CLOUDERA

WELCOME

CLOUD NATIVE DATA SERVICES

27th October 2021, 11 am IST

The Webinar will begin shortly

CLOUDERA

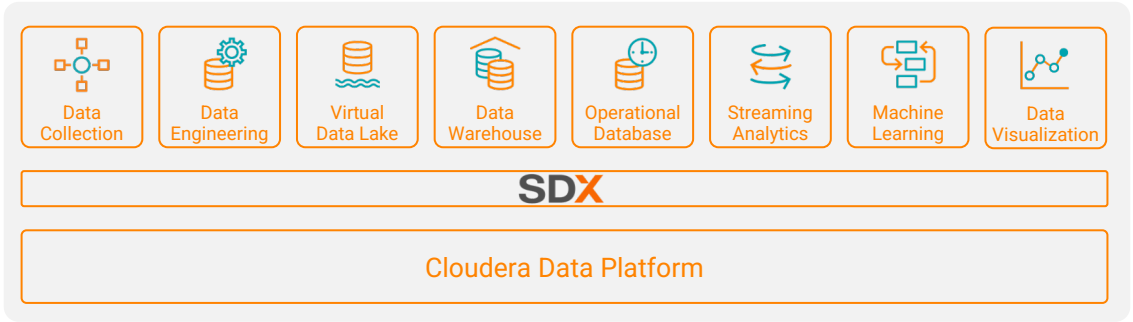
CLOUD NATIVE DATA SERVICES

Vinay Rayker| Partner Technology Lead

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

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

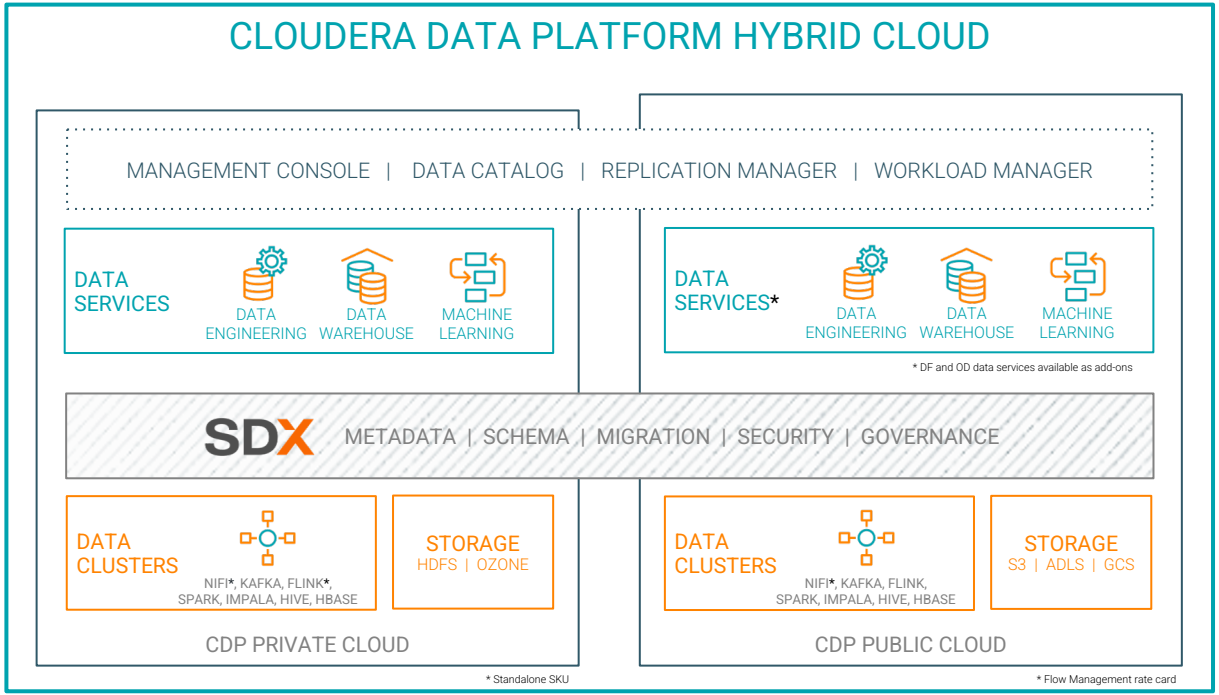
Secure & Governed
to simplify data protection, sharing and compliance

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

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

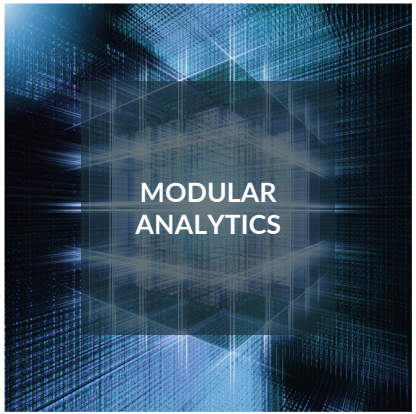
Data services are next generation, **practitioner-focused** analytic capabilities delivered as **modular cloud services**. Data services can be **standalone** offerings or combined into **integrated patterns** that deliver a seamless data lifecycle experience.

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



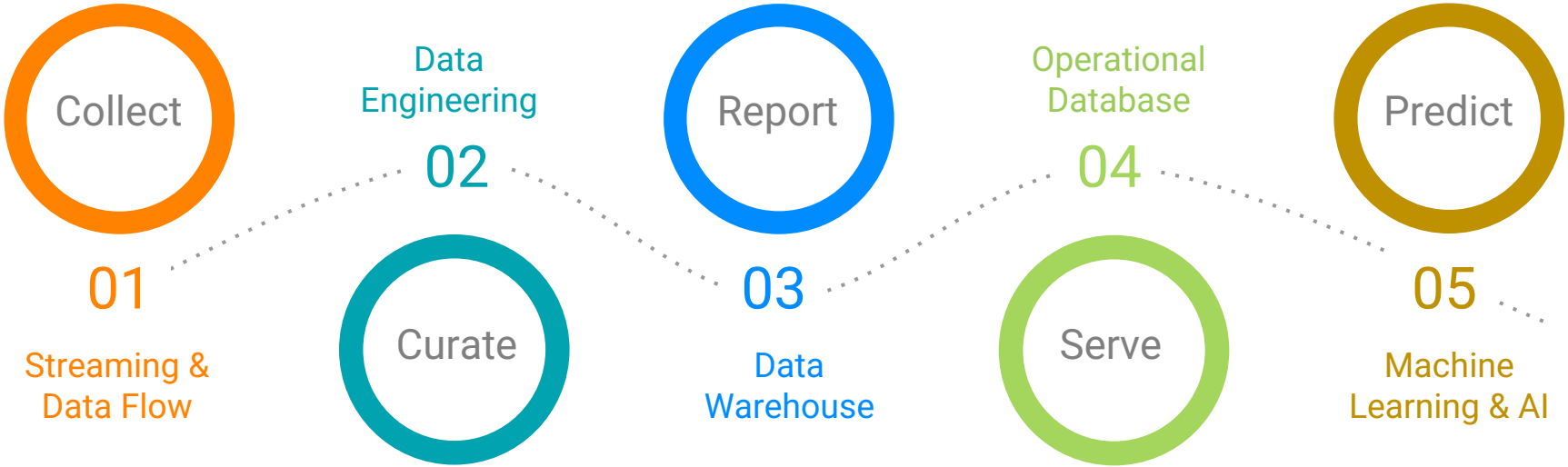
- DataFlow
- Data Engineering
- Data Warehouse
- Operational Database
- Machine Learning

- 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

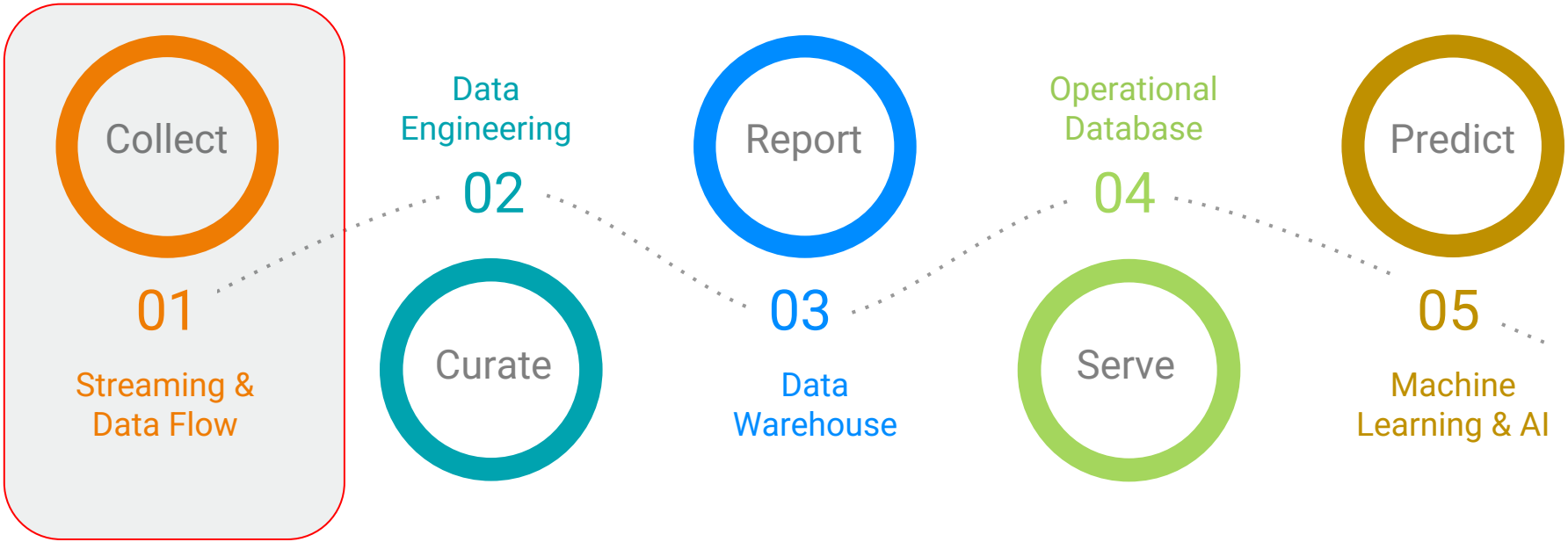
Manage and secure the data lifecycle in any cloud or datacenter



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

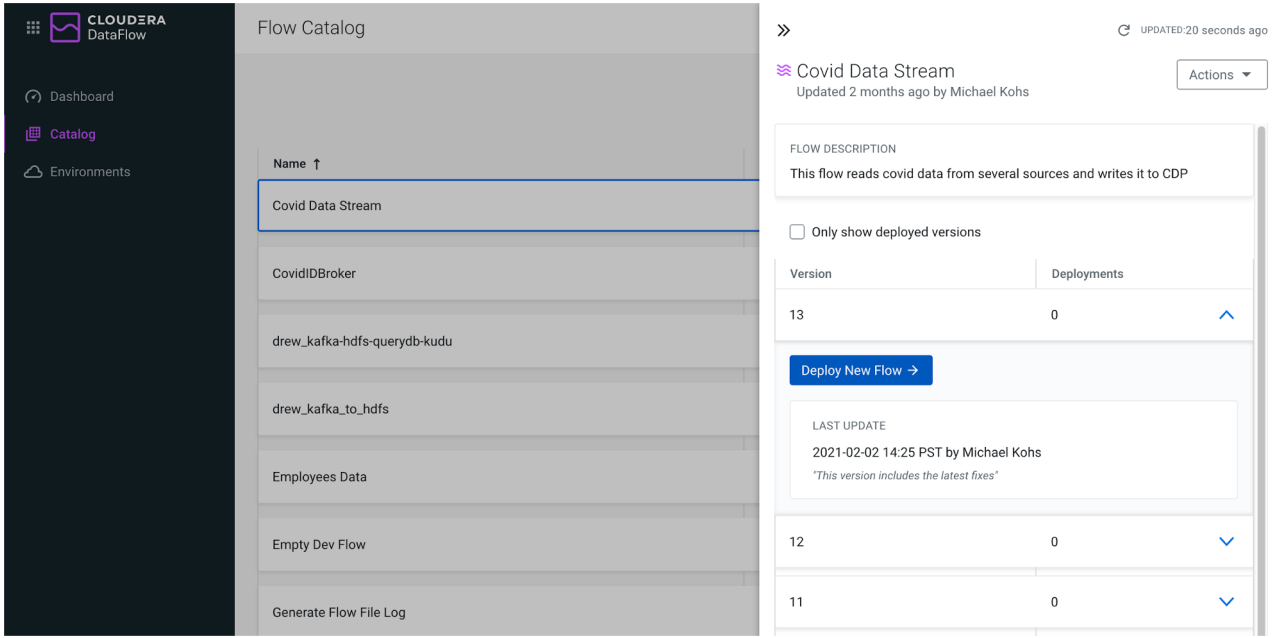
Central monitoring console for all your flows across environments

Monitor flow metrics and infrastructure usage

Define alerts for flows breaching assigned KPIs

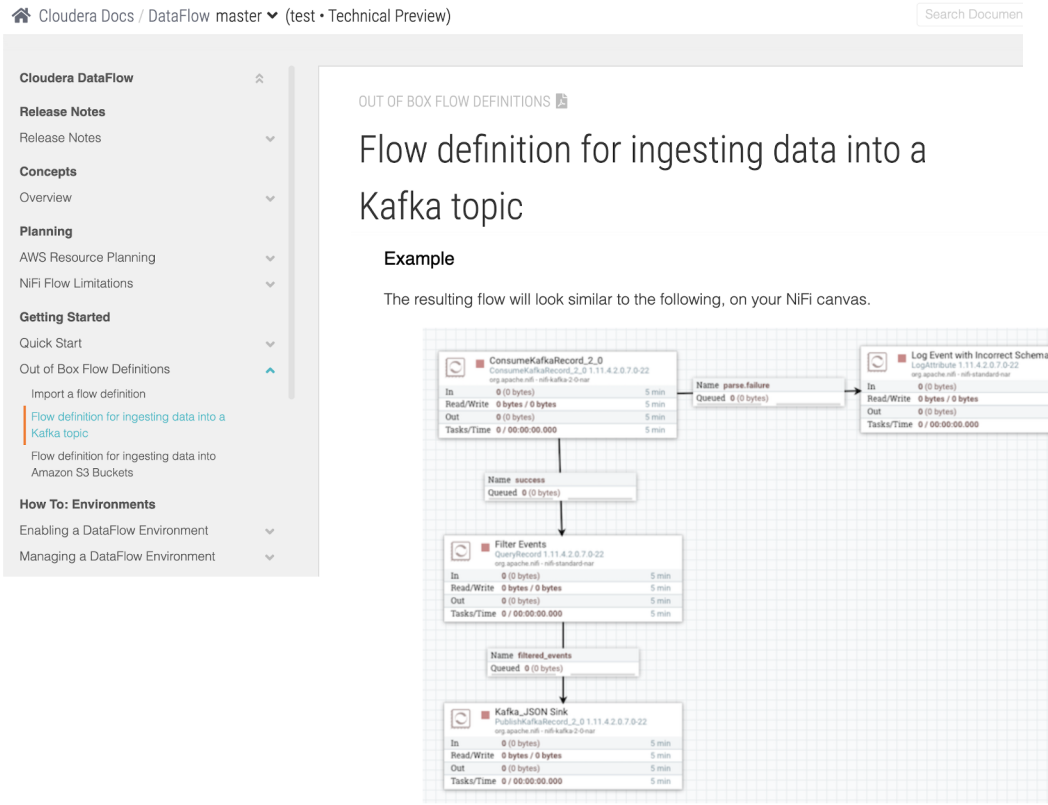
Flow Catalog

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



ReadyFlows

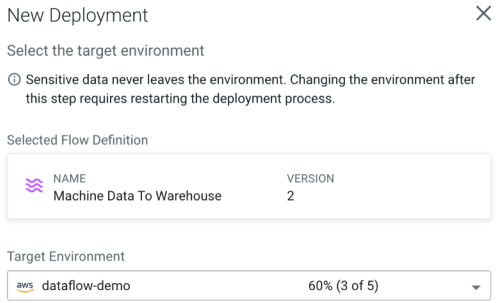
- Cloudera provided flow definitions
- Cover most common data flow use cases
- Can be deployed and adjusted as needed
- Made available through docs during Tech Preview



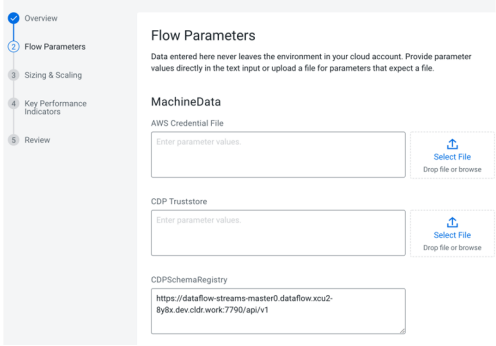
Deployment Wizard

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

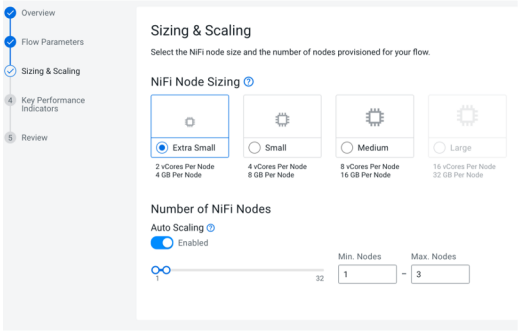
Start Deployment Wizard



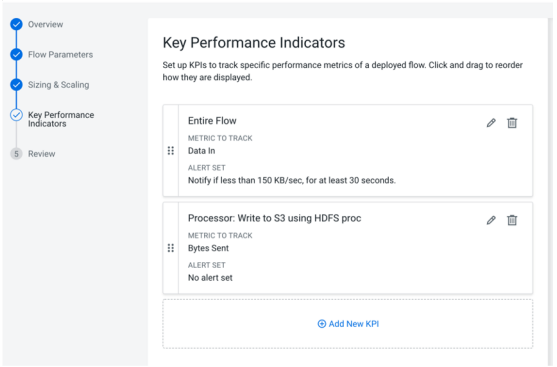
Provide Parameters



Configure Sizing & Scaling



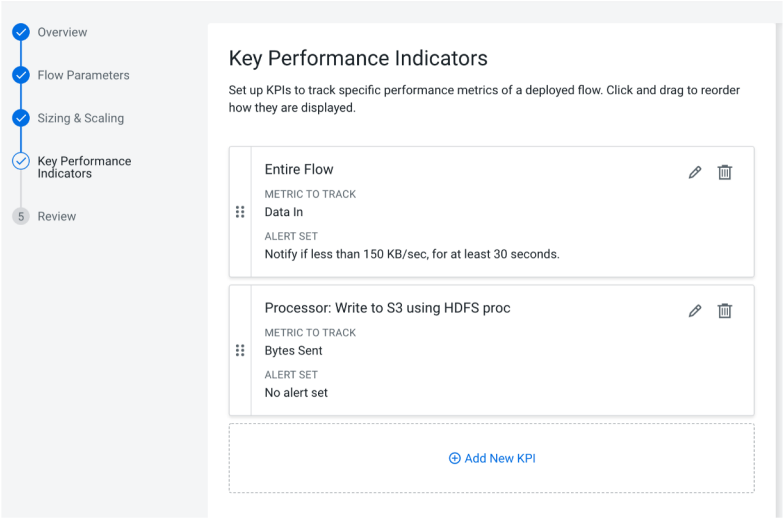
Define KPIs



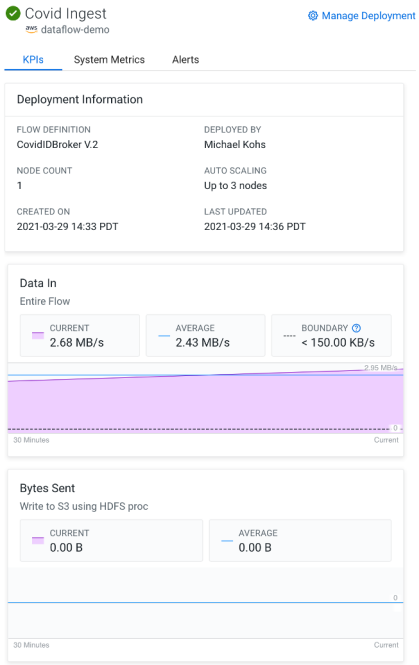
Key Performance Indicators

- 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

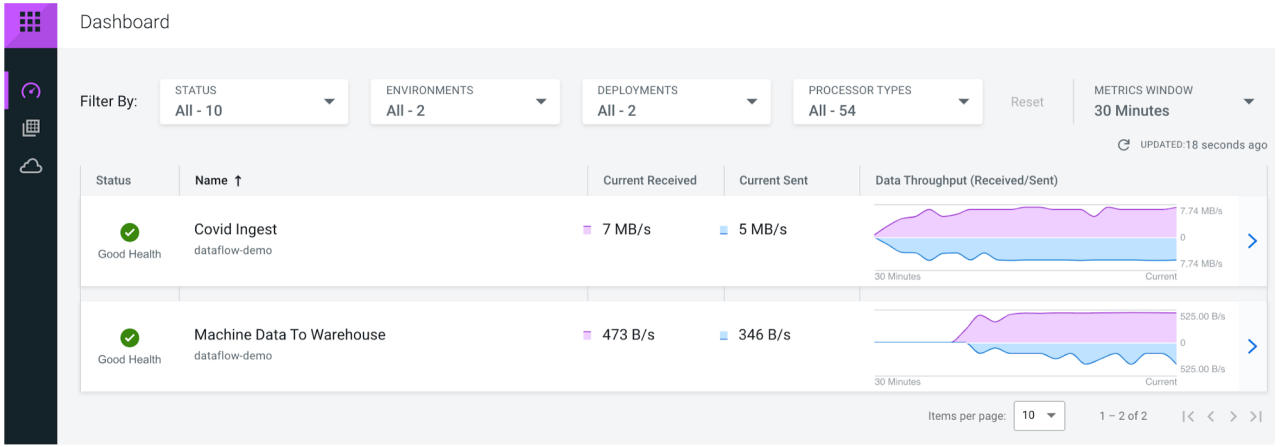


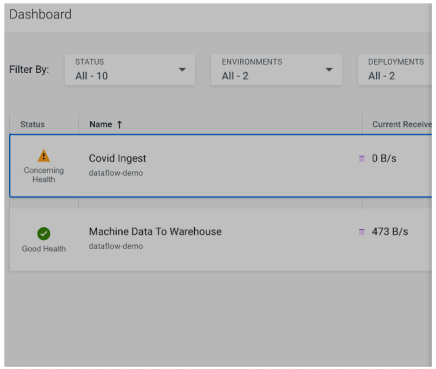
KPI Monitoring

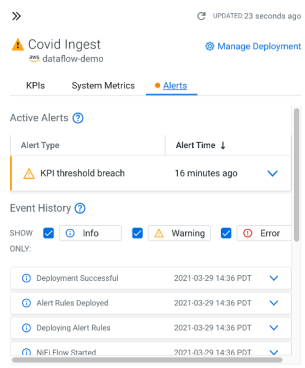


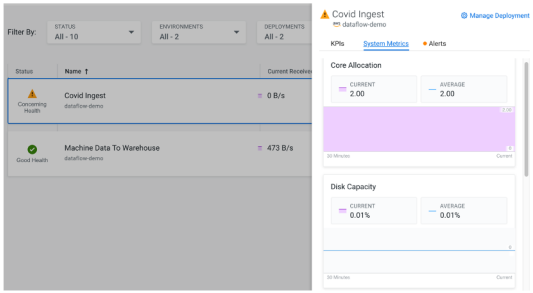
Dashboard

- 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

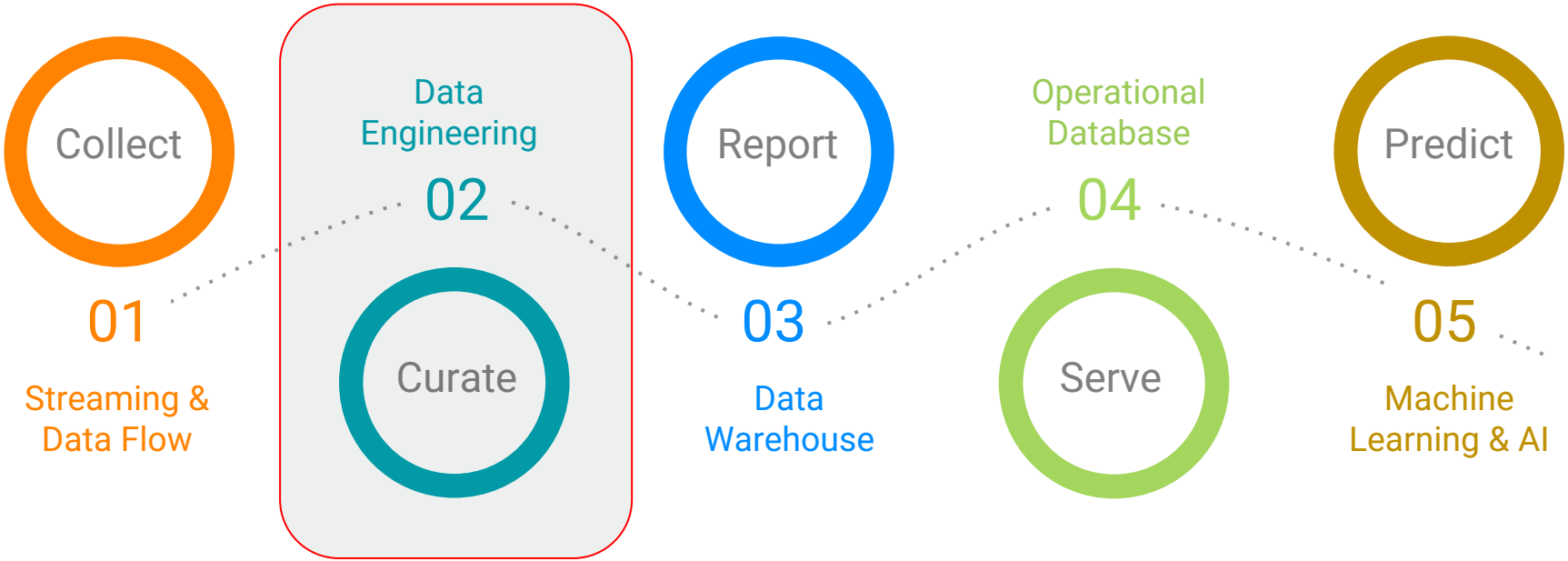








DATA ENGINEERING WITHIN THE DATA LIFECYCLE

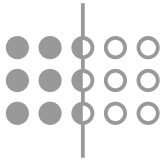


POWERED BY CLOUDERA SDX

Security | Governance | Lineage | Management | Automation

THE CHALLENGES WITH TRADITIONAL DATA ENGINEERING

Managing Spark Resources

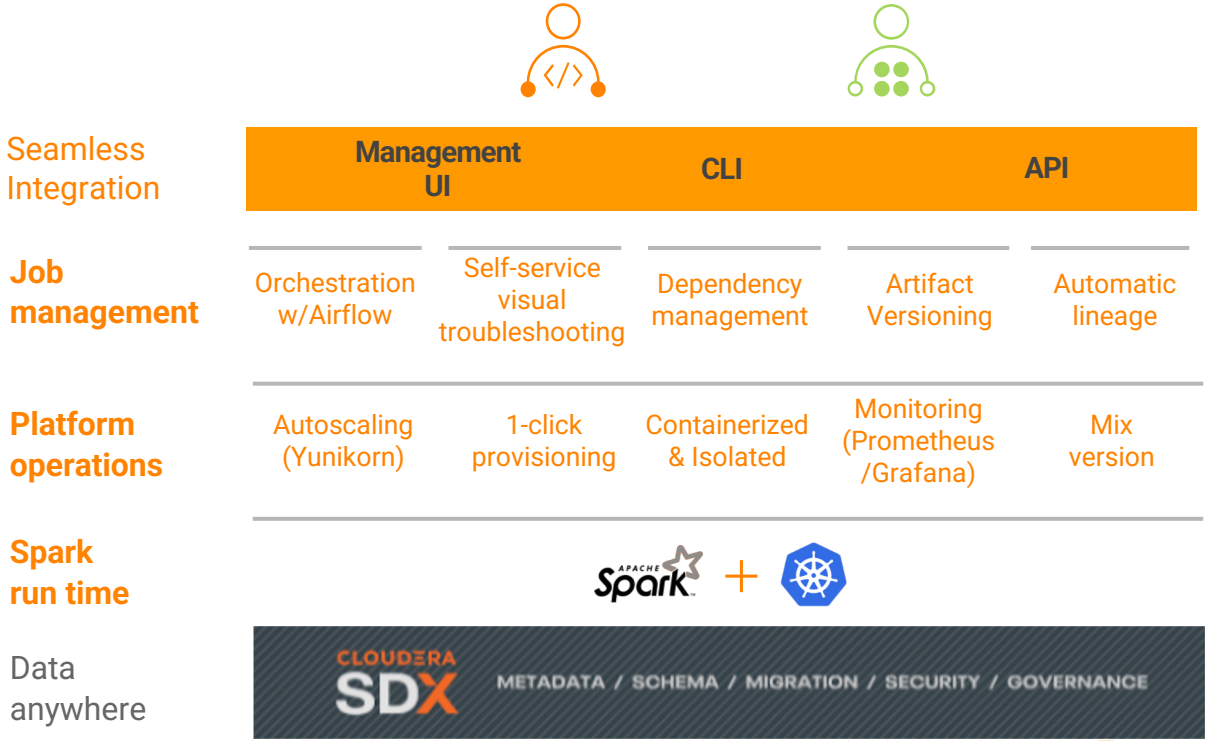


Orchestrating Complex Pipelines

Visibility & Troubleshooting

Secure & Fast Delivery

CLOUDERA DATA ENGINEERING



OPTIMIZED FOR ENTERPRISE WORKFLOWS

Managing Resources & Managing Jobs Across Teams

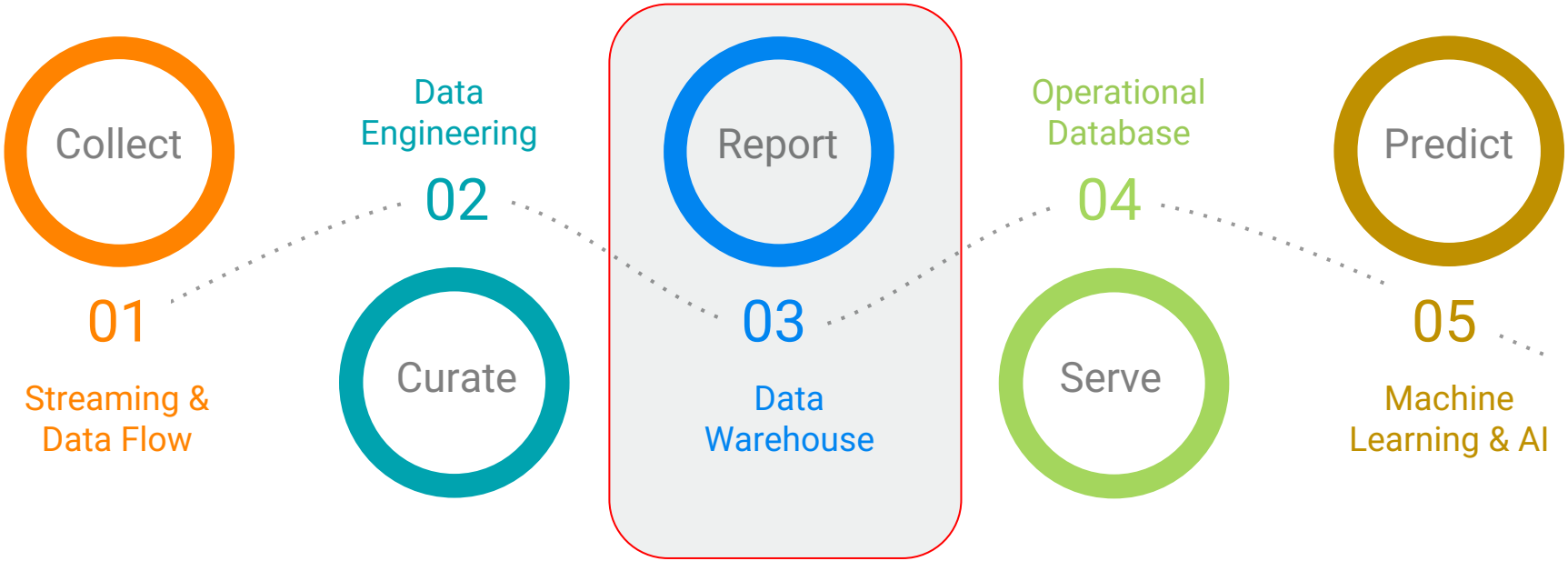
Platform Admins

- 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

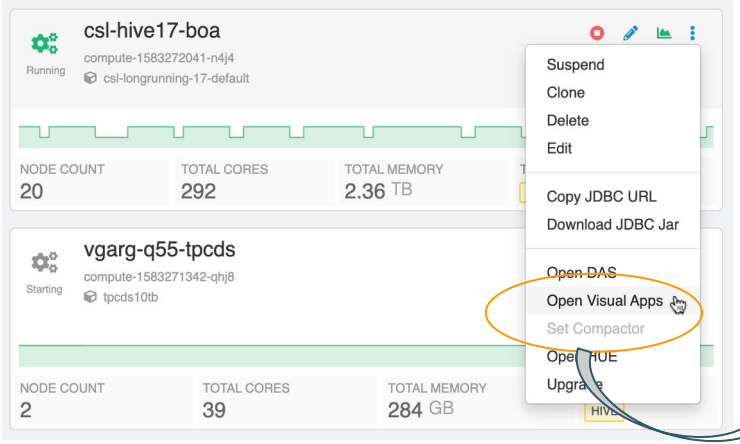
- **Users:**
  - **Self-Service** via Virtual Warehouses
  - **Meet SLAs** via autoscale, workload isolation, and automated caching
  - **Use Tool of Choice**

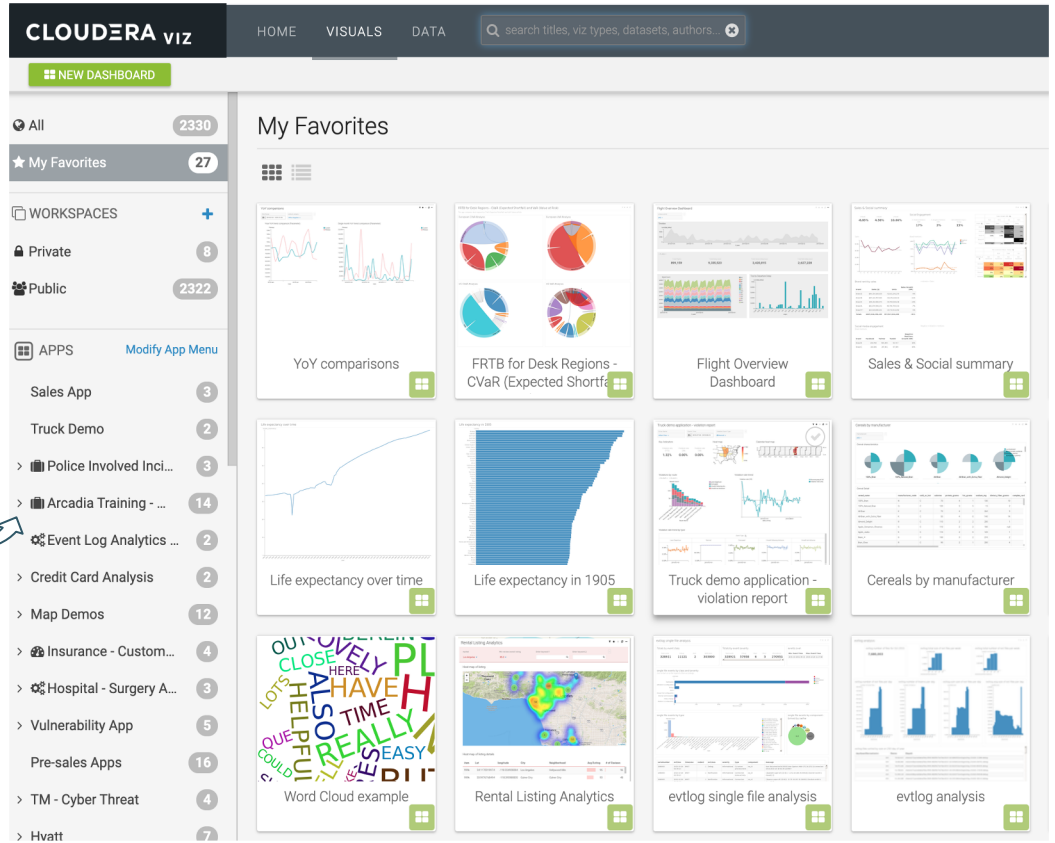
- **Platform/IT:**
  - **Keep up with the Need & Speed of Business** via fast provisioning, autoscale
  - **Meet Analytics Requirements** via engines for reporting, exploration, dashboards, and time series
  - **Integration with Data Lifecycle** via zero data copy & integration with full data lifecycle
  - **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 (cluster size increment)*	*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

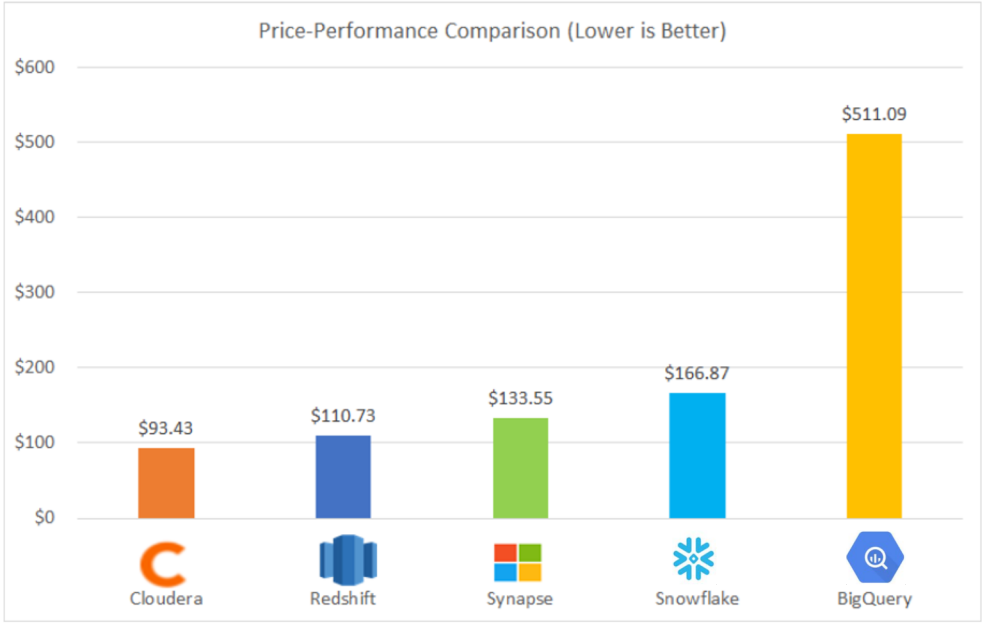
DW Viz in CDW





CDW - LEADER IN PRICE-PERFORMANCE COMPARISON

Comprehensive & Independent benchmark

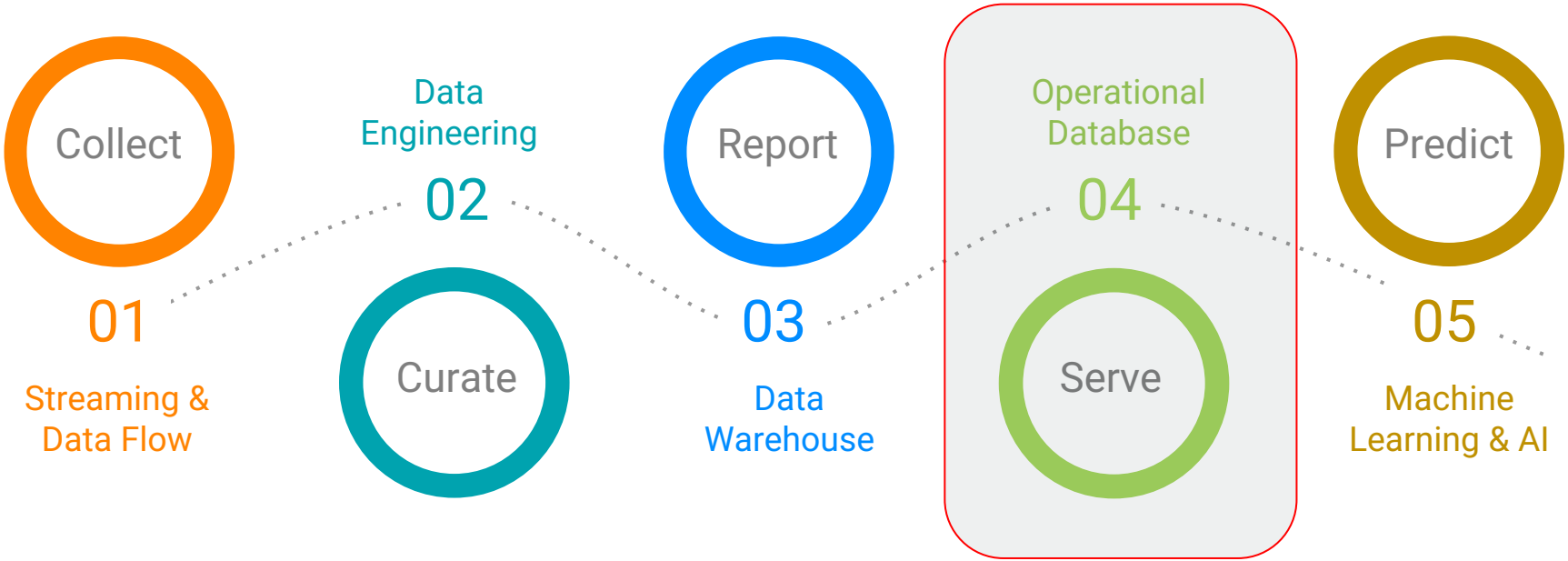


Cloudera Data Warehouse

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

SOURCE: *Cloud Data Warehouse Performance Testing By William Mcknight, January 2021*

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

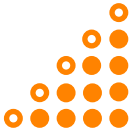
Easy and quick deployment for developers

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

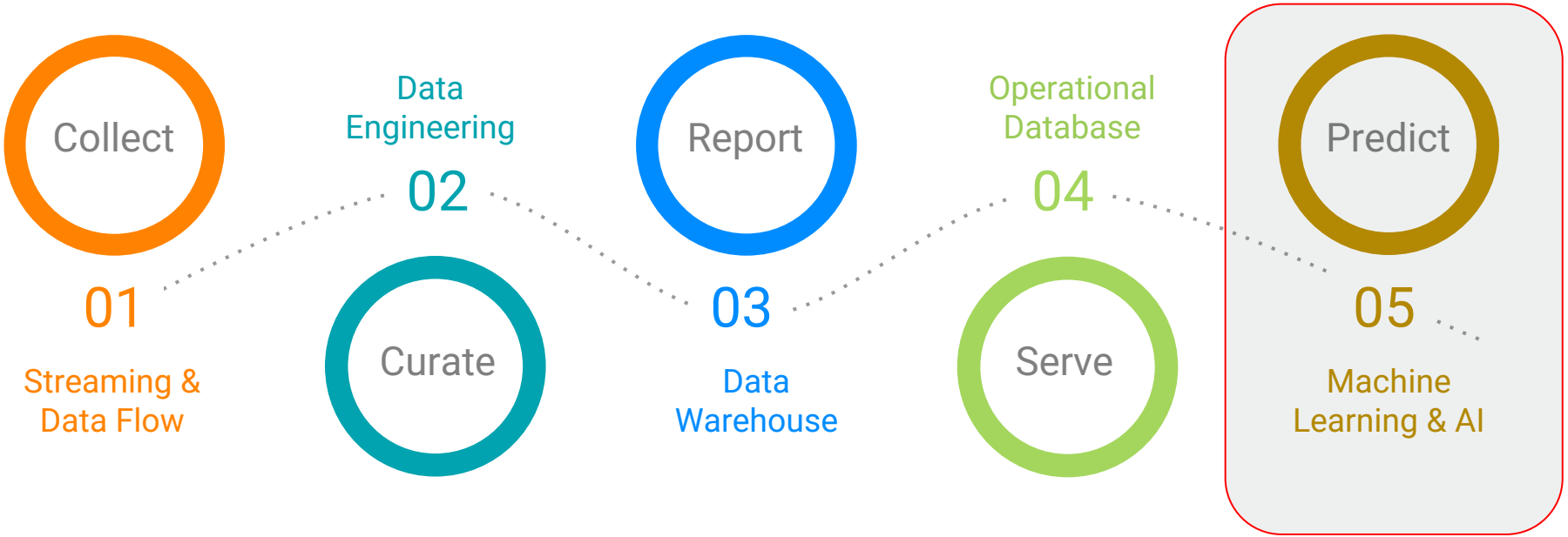
Auto-tuning

- Hot spotting
- Space management

Eliminate tuning

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

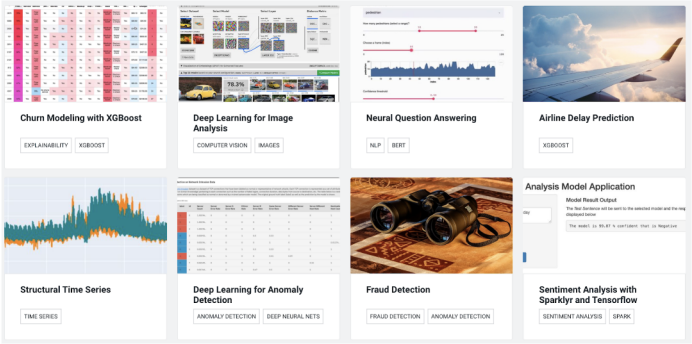
MACHINE LEARNING WITHIN THE DATA LIFECYCLE



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

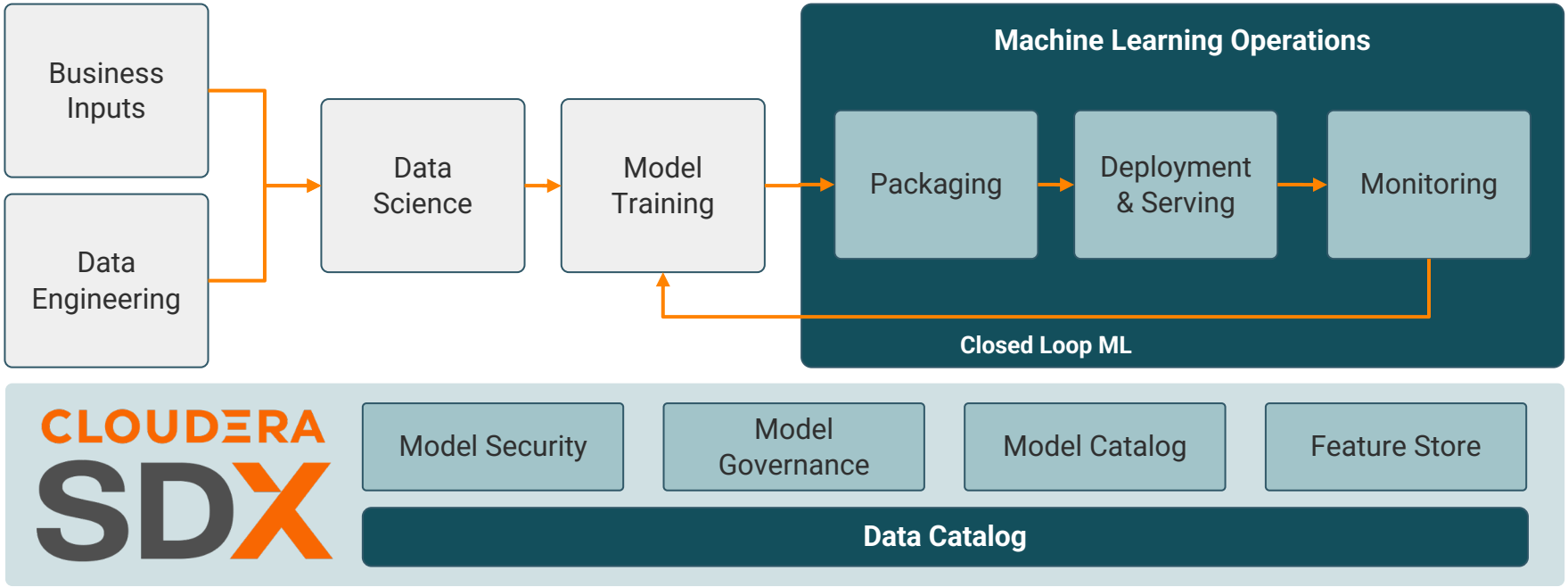
ENTERPRISE REQUIREMENTS

CLOUDERA SDX

MODEL SECURITY | MODEL GOVERNANCE | MODEL CATALOG | FEATURE STORE

DATA CATALOG

MACHINE LEARNING IN PRODUCTION

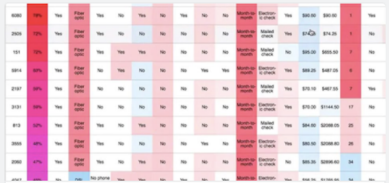


Applied ML Prototypes Catalog

Projects
Sessions
Experiments
Models
Jobs
Applications
Settings
Prototype Catalog
Admin

Search prototypes

Tags: all

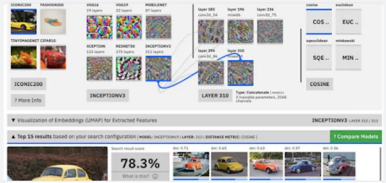


Churn Modeling with XGBoost

EXPLAINABILITY, XGBOOST

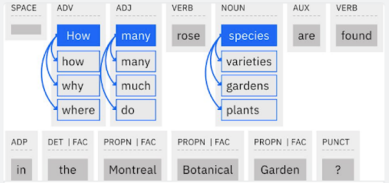
Airline Delay Prediction

XGBOOST



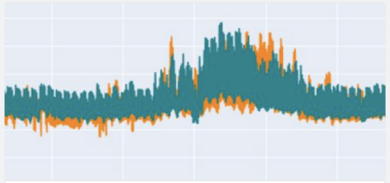
Deep Learning for Image Analysis

COMPUTER VISION IMAGES



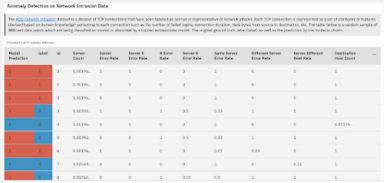
Neural Question Answering

NLP



Structural Time Series

TIME SERIES

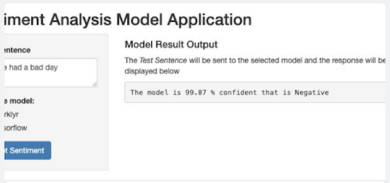


Deep Learning for Anomaly Detection

ANOMALY DETECTION DEEP NEURAL NETS

Fraud Detection

FRAUD DETECTION ANOMALY DETECTION



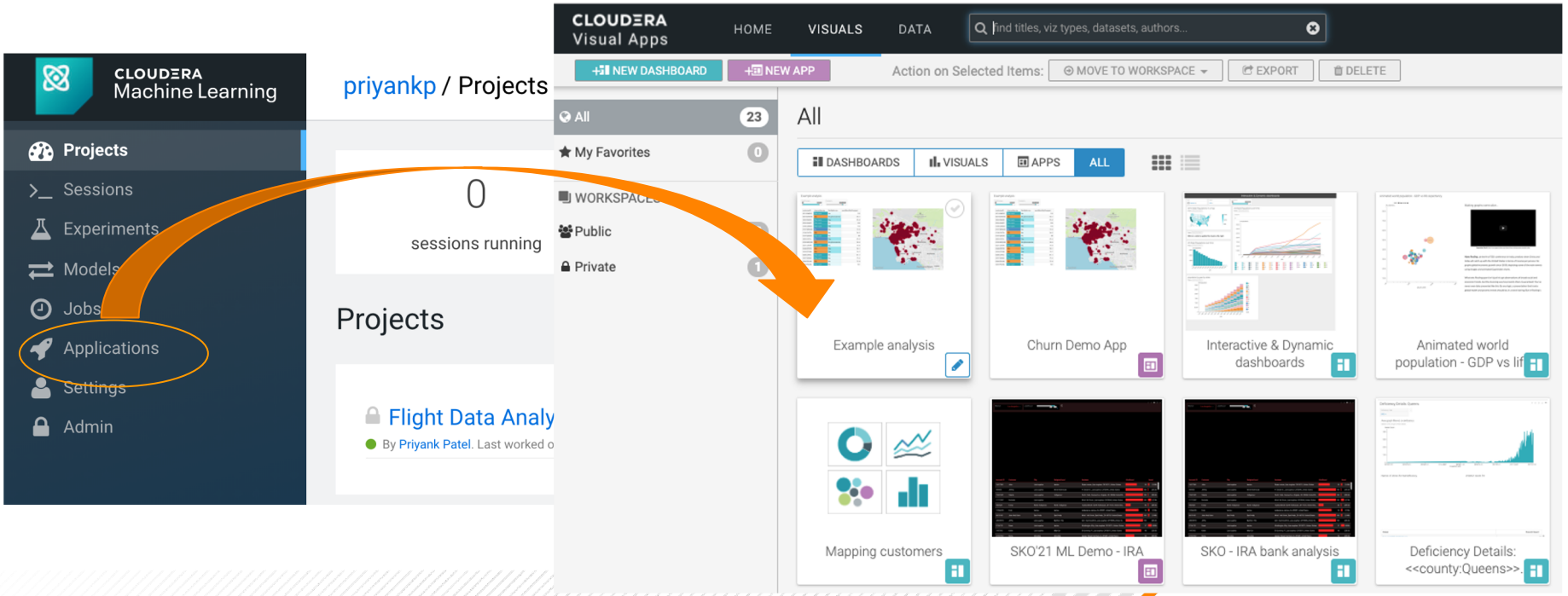
Sentiment Analysis with Sparklyr and Tensorflow

SENTIMENT ANALYSIS SPARK

License expires in 59 days

Integrated Business Intelligence

Native Data Visualizations in CML



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