CLOUDERA

Cloudera Data Engineering

The Enterprise Solution for Modern Data Engineering

Cloudera Team

Suresh MR
Director - Channels & Alliances

Vinay Rayker
Partner Technology Leader

Puneet Joshi
Partner Solutions Engineer

Pannag Katti
Partner Solutions Engineer

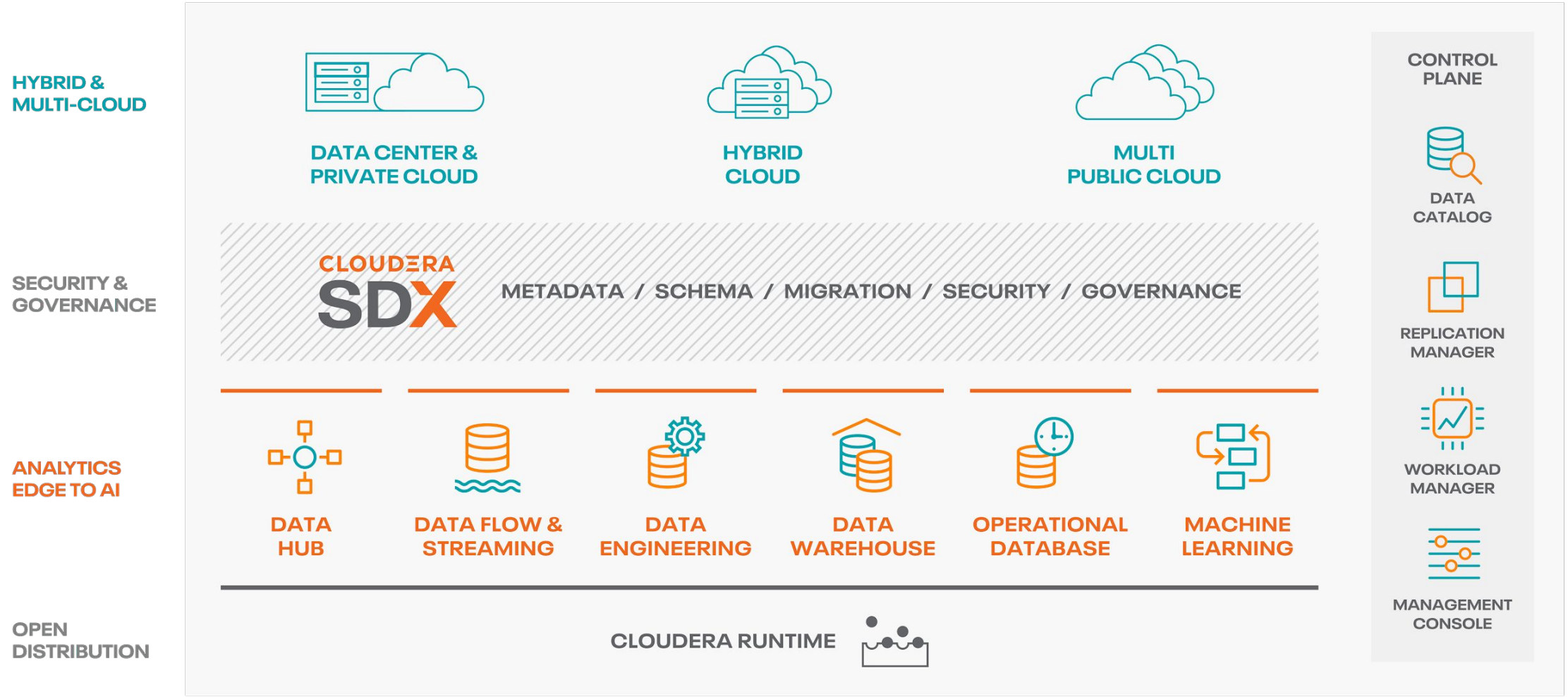
Manick Mehra
Partner Solutions Engineer

AGENDA

- **Introductions**
- **Cloudera Data Engineering - Overview (25 minutes)**
- **Workshop Overview & Detailed Instructions (20 minutes)**
- **Hands-on exercises (90 minutes)**
- **Q&A (throughout the workshop)**

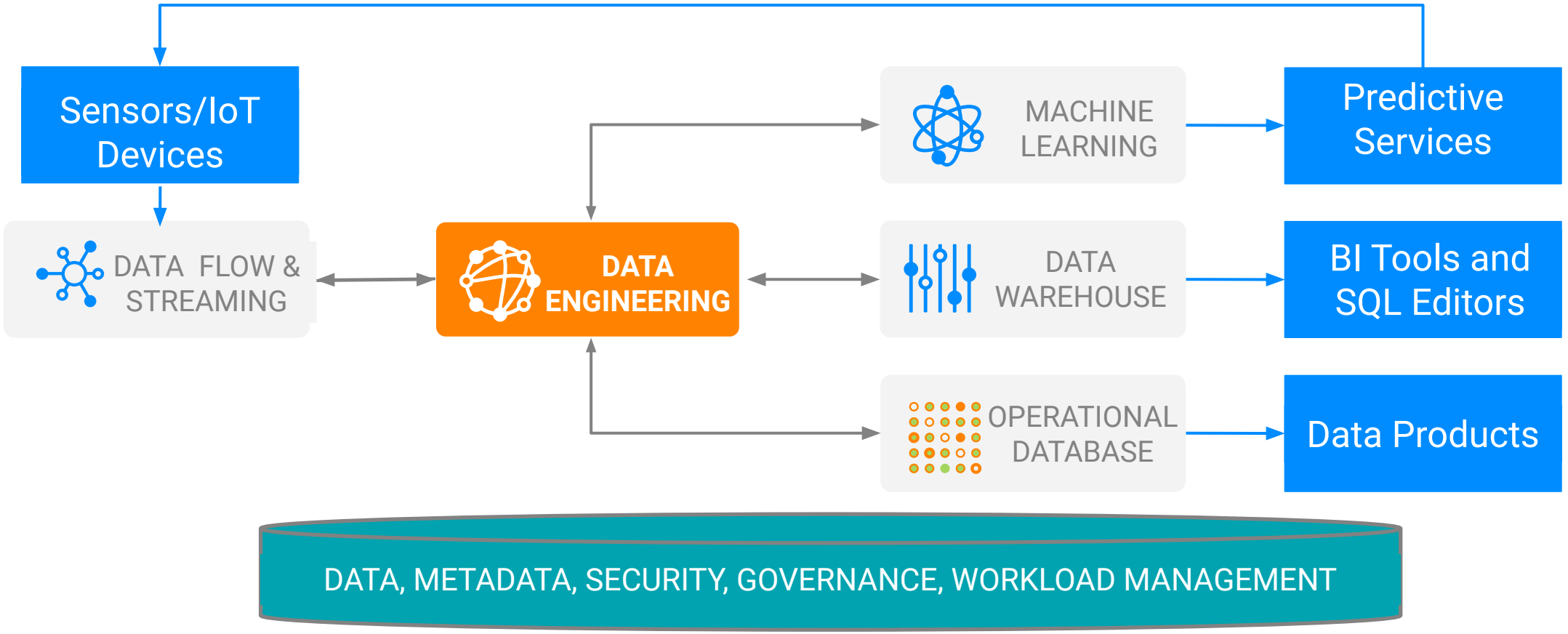
Cloudera Data Platform

CLOUDERA DATA PLATFORM

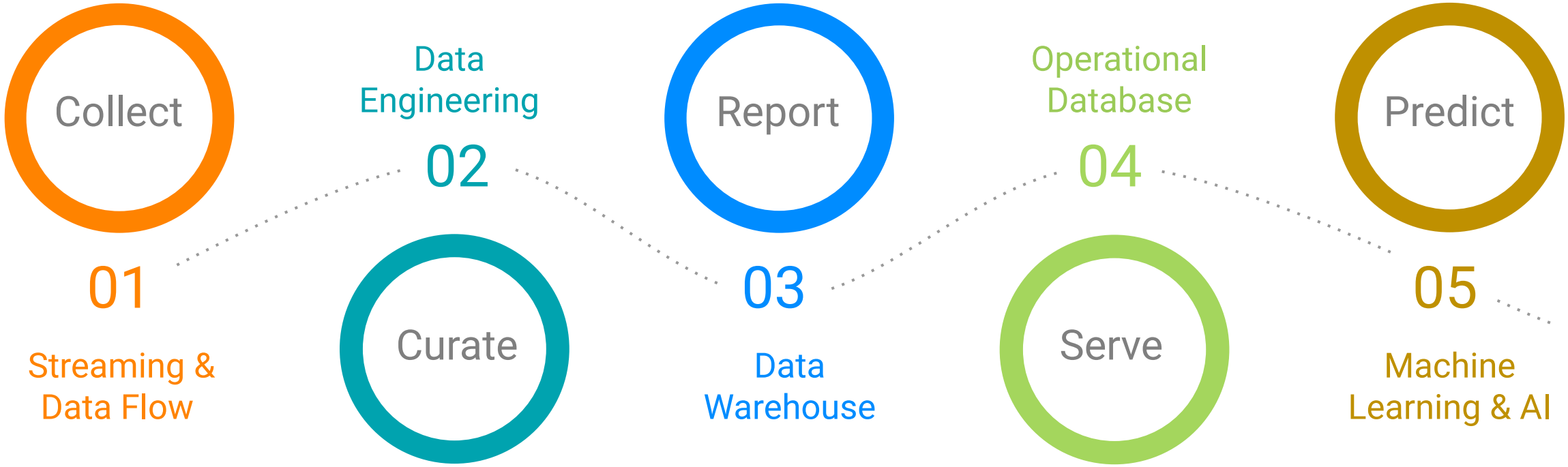


DATA ENGINEERING IS CENTRAL

To Powering Business Analytics, ML, and Data Products



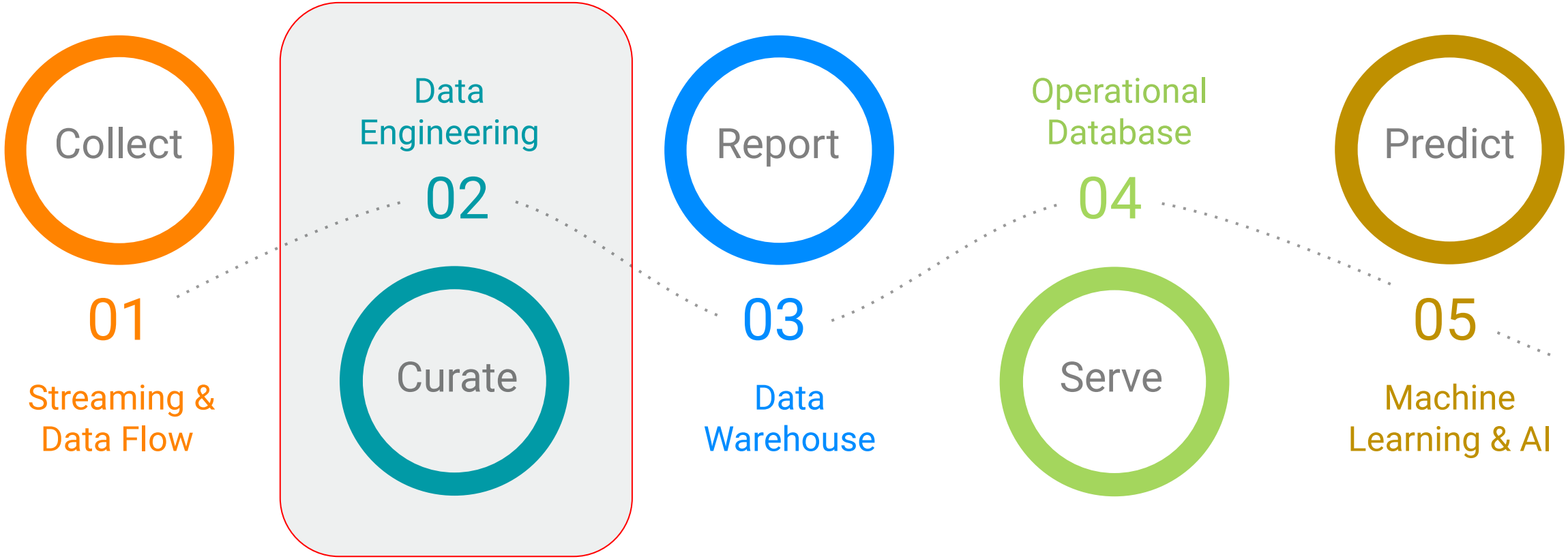
DATA ENGINEERING WITHIN THE DATA LIFECYCLE



POWERED BY CLOUDERA SDX

Security | Governance | Lineage | Management | Automation

DATA ENGINEERING WITHIN THE DATA LIFECYCLE

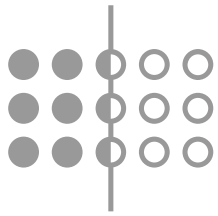


POWERED BY CLOUDERA SDX

Security | Governance | Lineage | Management | Automation

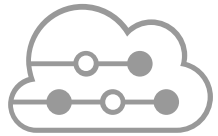
THE CHALLENGES WITH TRADITIONAL DATA ENGINEERING

Managing Spark Resources



Orchestrating Complex Pipelines

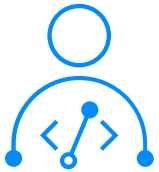
Visibility & Troubleshooting



Secure & Fast Delivery

CLOUDERA DATA ENGINEERING

An Integrated, Purpose Built Experience for Data Engineers

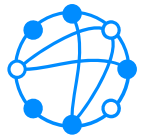


OPTIMIZED FOR DATA ENGINEERS

Streamlined service for scheduling, monitoring, debugging, and promoting data pipelines quickly & securely.

EVERYTHING YOU NEED TO POWER ANALYTICS

- Inherited governance
- Deliver data pipelines to CDW, CML, or COD easily
- Portable and flexible



COMPLETE DATA PIPELINE MANAGEMENT

- Monitoring & alerting for catching issues early
- Visual troubleshooting
- Governed and secure workflows with SDX

CLOUDERA DATA ENGINEERING

Spark run time

CLOUDERA DATA ENGINEERING

Platform operations	Autoscaling	1-click provisioning	Containerized & Isolated	Monitoring (Prometheus /Grafana)	Mix version
Spark run time					

CLOUDERA DATA ENGINEERING

Job management	Orchestration w/Airflow	Self-service visual troubleshooting	Dependency management	Artifact Versioning	Automatic lineage
Platform operations	Autoscaling	1-click provisioning	Containerized & Isolated	Monitoring (Prometheus /Grafana)	Mix version
Spark run time					

CLOUDERA DATA ENGINEERING

Seamless Integration

Management UI	CLI	API

Job management	Orchestration w/Airflow	Self-service visual troubleshooting	Dependency management	Artifact Versioning	Automatic lineage
Platform operations	Autoscaling	1-click provisioning	Containerized & Isolated	Monitoring (Prometheus /Grafana)	Mix version
Spark run time					

CLOUDERA DATA ENGINEERING

Seamless Integration

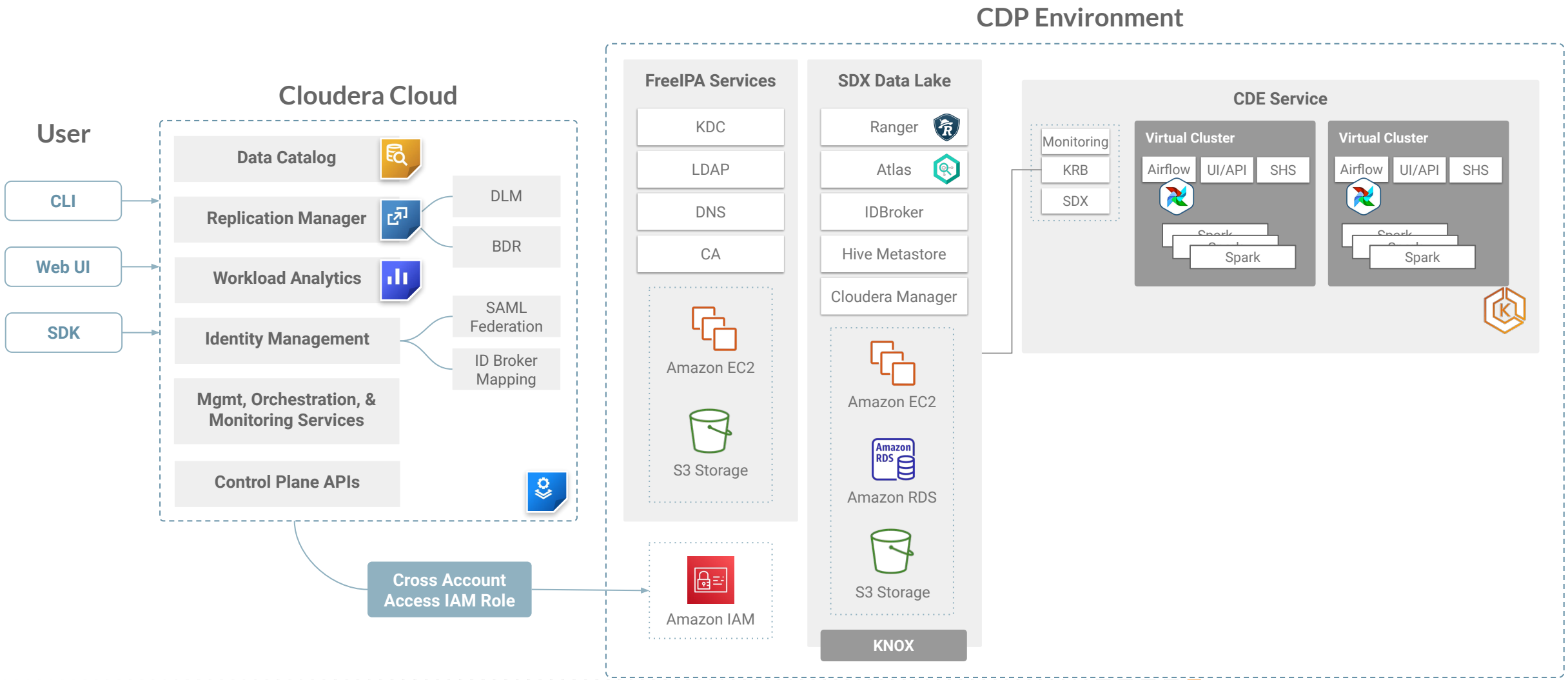
Management UI | **CLI** | **API**

Job management	Orchestration w/Airflow	Self-service visual troubleshooting	Dependency management	Artifact Versioning	Automatic lineage
Platform operations	Autoscaling (Yunikorn)	1-click provisioning	Containerized & Isolated	Monitoring (Prometheus /Grafana)	Mix version
Spark run time	Apache Spark + Kubernetes				

Data anywhere

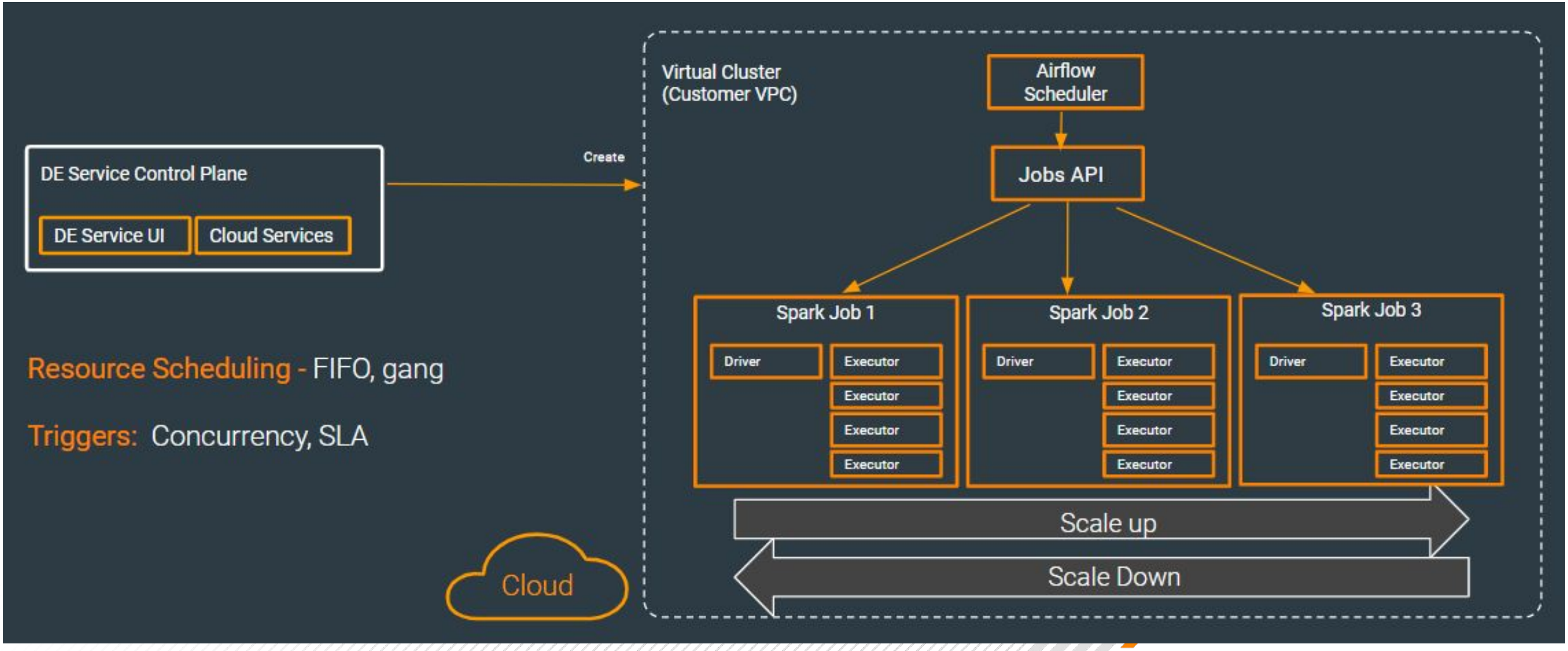
CLOUDERA SDX — METADATA / SCHEMA / MIGRATION / SECURITY / GOVERNANCE

CDP - AWS HIGH LEVEL ARCHITECTURE



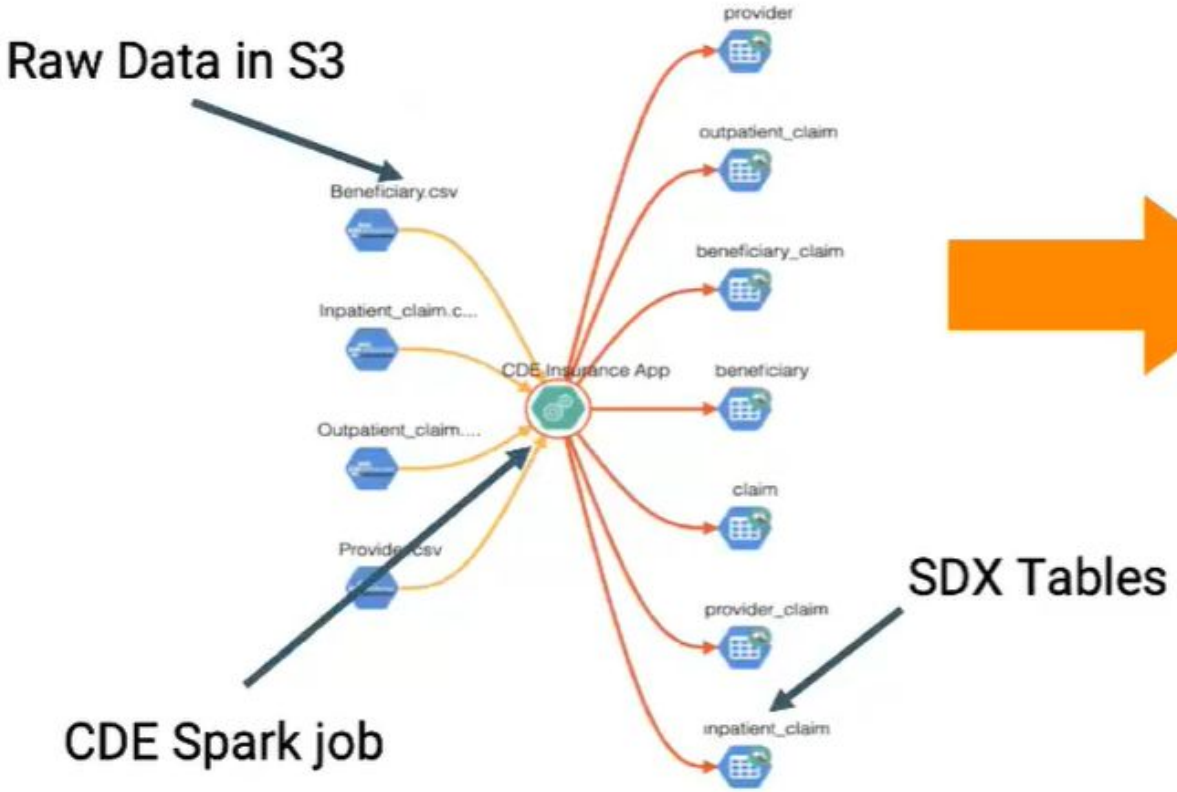
CDE Environment Scaling

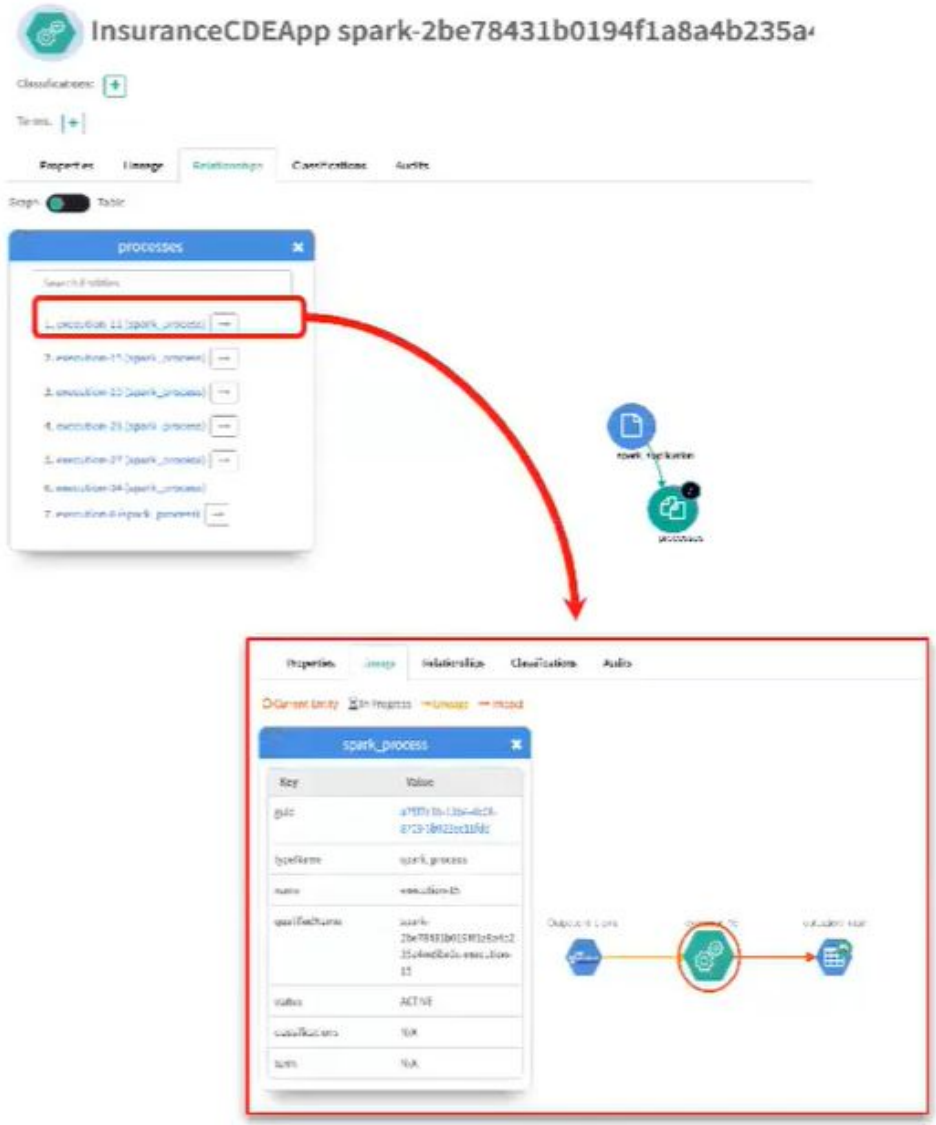
Automatic Resource Scaling



SPARK ATLAS CONNECTOR

Better spark session lineage
Single to Multi-process entities

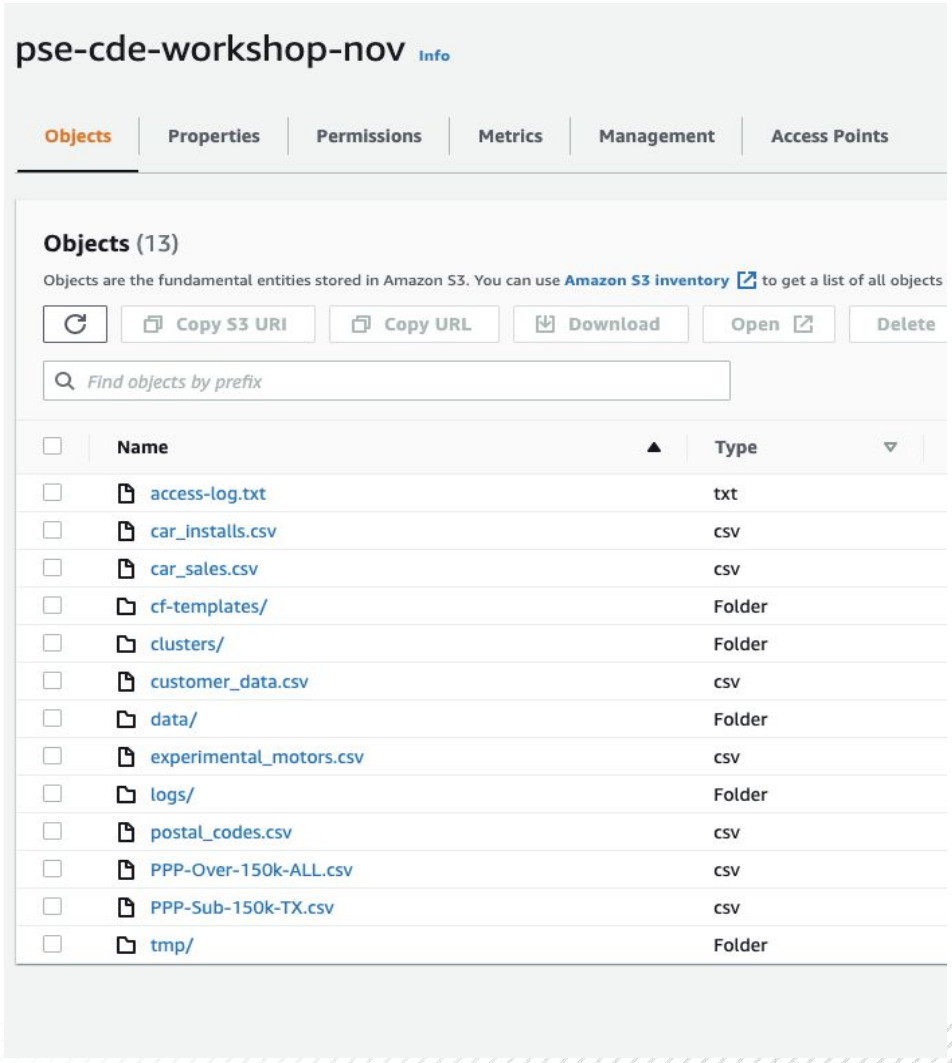




Workshop Quick Tour

- **Lab 0 - Prerequisites**
- **Lab 1 - Walkthrough of CDE service**
- **Lab 2 - Create and trigger ad-hoc Spark jobs**
- **Lab 3 - Add schedule to the ad-hoc Spark jobs**
- **Lab 4 - Orchestrate a set of jobs using Airflow**
- **Lab 5 - Install and Configure CDE CLI**
- **Lab 6 - Run jobs using CDE CLI**
- **Lab 7 - Data Lineage and Auto-Scaling**

Datasets


pse-cde-workshop-nov Info
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (13)
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects
Copy S3 URI
Copy URL
Download
Open
Delete
Find objects by prefix
Name
Type
access-log.txt txt
car_installs.csv csv
car_sales.csv csv
cf-templates/ Folder
clusters/ Folder
customer_data.csv csv
data/ Folder
experimental_motors.csv csv
logs/ Folder
postal_codes.csv csv
PPP-Over-150k-ALL.csv csv
PPP-Sub-150k-TX.csv csv
tmp/ Folder

Use cases

- **Log Data Cleansing using Spark**

- **Analyze the Paycheck Protection Program Data**
  - **Report 1: Breakdown of all cities in Texas that retained jobs**
  - **Report 2: Breakdown of company type that retained jobs**

- **PySpark job to enrich your data using an existing data warehouse**

Get Set, GO!

https://github.com/vrayker/CDE_Workshop/blob/main/CDE_Workshop_Student_Guide%20(2).pdf