



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



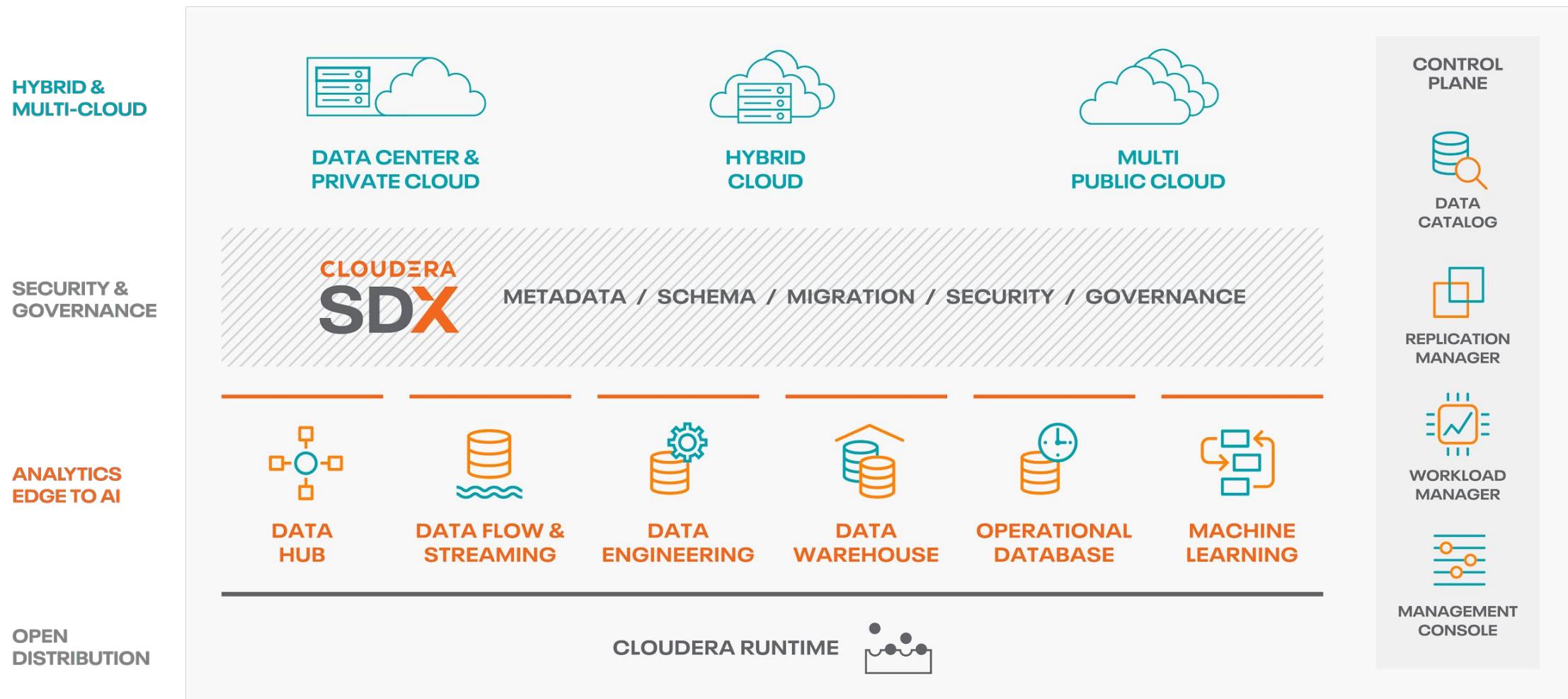
Manick Mehra
Partner Solutions
Engineer

AGENDA

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

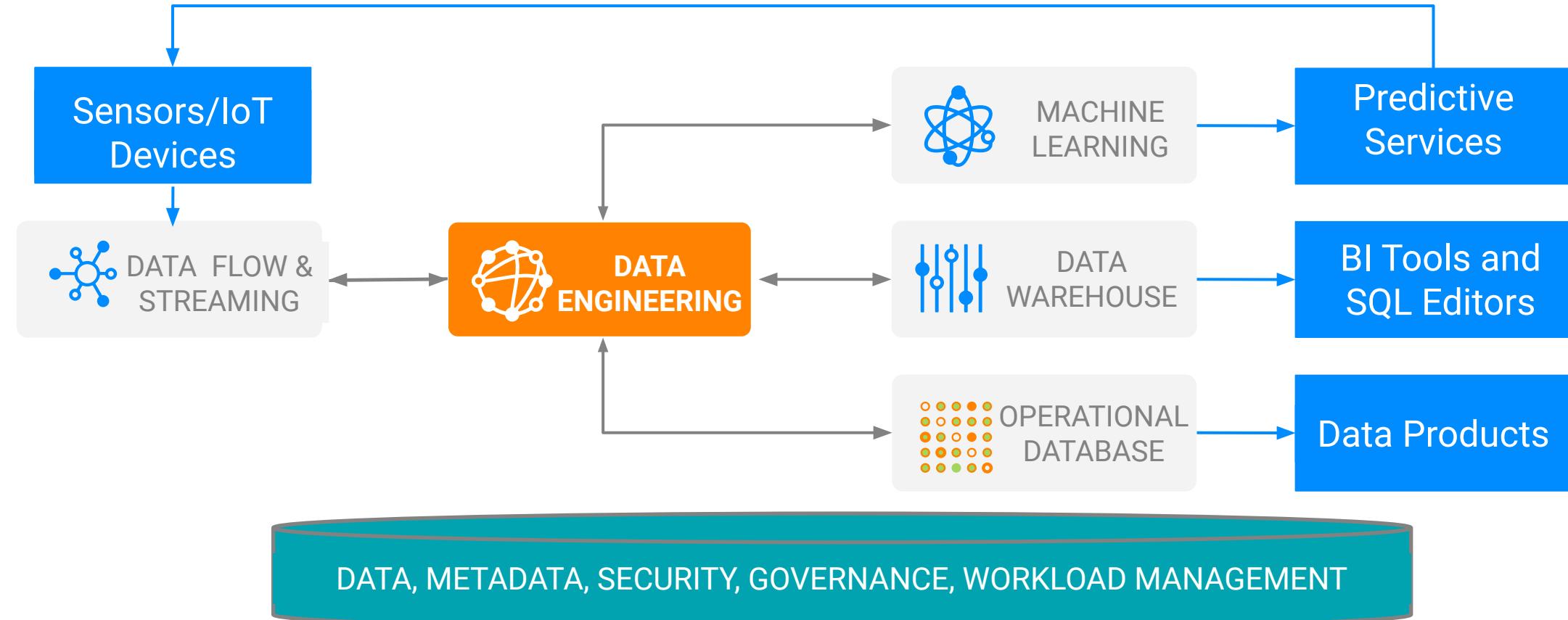
Cloudera Data Platform

CLOUDERA DATA PLATFORM



DATA ENGINEERING IS CENTRAL

To Powering Business Analytics, ML, and Data Products



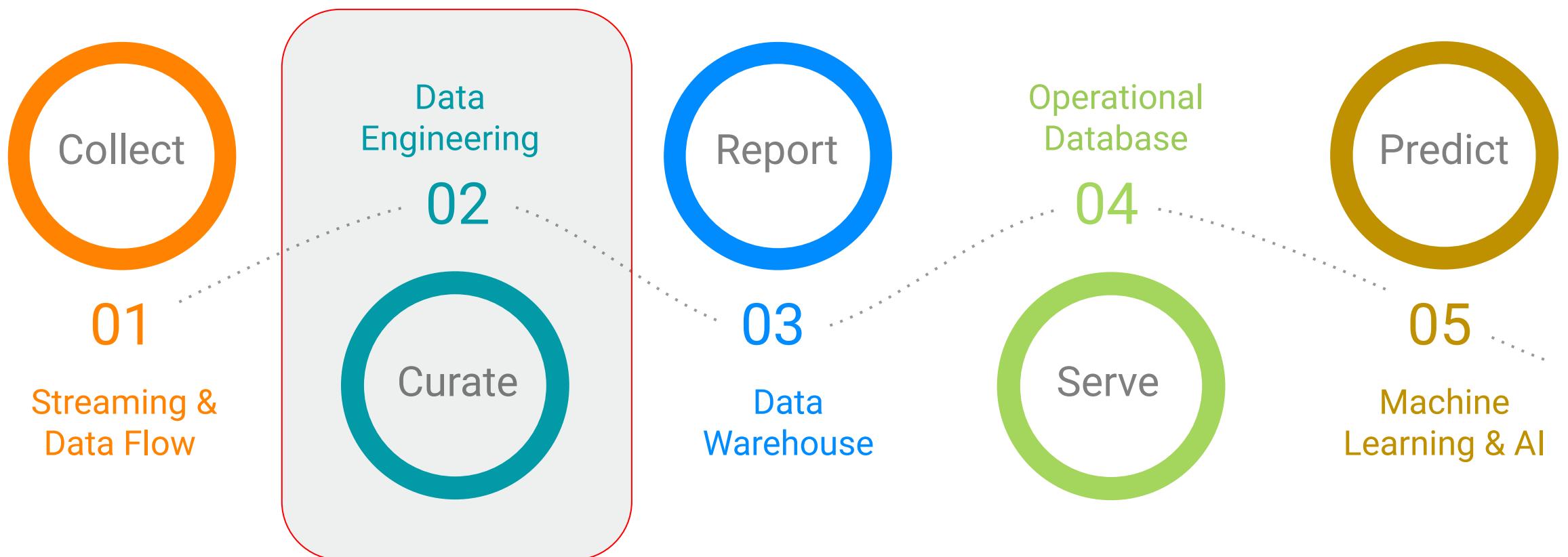
DATA ENGINEERING WITHIN THE DATA LIFECYCLE



POWERED BY **CLOUDERA
SDX**

Security | Governance | Lineage | Management | Automation

DATA ENGINEERING WITHIN THE DATA LIFECYCLE



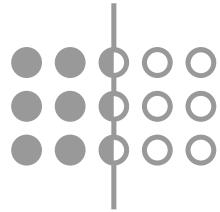
POWERED BY **CLOUDERA**
SDX

Security | Governance | Lineage | Management | Automation

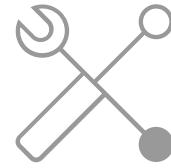
THE CHALLENGES WITH TRADITIONAL DATA ENGINEERING



**Managing Spark
Resources**



**Orchestrating Complex
Pipelines**



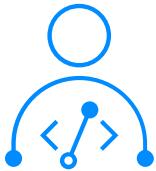
**Visibility &
Troubleshooting**



**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



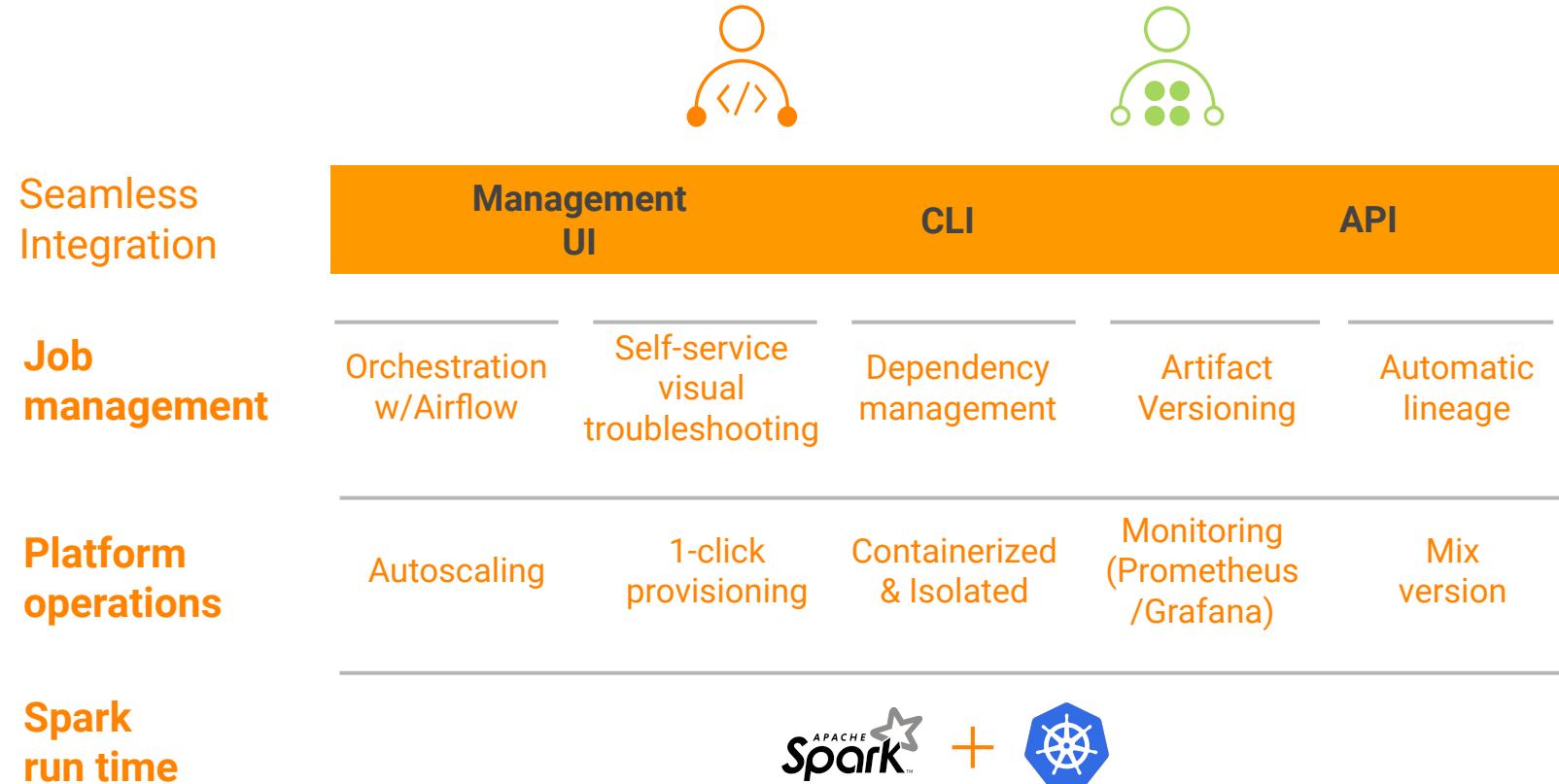
CLOUDERA DATA ENGINEERING

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

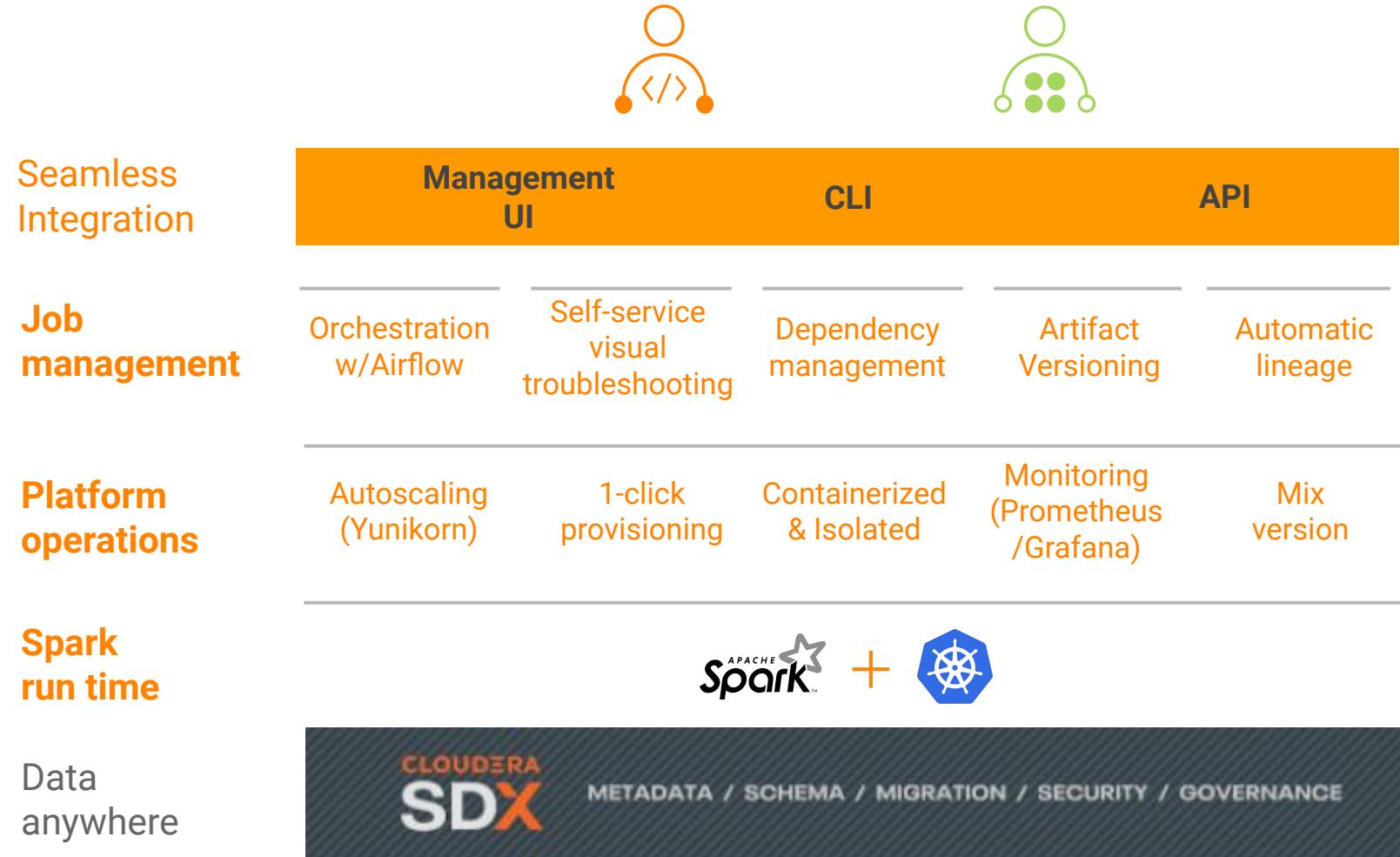
CLOUDERA DATA ENGINEERING

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



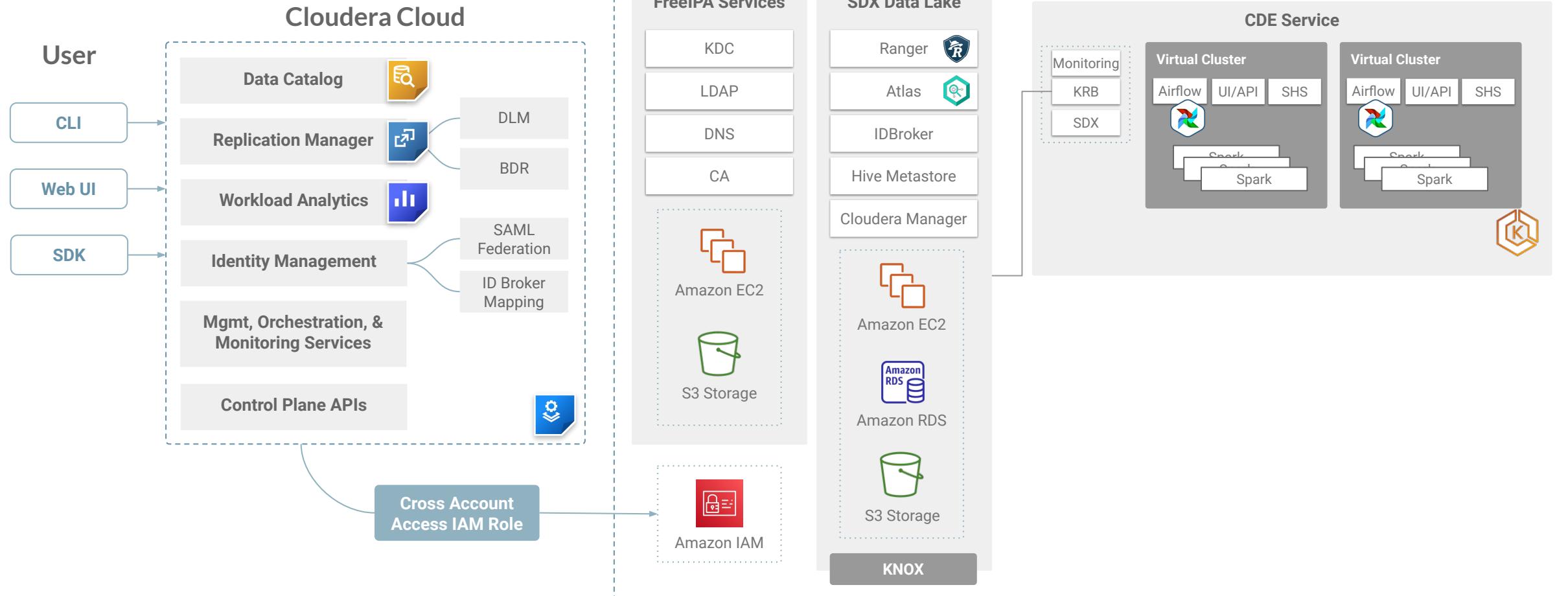
CLOUDERA DATA ENGINEERING



CDP - AWS HIGH LEVEL ARCHITECTURE

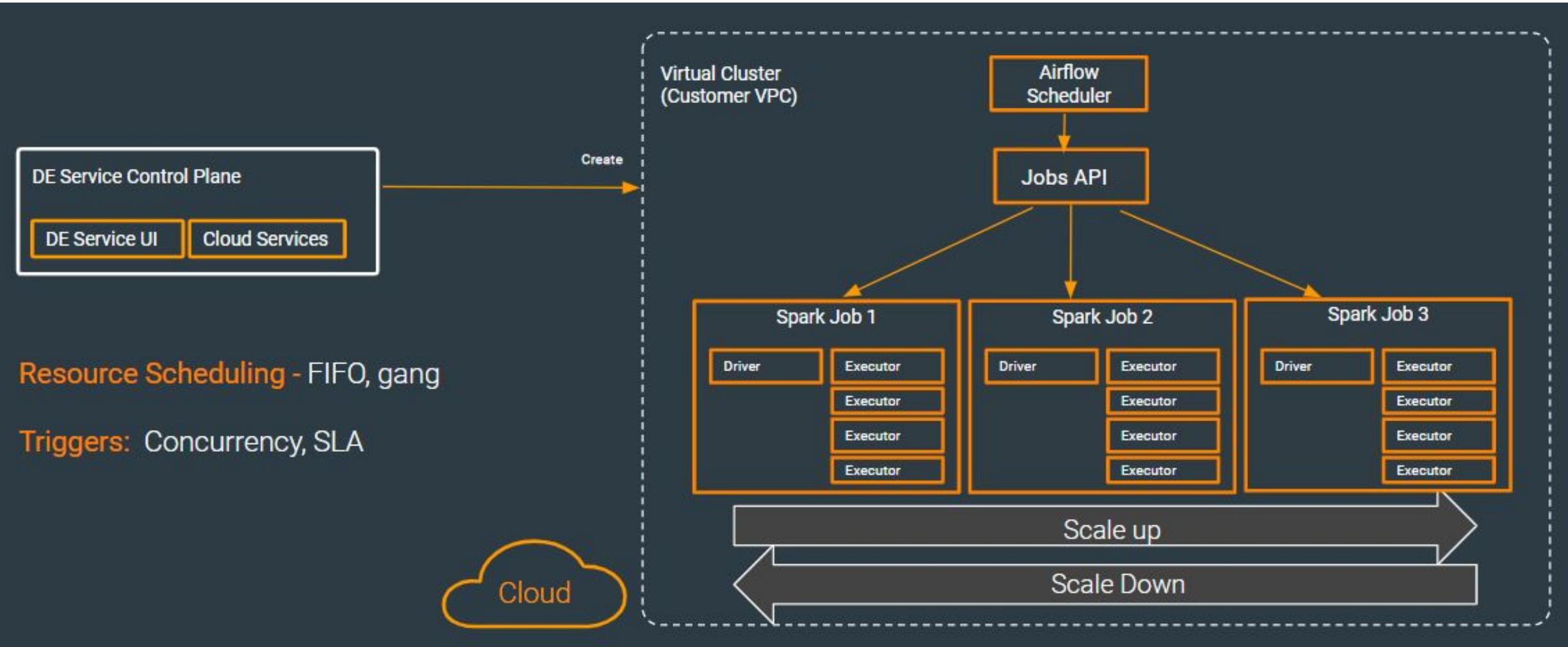


CDP Environment



CDE Environment Scaling

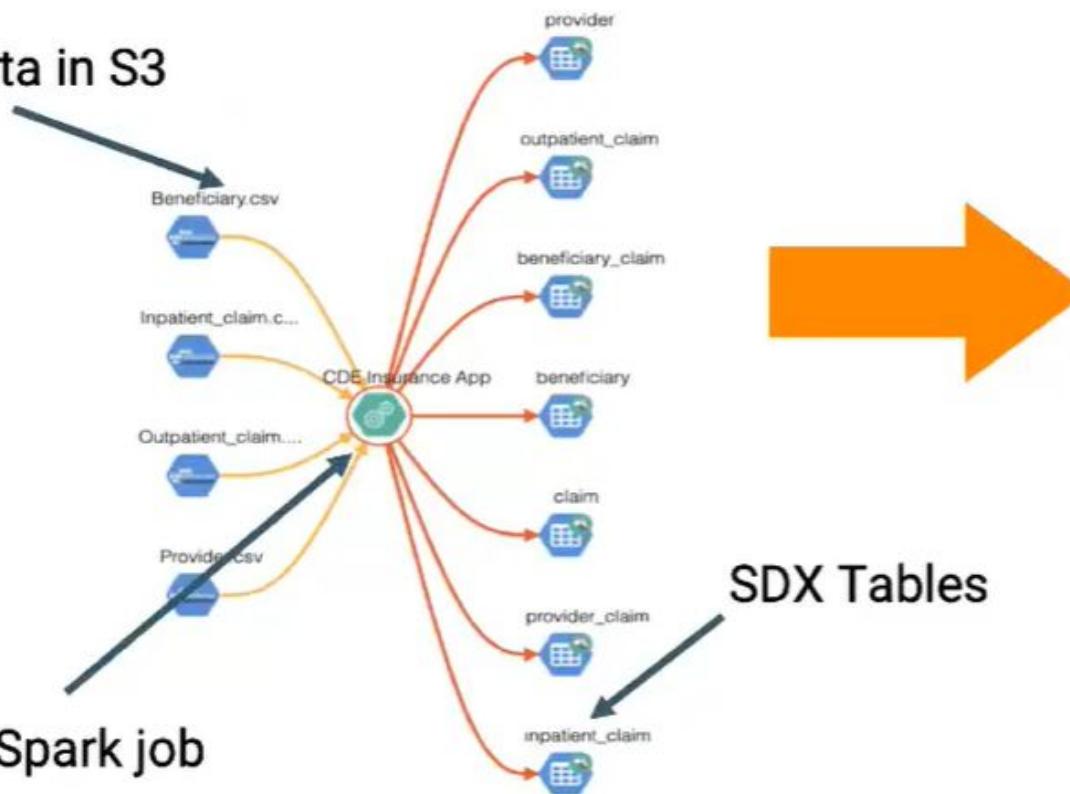
Automatic Resource Scaling



SPARK ATLAS CONNECTOR

Better spark session lineage
Single to Multi-process entities

Raw Data in S3



CDE Spark job

The screenshot shows the Cloudera Data Engine (CDE) interface. At the top, a header bar displays 'InsuranceCDEApp spark-2be78431b0194f1a8a4b235a4'. Below the header are tabs for 'Properties', 'Image', 'Relationships' (which is selected), 'Connections', and 'Audit'. A 'Search' toggle is set to 'Table'. A modal window titled 'processes' lists seven entries, each with a red box highlighting the first item: '1. execution-11 (spark_process)'. To the right of the modal, a 'work tracker' icon is connected to a 'spark_process' node. A large red curved arrow points from the 'spark_process' node in the modal to the 'spark_process' node in the main interface. The main interface shows a detailed view of the 'spark_process' entity with the following properties:

Key	Value
id	a79710c1-13e4-4c26-8723-1002ec115d
typeItem	spark_process
name	execution-11
qualFeature	spark-2be78431b0194f1a8a4b235a4-execution-11
status	ACTIVE
classifiers	10K
last	10K

Below the table, a small graph visualization shows the relationship between 'Outpatient claim', 'Execution-11', and 'Inpatient claim'.

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

<input type="checkbox"/>	Name	Type
<input type="checkbox"/>	access-log.txt	txt
<input type="checkbox"/>	car_installs.csv	csv
<input type="checkbox"/>	car_sales.csv	csv
<input type="checkbox"/>	cf-templates/	Folder
<input type="checkbox"/>	clusters/	Folder
<input type="checkbox"/>	customer_data.csv	csv
<input type="checkbox"/>	data/	Folder
<input type="checkbox"/>	experimental_motors.csv	csv
<input type="checkbox"/>	logs/	Folder
<input type="checkbox"/>	postal_codes.csv	csv
<input type="checkbox"/>	PPP-Over-150k-ALL.csv	csv
<input type="checkbox"/>	PPP-Sub-150k-TX.csv	csv
<input type="checkbox"/>	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](https://github.com/vrayker/CDE_Workshop/blob/main/CDE_Workshop_Student_Guide%20(2).pdf)