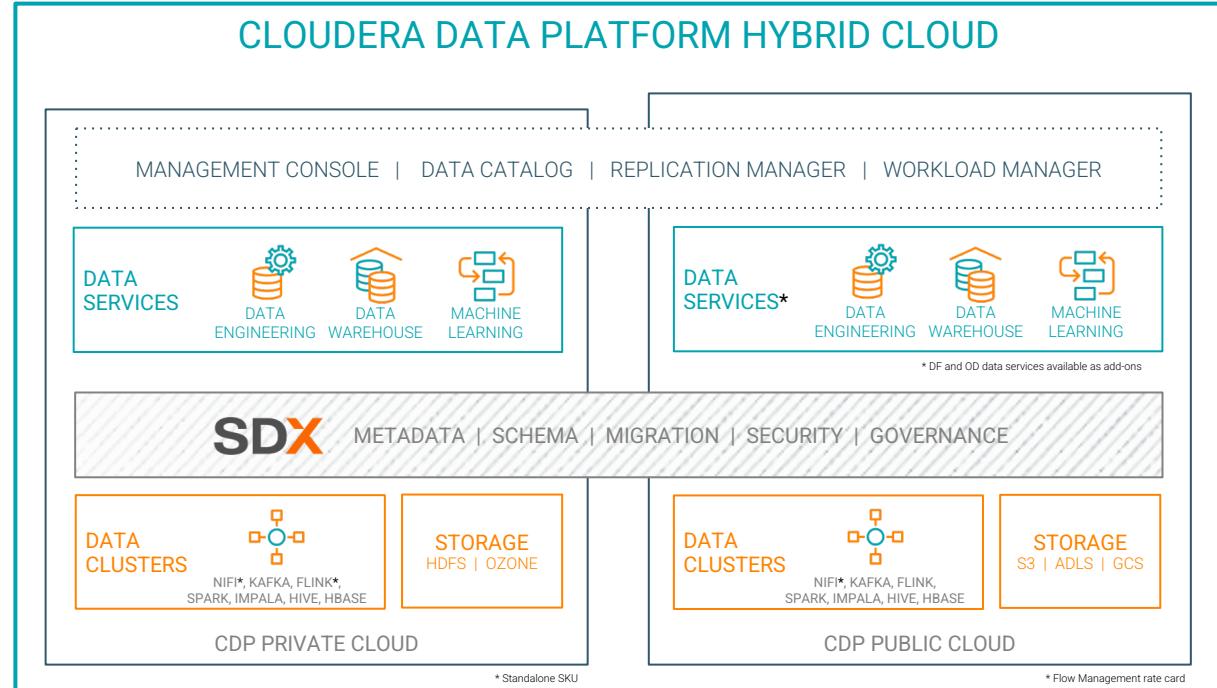


**Open & Extensible**  
to support more use cases faster and at lower cost

# CDP HYBRID CLOUD

Consistency and flexibility  
for data analytics across  
private & public clouds

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



## CLOUDERA DATA SERVICES

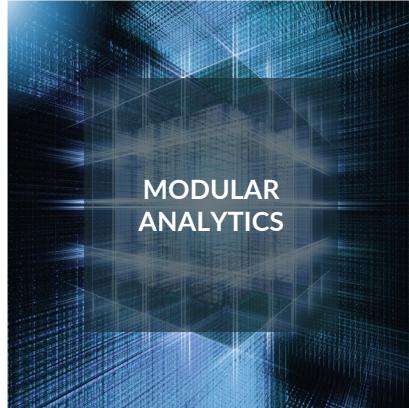
---

Data services are next generation, **practitioner-focused** analytic capabilities delivered as **modular cloud services**. Data services can be **standalone** offerings or combined into **integrated patterns** that deliver a seamless data lifecycle experience.

Cloudera Data Services share a **common management, security, and governance** framework and provide a consistent experience in any cloud.

# CLOUDERA DATA SERVICES

Build and run your data apps once and run them anywhere on a single data platform



- Designed for practitioners
- All-inclusive toolkits
- Platform integration

- DataFlow
- Data Engineering
- Data Warehouse
- Operational Database
- Machine Learning

- Private & public clouds
- Consistent user experience
- Common admin controls

- Built on SDX framework
- Simpler compliance
- End-to-end visibility

# CLOUDERA - THE HYBRID DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter



POWERED BY **CLOUDERA**  
**SDX**

Security | Governance | Lineage | Management | Automation

# DATA FLOW WITHIN THE DATA LIFECYCLE



POWERED BY **CLOUDERA**  
**SDX**

Security | Governance | Lineage | Management | Automation

# Cloudera DataFlow



## Flow Catalog

Reuse your existing NiFi flows by uploading them to the catalog

Keep track of your flow definitions and versions in a central catalog

Discover, search and reuse existing flows easily



## Flow Deployment

Allows easy flow deployment based on NiFi 1.13 across CDP environments (Dev, QA, Prod)

Define and assign KPIs to your flows

Automatic infrastructure scaling based on CPU utilization



## Flow Monitoring

Central monitoring console for all your flows across environments

Monitor flow metrics and infrastructure usage

Define alerts for flows breaching assigned KPIs

# Flow Catalog

- Central repository for flow definitions
- Import existing NiFi flows
- Manage flow definitions
- Initiate flow deployments

The screenshot shows the Cloudera DataFlow interface. On the left is a dark sidebar with the Cloudera logo and three navigation items: Dashboard, Catalog (which is selected and highlighted in blue), and Environments. The main area is titled "Flow Catalog". It displays a list of flow definitions with their names: Covid Data Stream, CovidIDBroker, drew\_kafka-hdfs-querydb-kudu, drew\_kafka\_to\_hdfs, Employees Data, Empty Dev Flow, and Generate Flow File Log. The "Covid Data Stream" item is currently selected, indicated by a blue border around its name. To the right of the list is a detailed view for the selected flow. This view includes a "FLOW DESCRIPTION" section with the text: "This flow reads covid data from several sources and writes it to CDP". Below this is a checkbox labeled "Only show deployed versions". A table follows, showing deployment information for different versions: Version 13 has 0 Deployments, Version 12 has 0 Deployments, and Version 11 has 0 Deployments. A "Deploy New Flow →" button is located above the table. At the bottom of the detailed view is a "LAST UPDATE" section showing "2021-02-02 14:25 PST by Michael Kohs" and the note "This version includes the latest fixes".

# ReadyFlows

- Cloudera provided flow definitions
- Cover most common data flow use cases
- Can be deployed and adjusted as needed
- Made available through docs during Tech Preview

Cloudera Docs / DataFlow master ▾ (test • Technical Preview) Search Document

**Cloudera DataFlow**

**Release Notes**  
Release Notes

**Concepts**  
Overview

**Planning**  
AWS Resource Planning  
NiFi Flow Limitations

**Getting Started**  
Quick Start  
Out of Box Flow Definitions  
Import a flow definition  
[Flow definition for ingesting data into a Kafka topic](#)  
Flow definition for ingesting data into Amazon S3 Buckets

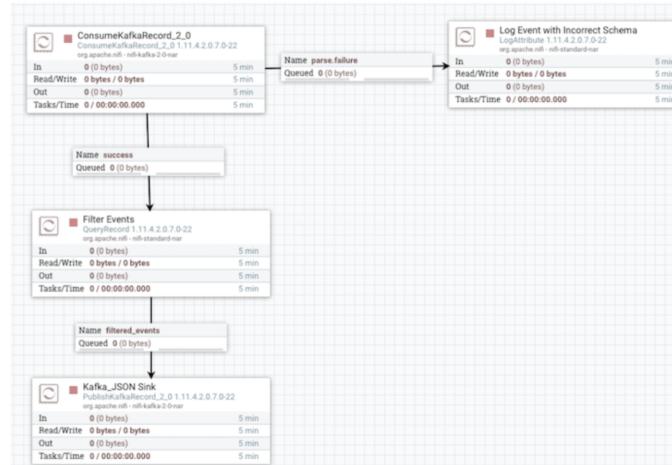
**How To: Environments**  
Enabling a DataFlow Environment  
Managing a DataFlow Environment

## OUT OF BOX FLOW DEFINITIONS

### Flow definition for ingesting data into a Kafka topic

#### Example

The resulting flow will look similar to the following, on your NiFi canvas.



# Deployment Wizard

- Turns flow definitions into flow deployments
- Guides users through providing required configuration
- Pick from pre-defined NiFi node sizes
- Define KPIs for the deployment

## Start Deployment Wizard

New Deployment X

Select the target environment

ⓘ Sensitive data never leaves the environment. Changing the environment after this step requires restarting the deployment process.

Selected Flow Definition

NAME	VERSION
Machine Data To Warehouse	2

Target Environment

aws dataflow-demo	60% (3 of 5)
-------------------	--------------

## Configure Sizing & Scaling

Overview  
Flow Parameters  
**Sizing & Scaling**  
Key Performance Indicators  
Review

**Sizing & Scaling**  
Select the NiFi node size and the number of nodes provisioned for your flow.

**NiFi Node Sizing**

<input checked="" type="radio"/> Extra Small	<input type="radio"/> Small	<input type="radio"/> Medium	<input type="radio"/> Large
2 vCores Per Node 4 GB Per Node	4 vCores Per Node 8 GB Per Node	8 vCores Per Node 16 GB Per Node	16 vCores Per Node 32 GB Per Node

**Number of NiFi Nodes**

Auto Scaling   
 Enabled

Min. Nodes: 1 - Max. Nodes: 3

## Provide Parameters

Flow Parameters

Data entered here never leaves the environment in your cloud account. Provide parameter values directly in the text input or upload a file for parameters that expect a file.

MachineData

AWS Credential File

Enter parameter values. Select File  
Drop file or browse

CDP Truststore

Enter parameter values. Select File  
Drop file or browse

CDPSchemaRegistry

https://dataflow-streams-master0.dataflow.xcu-2-8y8.dev.cldr.work:7790/api/v1

## Define KPIs

Key Performance Indicators

Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed.

Entire Flow METRIC TO TRACK: Data In ALERT SET: Notify if less than 150 KB/sec, for at least 30 seconds.
Processor: Write to S3 using HDFS proc METRIC TO TRACK: Bytes Sent ALERT SET: No alert set

+ Add New KPI

# Key Performance Indicators

- Visibility into flow deployments
- Track high level flow performance
- Track in-depth NiFi component metrics
- Defined in Deployment Wizard
- Monitoring & Alerts in Deployment Details

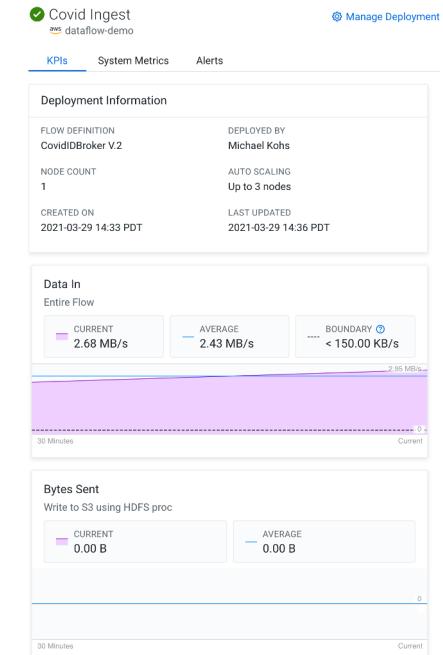
## KPI Definition in Deployment Wizard

The screenshot shows the 'Key Performance Indicators' step of a deployment wizard. On the left, a vertical navigation bar lists steps: Overview, Flow Parameters, Sizing & Scaling, Key Performance Indicators (which is the current step), and Review. The main area is titled 'Key Performance Indicators' and contains two cards:

- Entire Flow**
  - METRIC TO TRACK: Data In
  - ALERT SET: Notify if less than 150 KB/sec, for at least 30 seconds.
- Processor: Write to S3 using HDFS proc**
  - METRIC TO TRACK: Bytes Sent
  - ALERT SET: No alert set

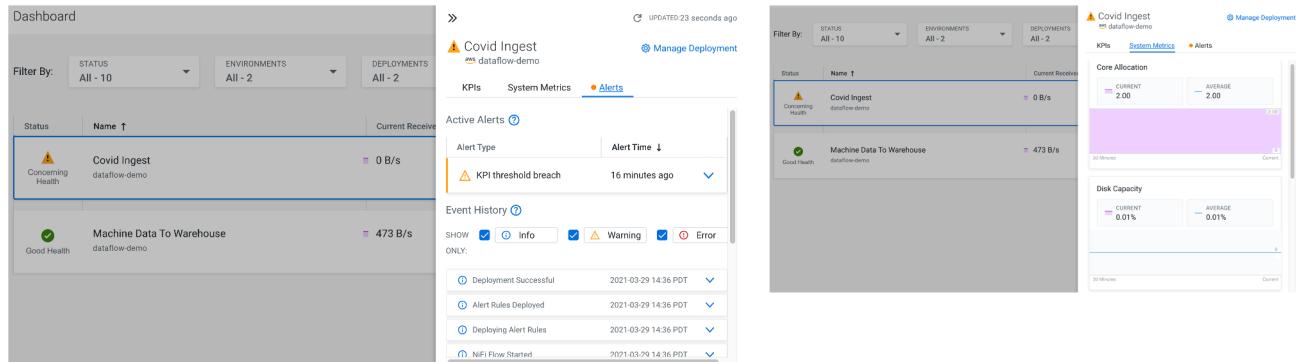
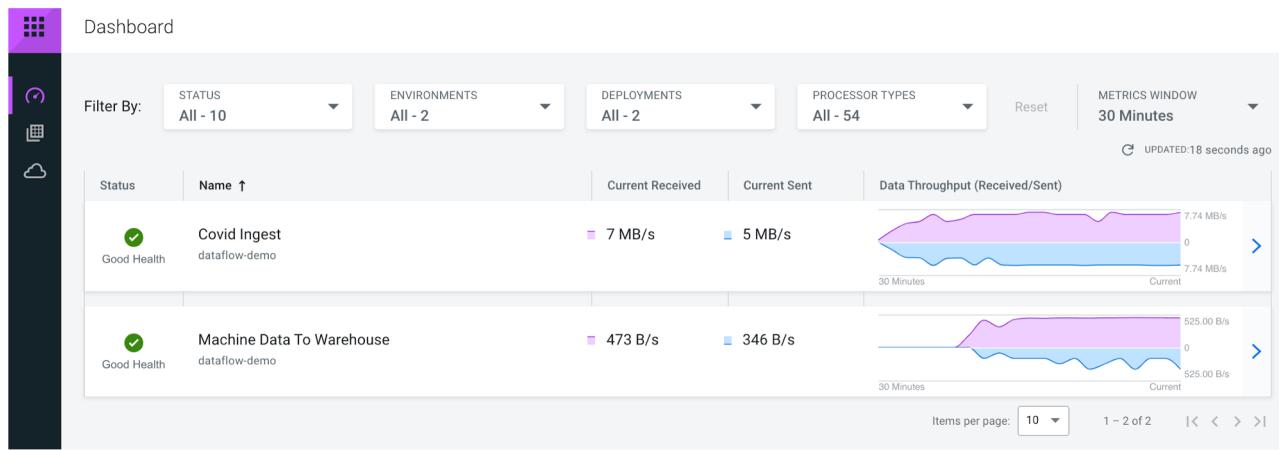
A blue 'Add New KPI' button is located at the bottom right of the card area.

## KPI Monitoring

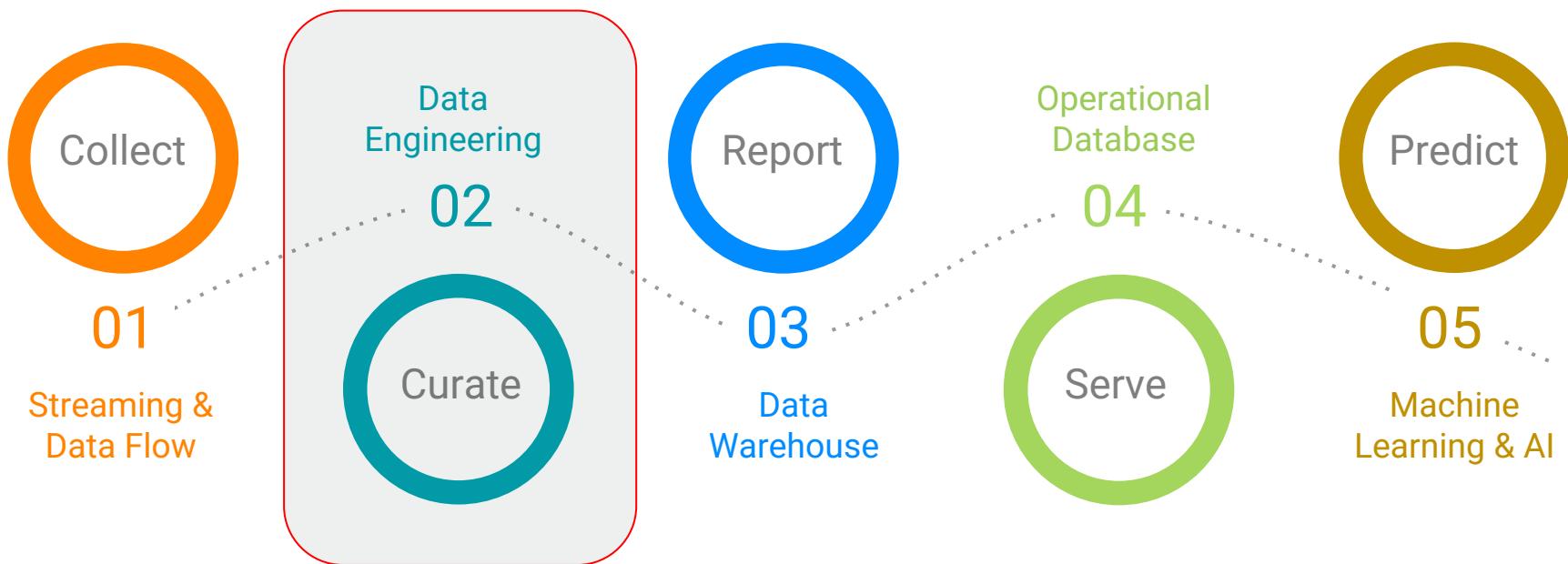


# Dashboard

- Central Monitoring View
- Monitors flow deployments across CDP environments
- Monitors flow deployment health & performance
- Drill into flow deployment to monitor system metrics and deployment events



# DATA ENGINEERING WITHIN THE DATA LIFECYCLE



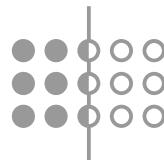
POWERED BY **CLOUDERA**  
**SDX**

Security | Governance | Lineage | Management | Automation

# THE CHALLENGES WITH TRADITIONAL DATA ENGINEERING



**Managing Spark  
Resources**



**Orchestrating Complex  
Pipelines**

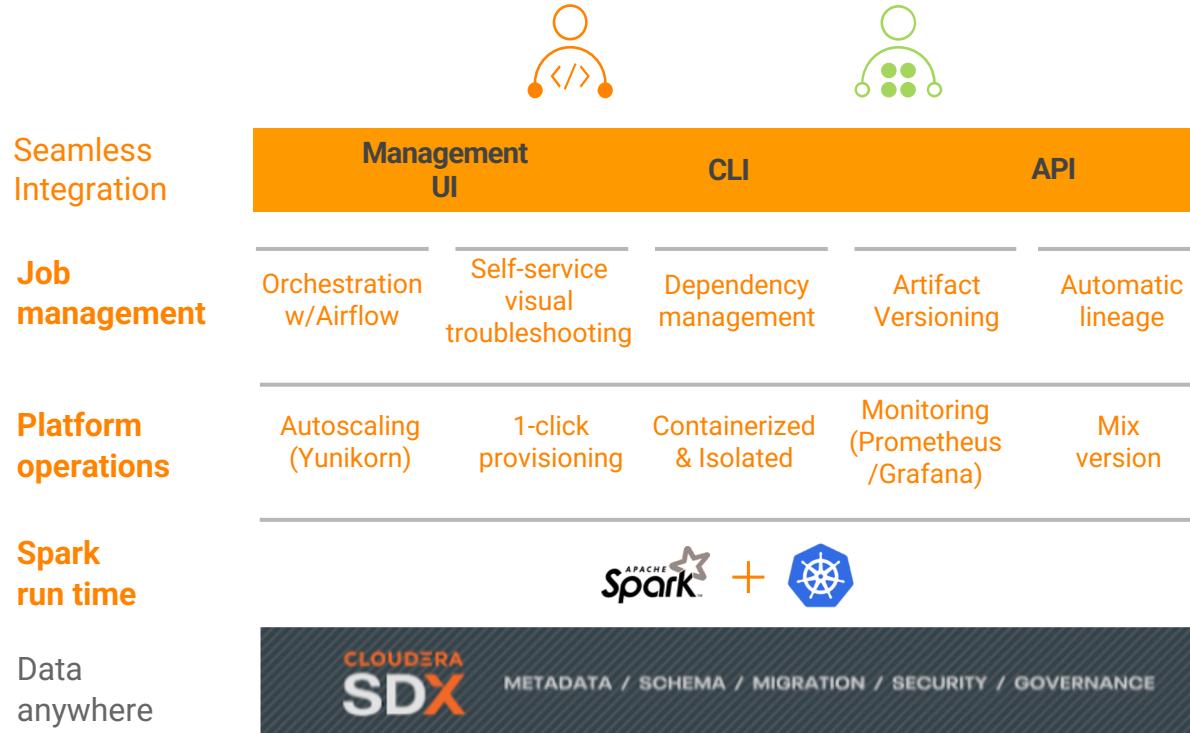


**Visibility &  
Troubleshooting**



**Secure & Fast  
Delivery**

# CLOUDERA DATA ENGINEERING



# OPTIMIZED FOR ENTERPRISE WORKFLOWS

## Managing Resources & Managing Jobs Across Teams



### Platform Admins

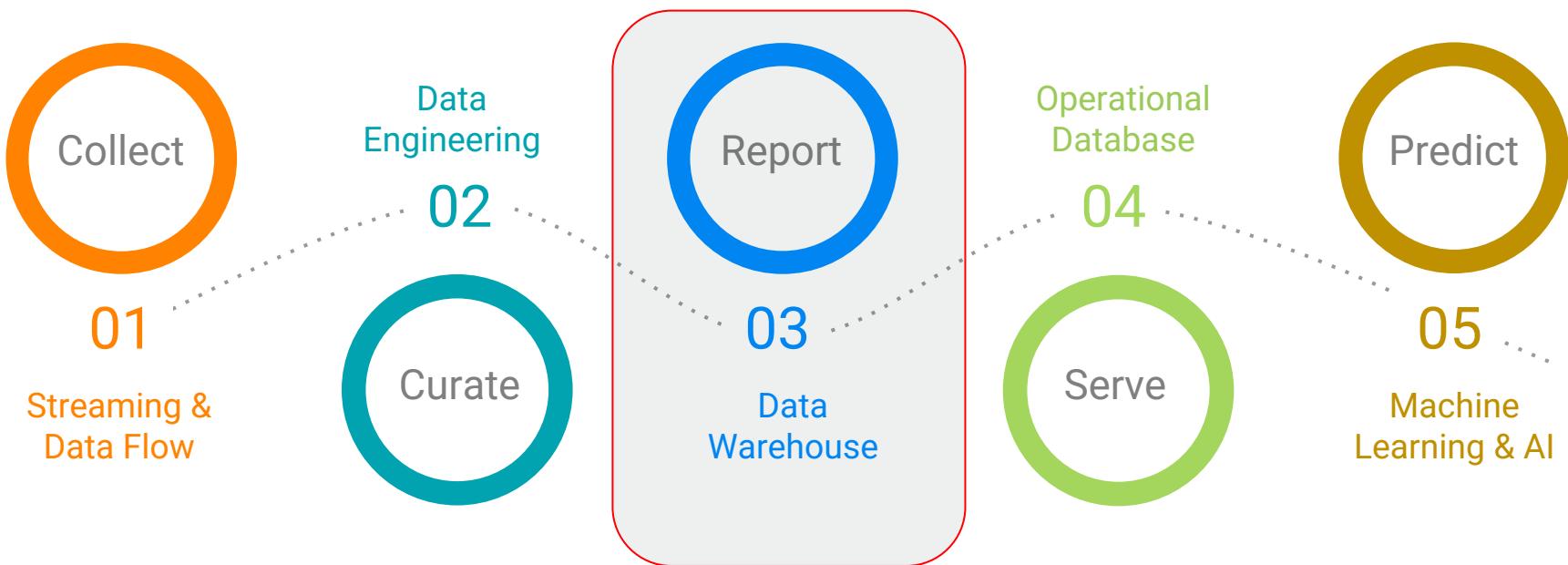
- Quickly provision new workloads
- Ensure isolation across LoB
- Control costs through on-demand autoscaling & resource guardrails
- Monitor resource usage over time
- Centralized Access controls & governance



### Data Engineers

- Easy, centralized deployment & monitoring of jobs
- Self-service troubleshooting with rich visual analysis
- Powerful workflow scheduling
- Automatic lineage capture
- Multiple versions of Spark
- Rich APIs for automation

# DATA WAREHOUSE WITHIN THE DATA LIFECYCLE



POWERED BY **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**

---

# Key CDW Features

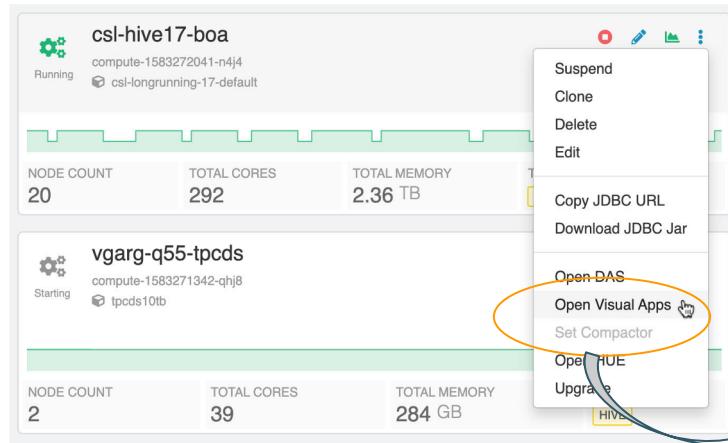
## What CDW Enables You to Do

- **Users:**
  - **Self-Service** via Virtual Warehouses
  - **Meet SLAs** via autoscale, workload isolation, and automated caching
  - **Use Tool of Choice**
- **Platform/IT:**
  - **Keep up with the Need & Speed of Business** via fast provisioning, autoscale
  - **Meet Analytics Requirements** via engines for reporting, exploration, dashboards, and time series
  - **Integration with Data Lifecycle** via zero data copy & integration with full data lifecycle
  - **Security & Governance** via integration with SDX
  - **Control Cost** via autosuspend and autostart

# COMPARISON OF SERVICE VS TEMPLATE DEPLOYMENT

	DW Service	DW Template
Installation	None	None
Provisioning / Deprovisioning	Automatic	Manual, Automate using CDP CLI
Sizing (in nodes)	Small (10), Medium (20), Large (40)	> 3
Node Type	Memory-Optimized, Preselected and tuned	Memory-Optimized, Customer Choice, Manual Tuning
Scaling	Automatic ( <i>cluster size increment</i> )	Manual, Automate using CDP CLI
Caching	Results, Data, File Handle, Warm Compute Nodes	Results, Data, File Handle
Suspend/Resume	Automatic	Manual, Automate using CDP CLI
Multiple Database Catalogs	Allowed	Not Allowed
Shared Data Experience	Inherited from CDP - Ranger, Knox, Atlas, HMS	Inherited from CDP - Ranger, Knox, Atlas, HMS

# DW Viz in CDW



CLOUDERA VIZ

HOME VISUALS DATA

search titles, viz types, datasets, authors...

All 2330

My Favorites 27

WORKSPACES + 8

Private 2322

Public

APPS Modify App Menu

- Sales App 3
- Truck Demo 2
- > Police Involved Inci... 3
- > Arcadia Training - ... 14
- > Event Log Analytics ... 2
- > Credit Card Analysis 2
- > Map Demos 12
- > Insurance - Custom... 4
- > Hospital - Surgery A... 3
- > Vulnerability App 5
- Pre-sales Apps 16
- > TM - Cyber Threat 4
- Hvatt 7

## My Favorites

YoY comparisons

FRT for Desk Regions - CVaR (Expected Shortfall)

Flight Overview Dashboard

Sales & Social summary

Life expectancy over time

Life expectancy in 1905

Truck demo application - violation report

Cereals by manufacturer

Word Cloud example

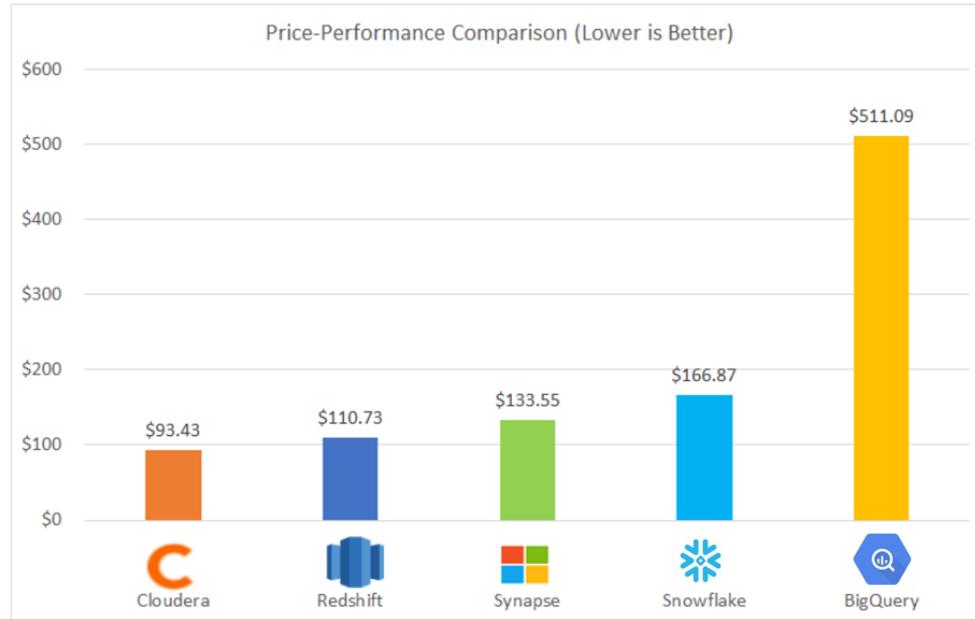
Rental Listing Analytics

evtlog single file analysis

evtlog analysis

# CDW - LEADER IN PRICE-PERFORMANCE COMPARISON

Comprehensive & Independent benchmark

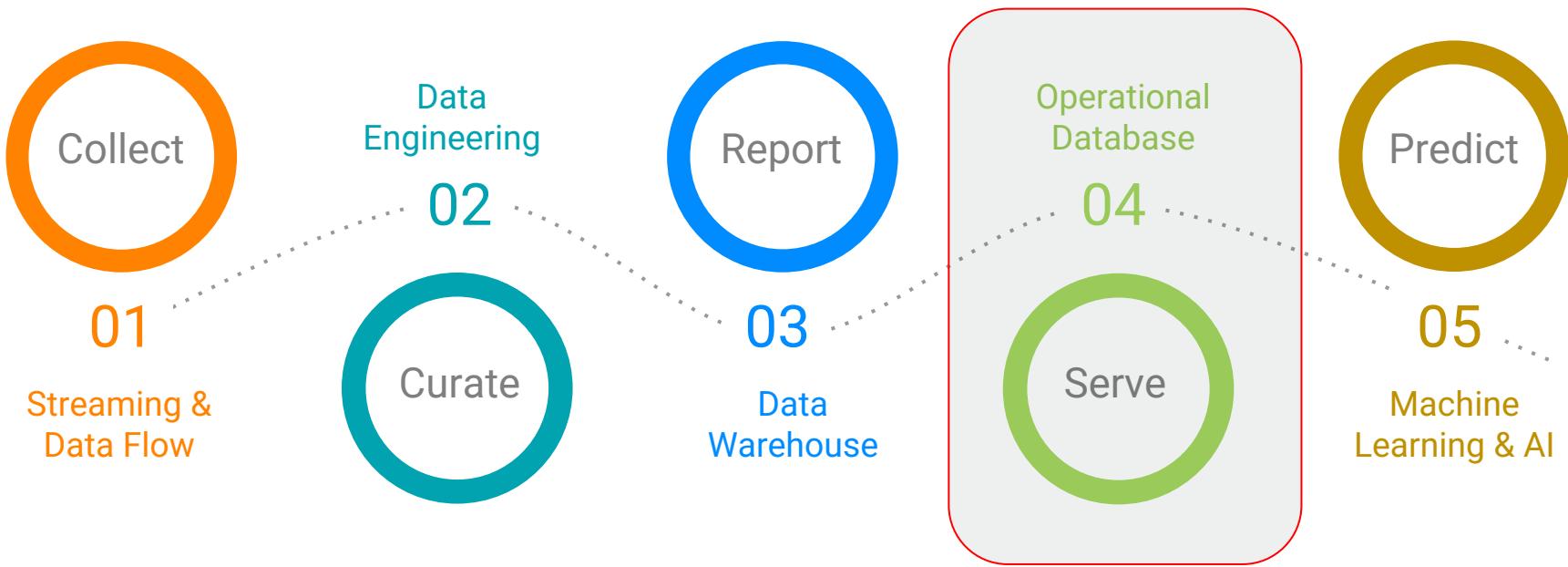


SOURCE: [Cloud Data Warehouse Performance Testing By William McKnight, January 2021](#)

## Cloudera Data Warehouse

- 20% cheaper than Redshift
- 40% cheaper than Synapse
- 80% cheaper than Snowflake
- 5.5x cheaper than BigQuery

# OPERATIONAL DATABASE WITHIN THE DATA LIFECYCLE



POWERED BY **CLOUDERA**  
**SDX**

Security | Governance | Lineage | Management | Automation

# WHAT TO EXPECT IN CDP PUBLIC CLOUD

Allow developers to spend time where it matters

**Easy and quick deployment for developers**



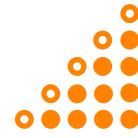
3 Clicks



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*

# 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

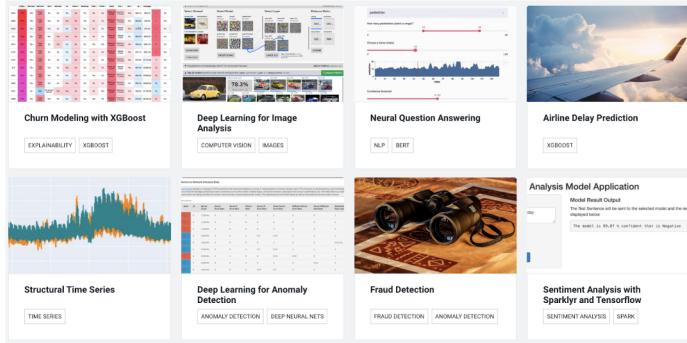
# MACHINE LEARNING WITHIN THE DATA LIFECYCLE



POWERED BY **CLOUDERA**  
**SDX**

Security | Governance | Lineage | Management | Automation

# ML Across the Lifecycle



Quickly Stand Up ML Applications



From Lab to AI Factory

While Governed and Secure

## ENTERPRISE REQUIREMENTS

**CLOUDERA**  
**SDX**

MODEL SECURITY

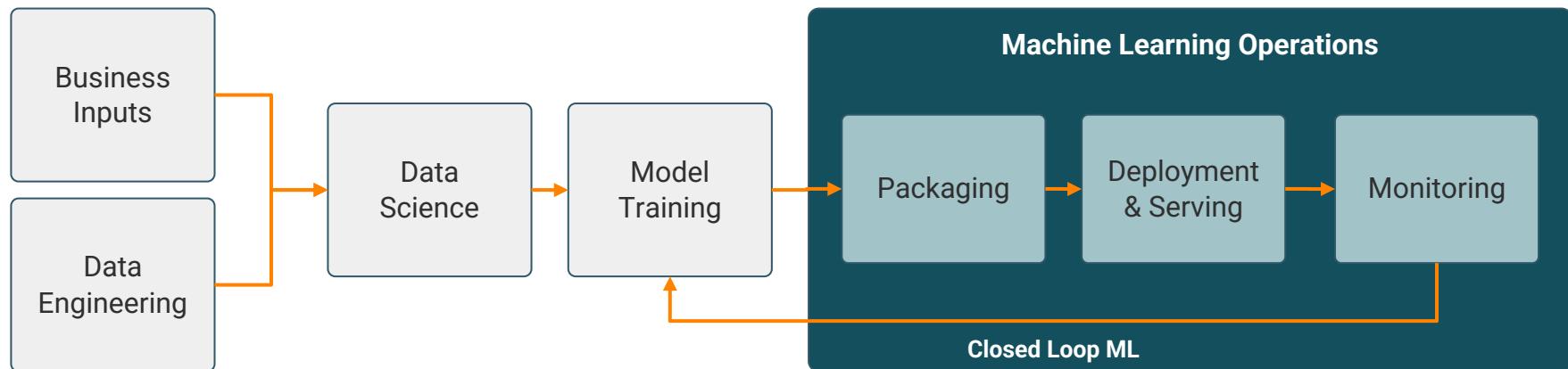
MODEL  
GOVERNANCE

MODEL CATALOG

FEATURE STORE

DATA CATALOG

# MACHINE LEARNING IN PRODUCTION





# Integrated Business Intelligence

## Native Data Visualizations in CML

The image shows two screenshots of the Cloudera Machine Learning (CML) interface. On the left, the 'Applications' section of the CML dashboard is displayed, featuring a sidebar with 'Projects' and various ML components like Sessions, Experiments, Models, Jobs, Applications, and Admin. The 'Applications' item is highlighted with a yellow oval and a large orange curved arrow pointing towards the right screenshot. The right screenshot shows the 'Visual Apps' interface, which includes a search bar, navigation tabs for HOME, VISUALS, and DATA, and a list of visualizations such as 'Example analysis', 'Churn Demo App', 'Interactive & Dynamic dashboards', and 'Animated world population - GDP vs life'. A small number '1' is visible near the bottom right of the CML logo in the top right corner of the right screenshot.

CLOUDERA Machine Learning

Projects

- Sessions
- Experiments
- Models
- Jobs
- Applications**
- Settings
- Admin

Flight Data Analy

By Priyank Patel. Last worked on

CLOUDERA Visual Apps

HOME VISUALS DATA

Find titles, viz types, datasets, authors...

priyankp / Projects

All 23

My Favorites 0

WORKSPACES 1

Public

Private

0 sessions running

All

DASHBOARDS VISUALS APPS ALL

- Example analysis
- Churn Demo App
- Interactive & Dynamic dashboards
- Animated world population - GDP vs life
- Mapping customers
- SKO'21 ML Demo - IRA
- SKO - IRA bank analysis
- Deficiency Details: <<county:Queens>>

© 2021 Cloudera, Inc. All rights reserved. 33

# KEY DIFFERENTIATION

Cloudera Data Platform

HYBRID & MULTI

vs

CLOUD ONLY

DATA LIFECYCLE

vs

SINGLE FUNCTION

SECURE & GOVERNED

vs

FRAGMENTED SECURITY

OPEN SOURCE

vs

PROPRIETARY LOCK-IN

PLATFORM

vs

POINT SOLUTION

# THANK YOU

CLOUDERA