



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



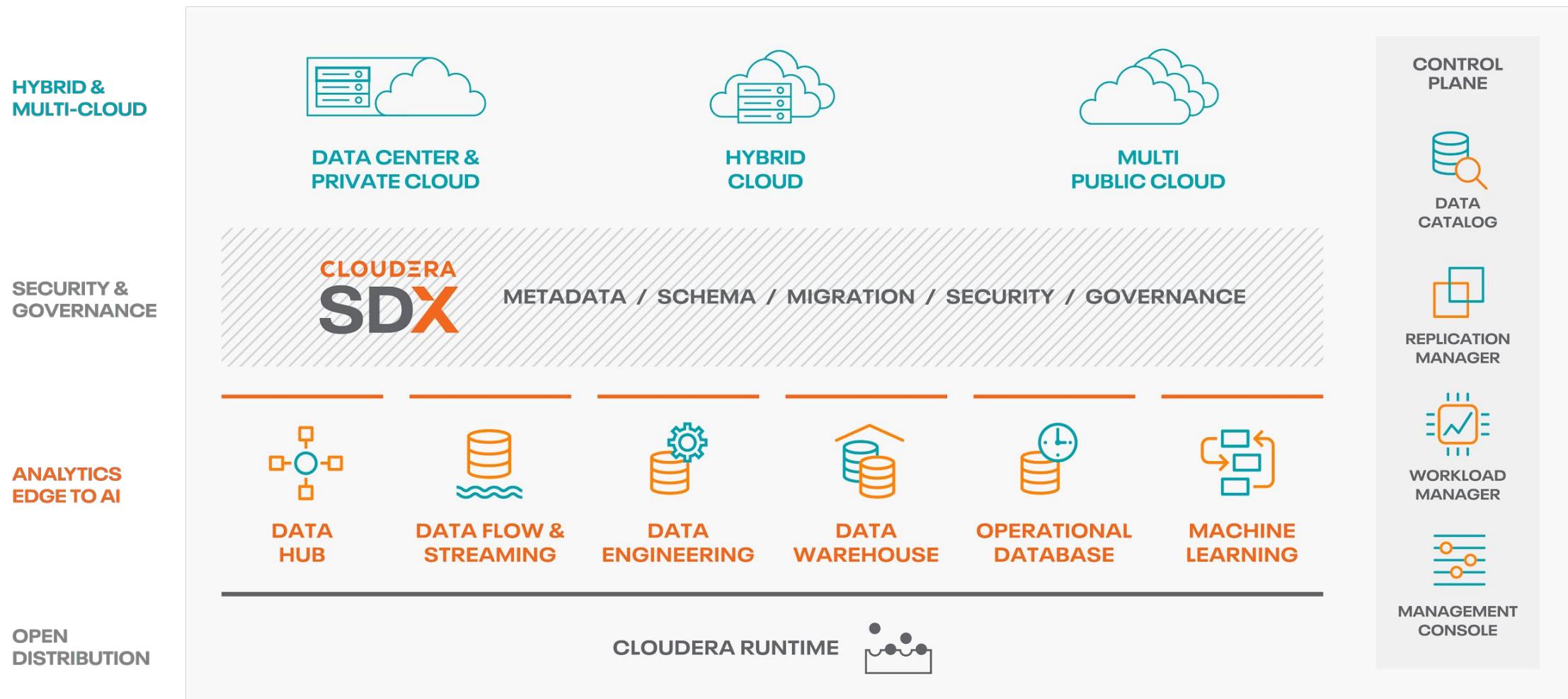
Manick Mehra
Partner Solutions
Engineer

AGENDA

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

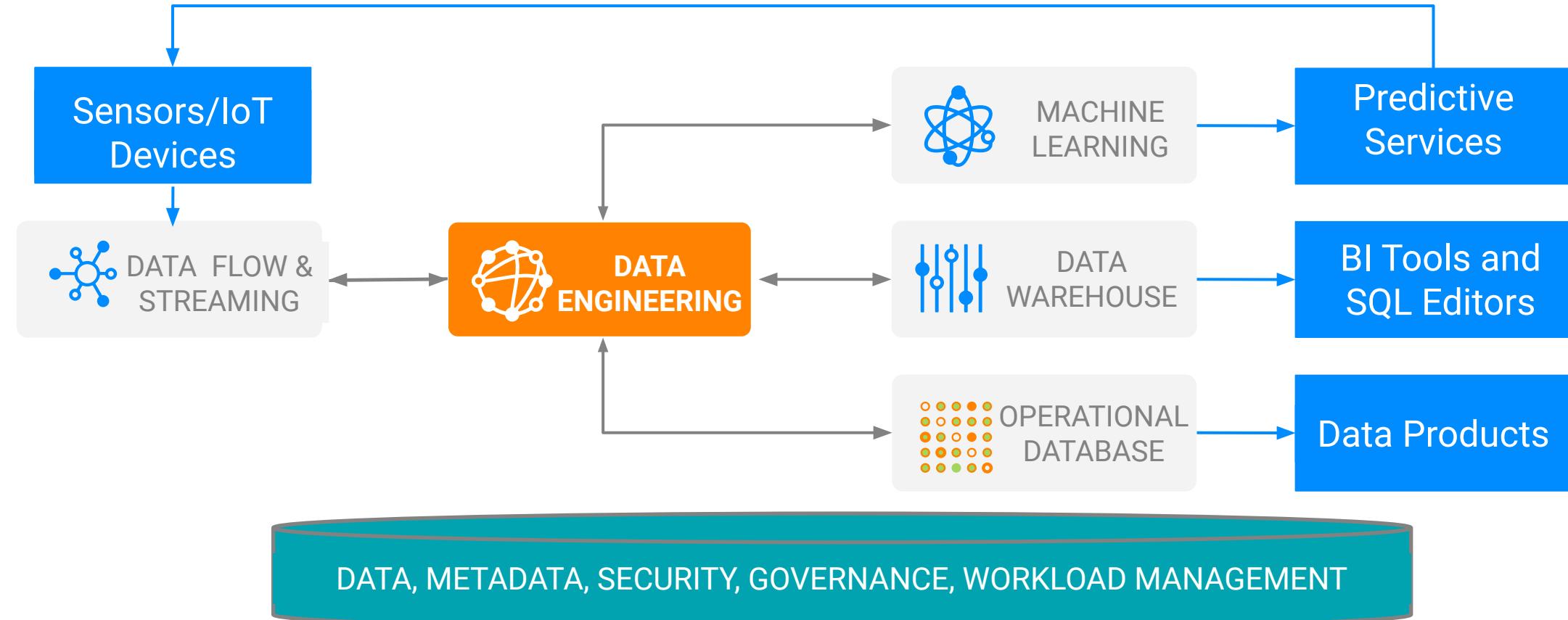
Cloudera Data Platform

CLOUDERA DATA PLATFORM



DATA ENGINEERING IS CENTRAL

To Powering Business Analytics, ML, and Data Products



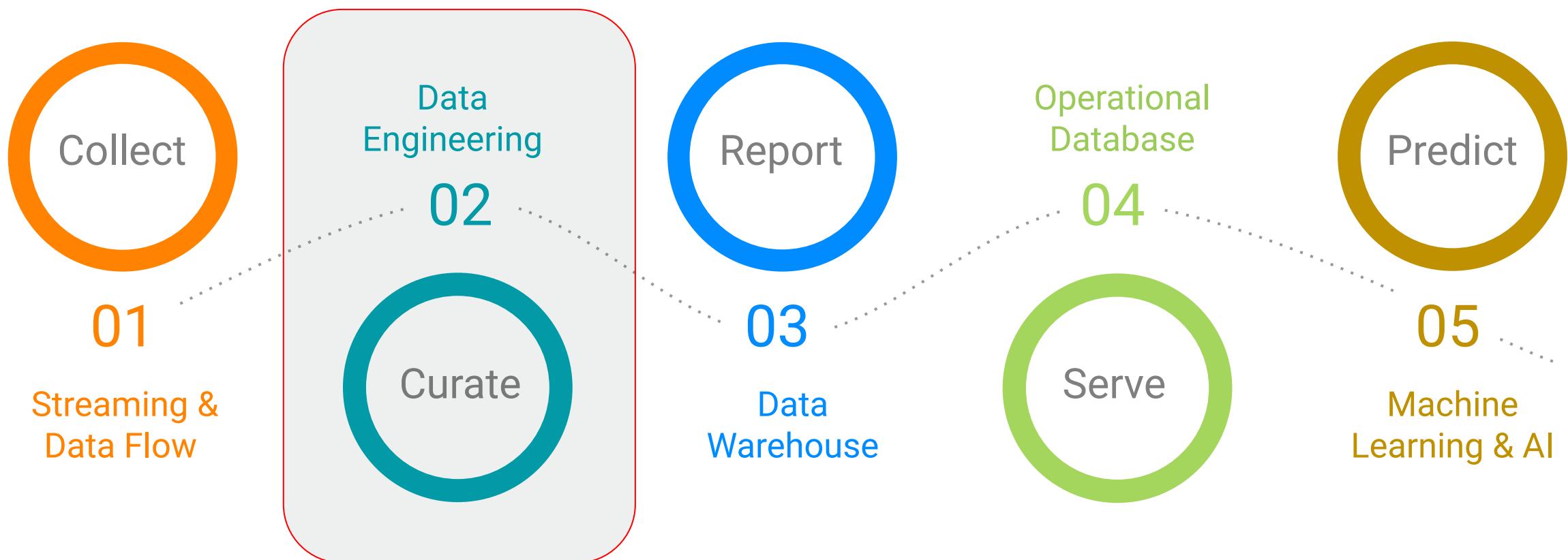
DATA ENGINEERING WITHIN THE DATA LIFECYCLE



POWERED BY **CLOUDERA
SDX**

Security | Governance | Lineage | Management | Automation

DATA ENGINEERING WITHIN THE DATA LIFECYCLE



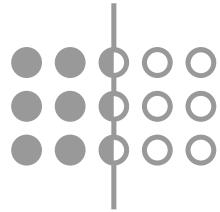
POWERED BY **CLOUDERA**
SDX

Security | Governance | Lineage | Management | Automation

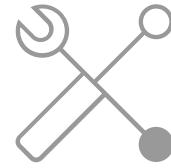
THE CHALLENGES WITH TRADITIONAL DATA ENGINEERING



**Managing Spark
Resources**



**Orchestrating Complex
Pipelines**



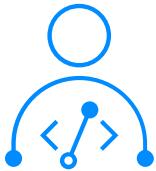
**Visibility &
Troubleshooting**



**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



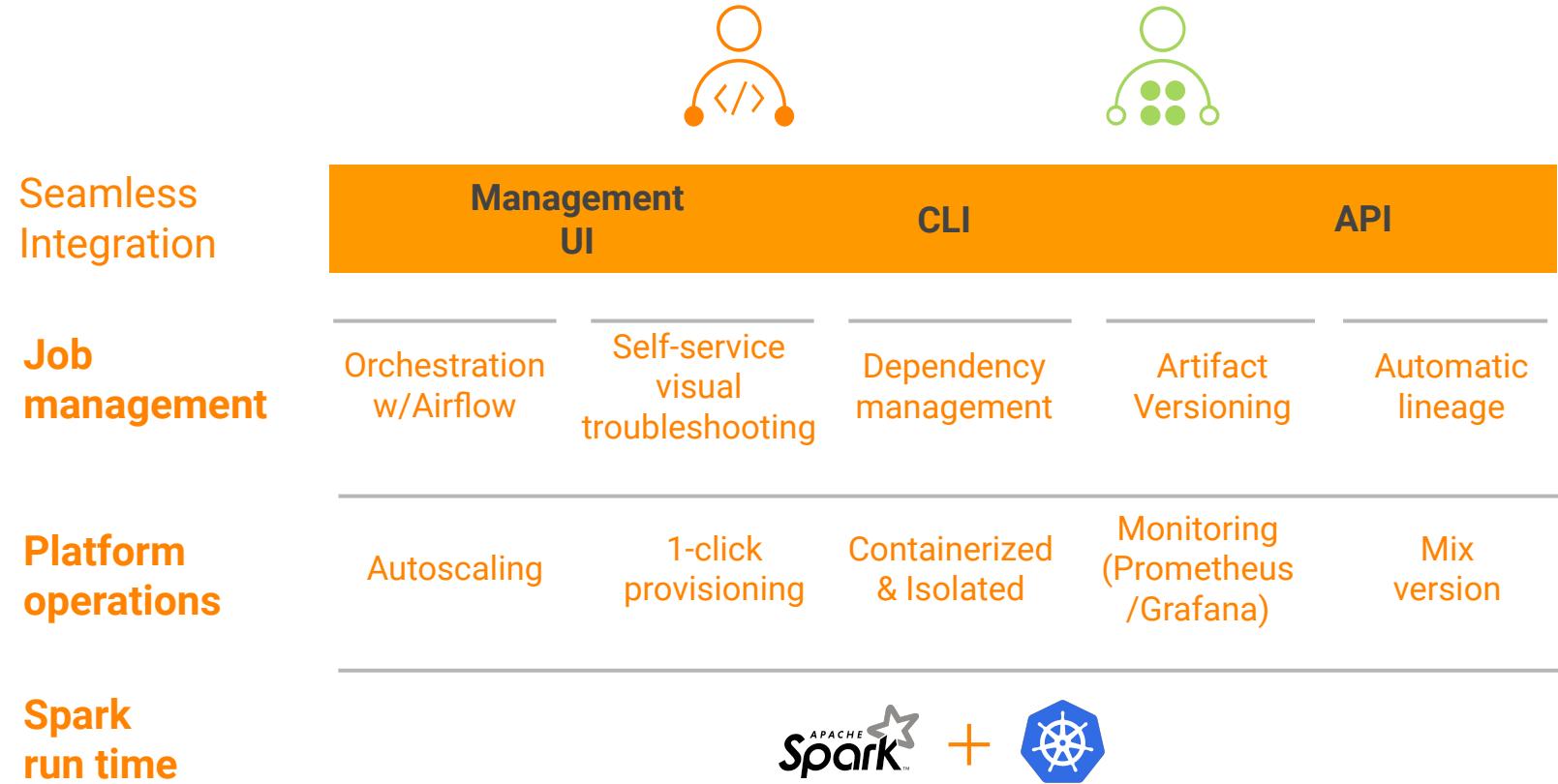
CLOUDERA DATA ENGINEERING

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

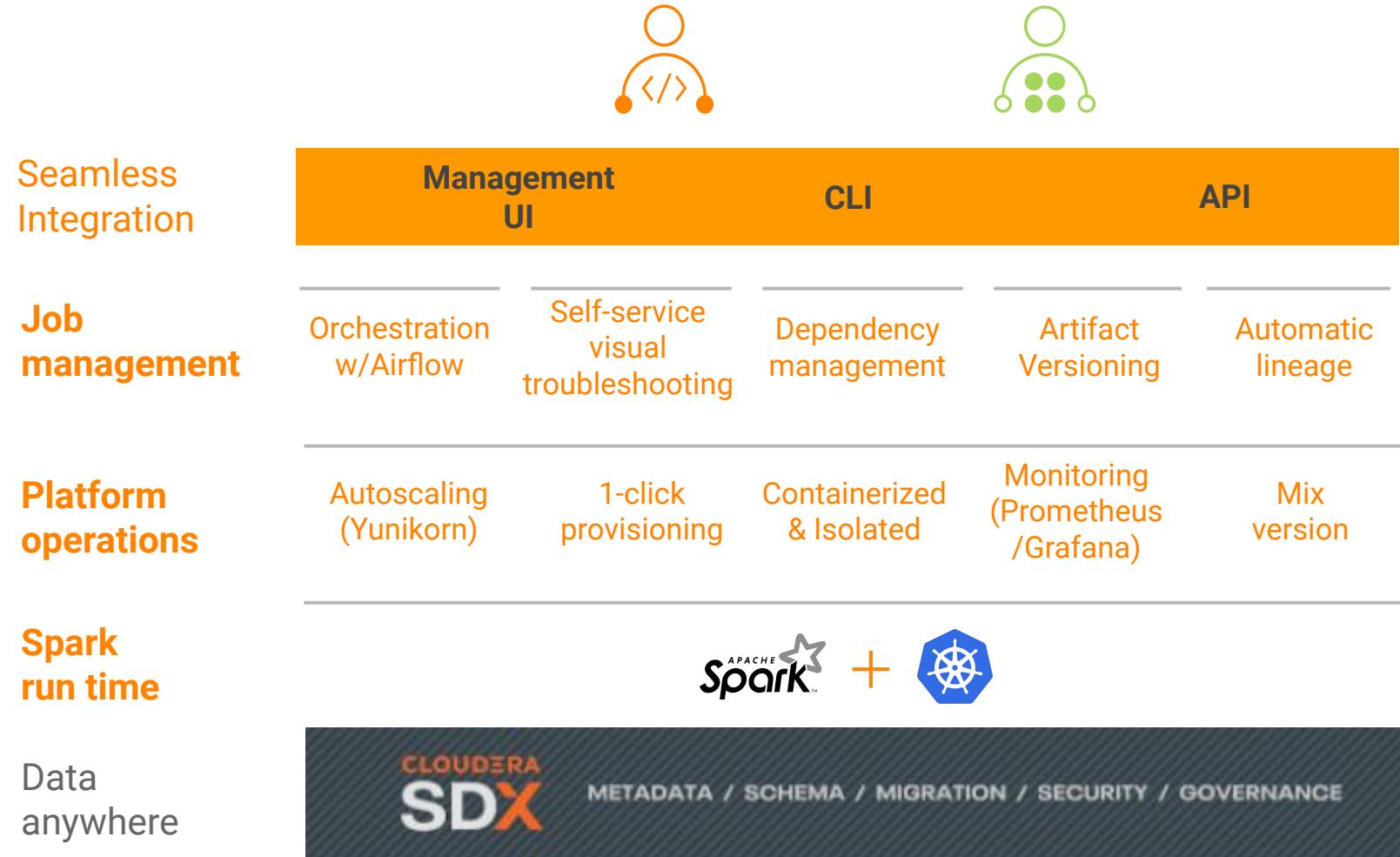
CLOUDERA DATA ENGINEERING

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



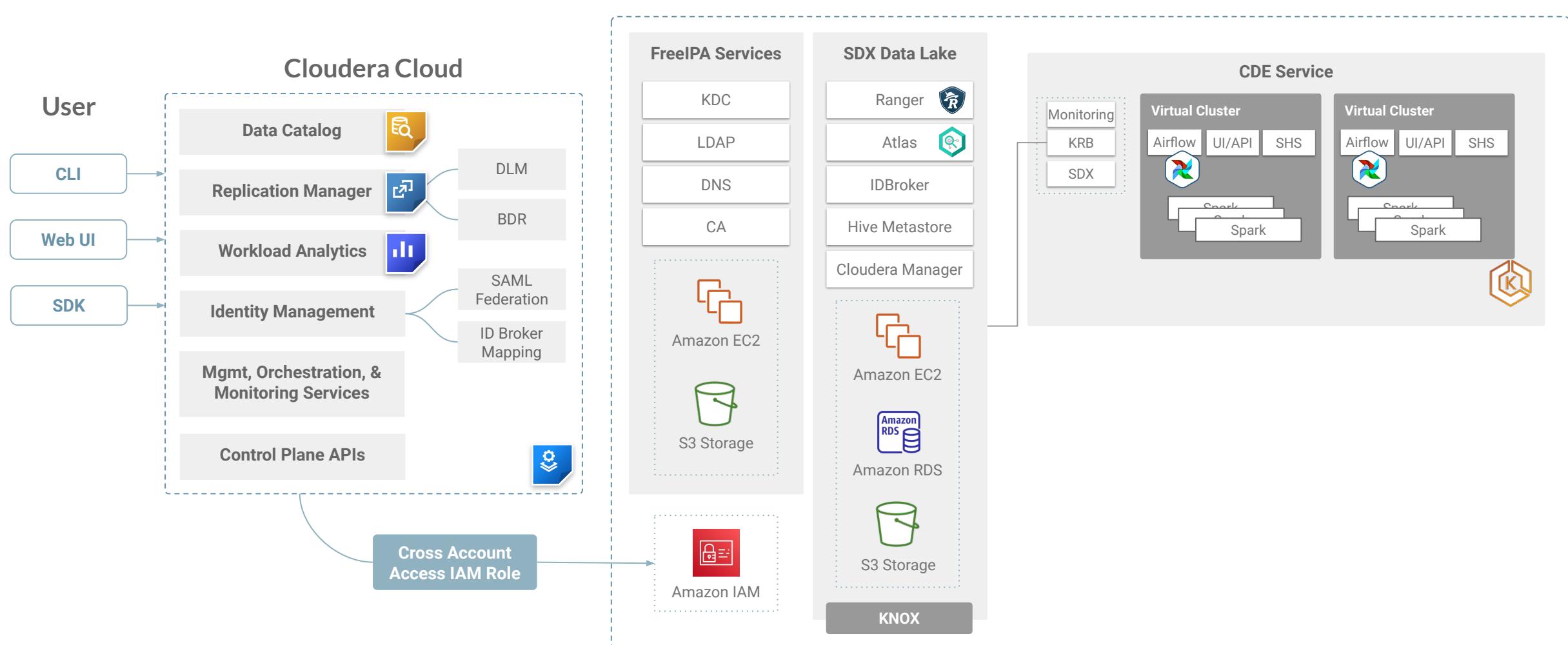
CLOUDERA DATA ENGINEERING



CDP - AWS HIGH LEVEL ARCHITECTURE

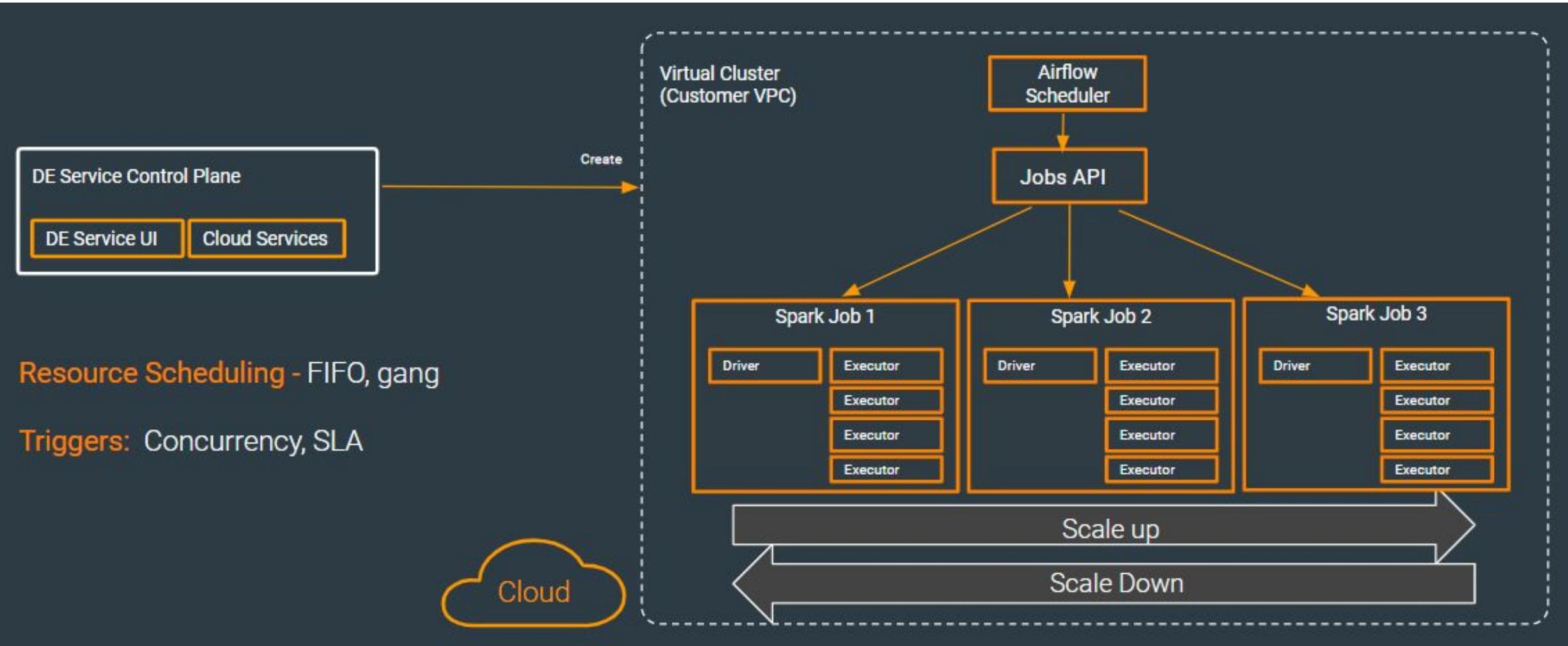


CDP Environment



CDE Environment Scaling

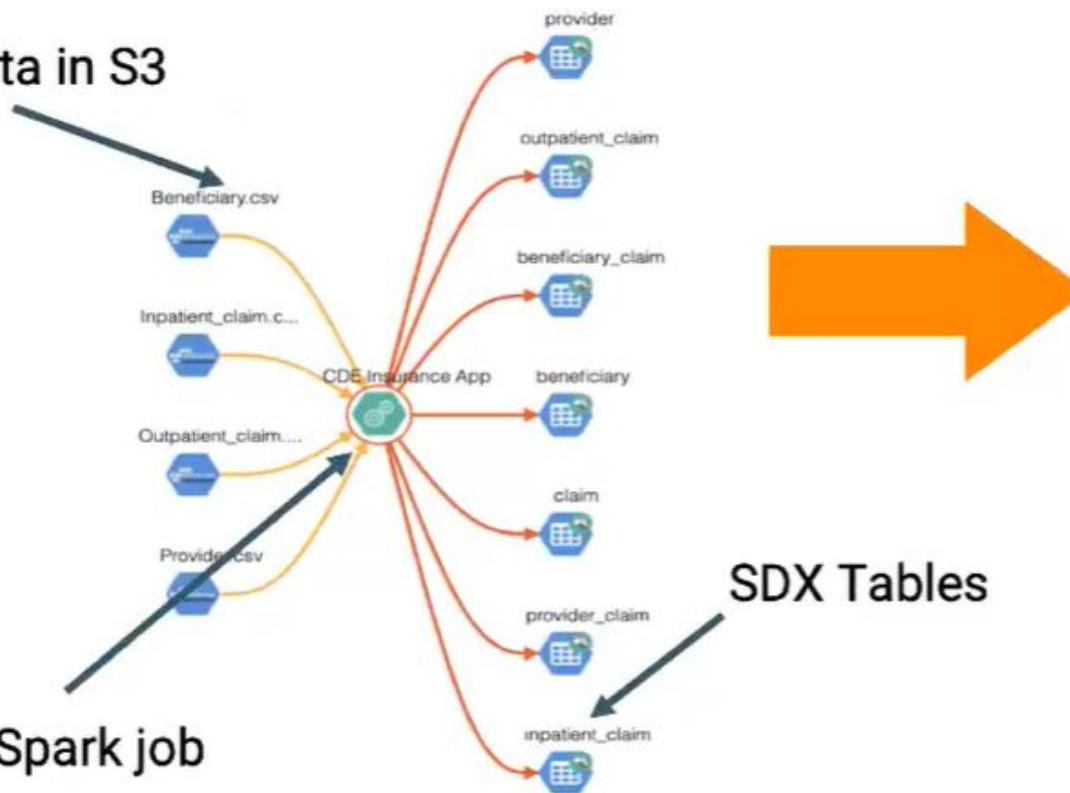
Automatic Resource Scaling



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'. Under the 'Relationships' tab, a sub-tab 'Table' is selected. A large red box highlights a dropdown menu titled 'processes' containing seven items, each preceded by a small icon: 1. execution-11 (spark_process), 2. execution-15 (spark_process), 3. execution-22 (spark_process), 4. execution-23 (spark_process), 5. execution-27 (spark_process), 6. execution-28 (spark_process), and 7. execution-30 (spark_process). A red curved arrow points from this menu to a specific entry in the list. To the right of the dropdown, there is a 'work backlog' section with a teal circular icon labeled 'spark_process'. Below the dropdown is a detailed view of a single process entry:

Key	Value
id	a79710c1-13e4-4c26-8723-1002ec115d
typeItem	spark_process
name	execution-15
qualFeature	spark-2be78431b0194f1a8a4b235a4-execution-15
status	ACTIVE
classifiers	10K
last	10A

Below this table is a small graph visualization showing the relationships between entities: 'Outpatient claim' is connected to 'Execution-15' (highlighted with a red circle), which is connected to 'Inpatient claim'.

THANK YOU

CLOUDERA