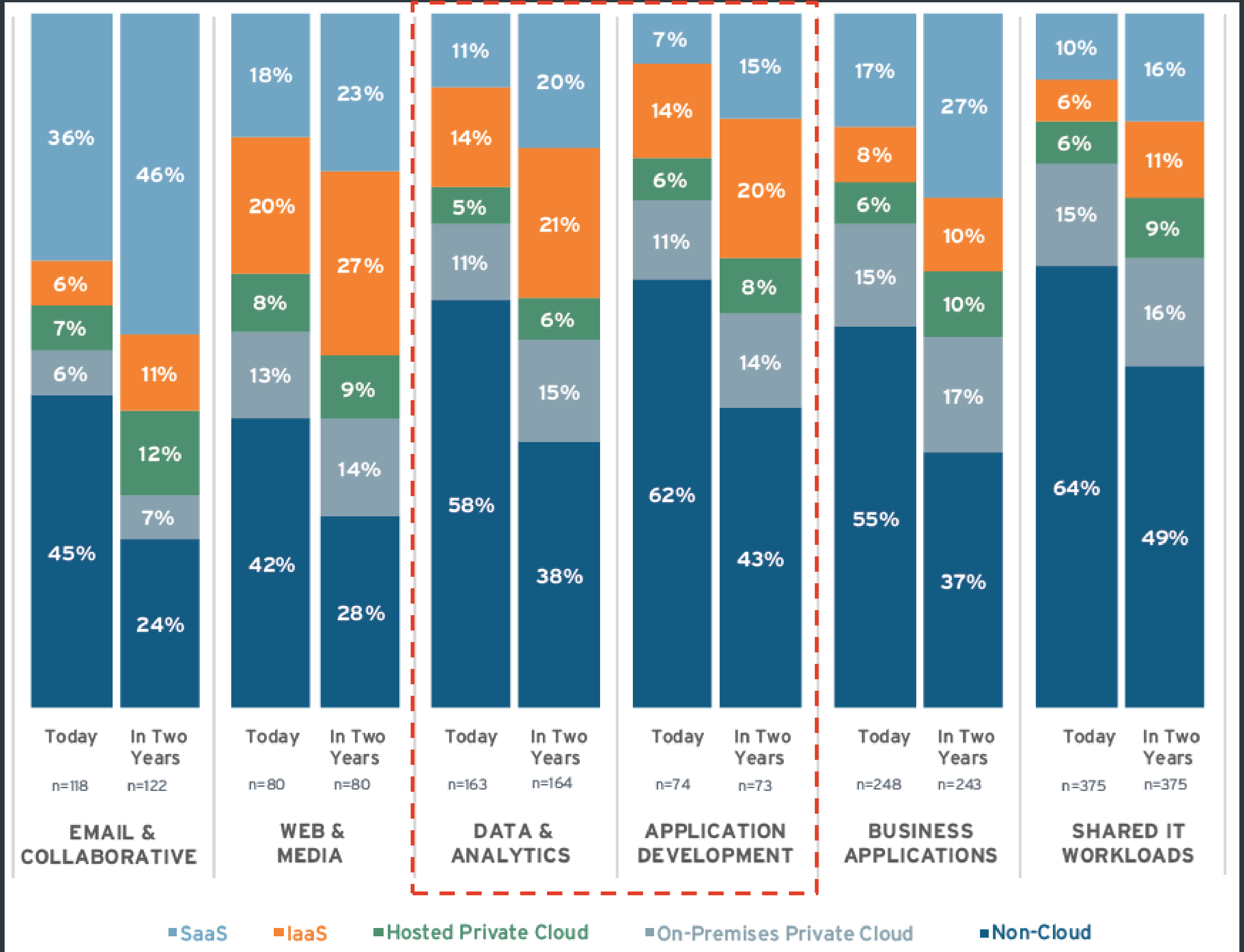
cloudera

Cloudera Altus: Big Data in der Cloud einfach gemacht

Michael Kohs | Sales Engineer | mkohs@cloudera.com

Shift to cloud: an analyst view



Key Points

- Cloud deployments will be the dominant environment in every category
- Every cloud deployment environment will see increases in every workload category
- Analytics and App Development areas expected strong gains

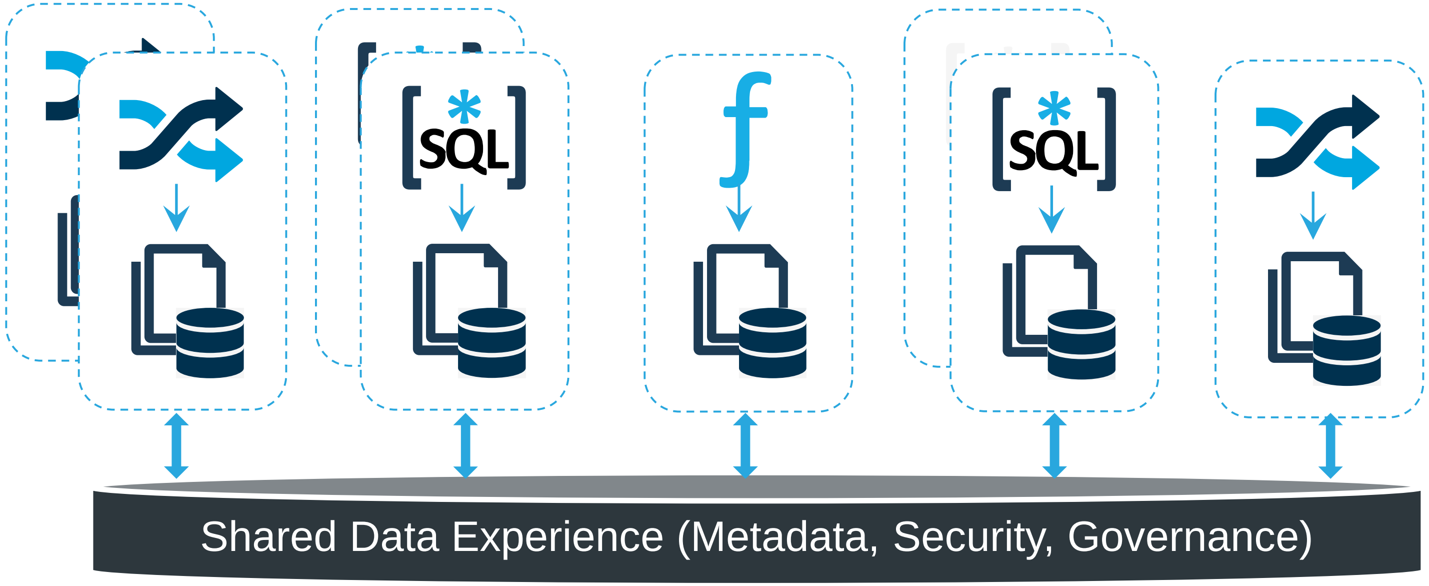
Source: 451 Research, Voice of the Enterprise: Workloads and Key Projects, Cloud Transformation, 2017.

cloudera

My organization is moving to the cloud, why should we consider Cloudera?

Traditional on-premises deployments perform reasonably well

One physical cluster provides a shared data experience to multiple workloads and tenants

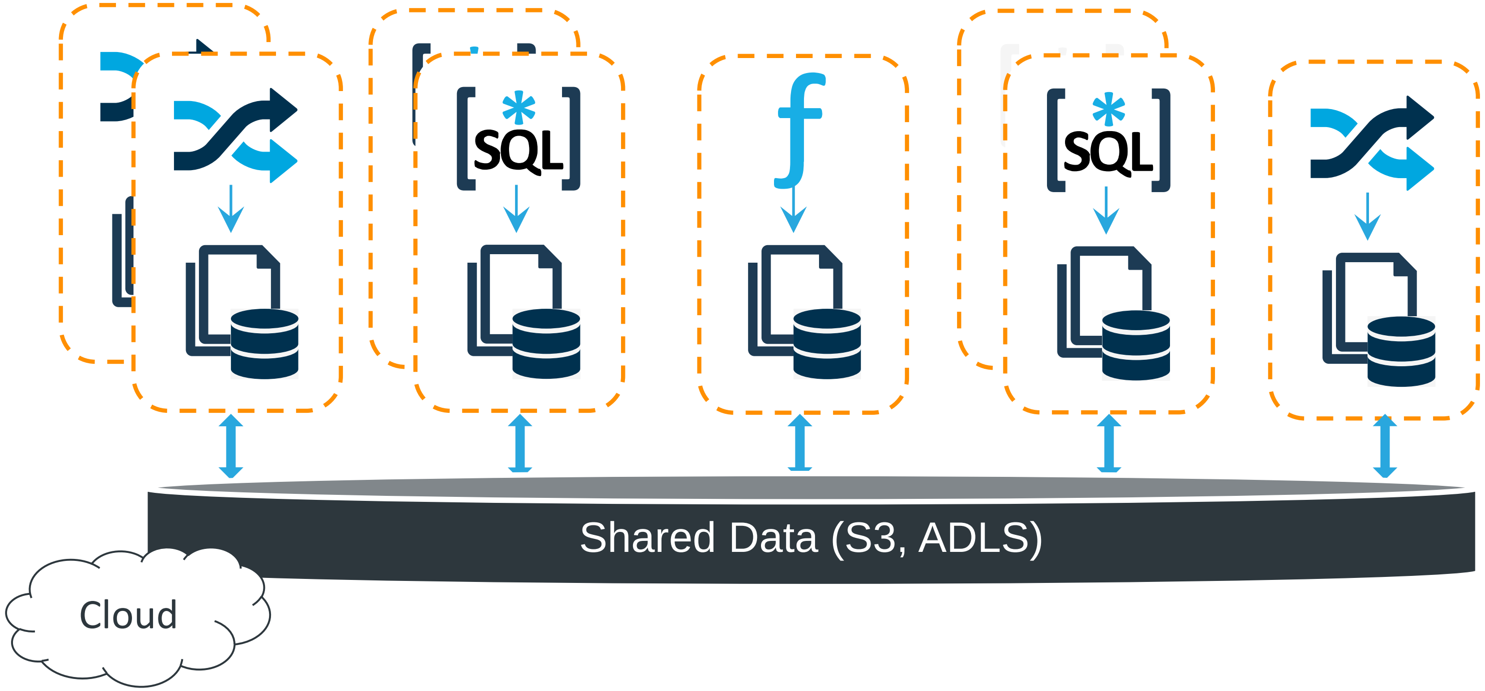


Shared Data Experience (Metadata, Security, Governance)

●	Strong multi-function support
●	Strong shared data experience
●	Strong operational model
◒	Moderate cost management
◒	Moderate tenant isolation
◒	Moderate workload elasticity
○	Weak on self service
○	Weak on speed of deployment

But not good enough for tomorrow

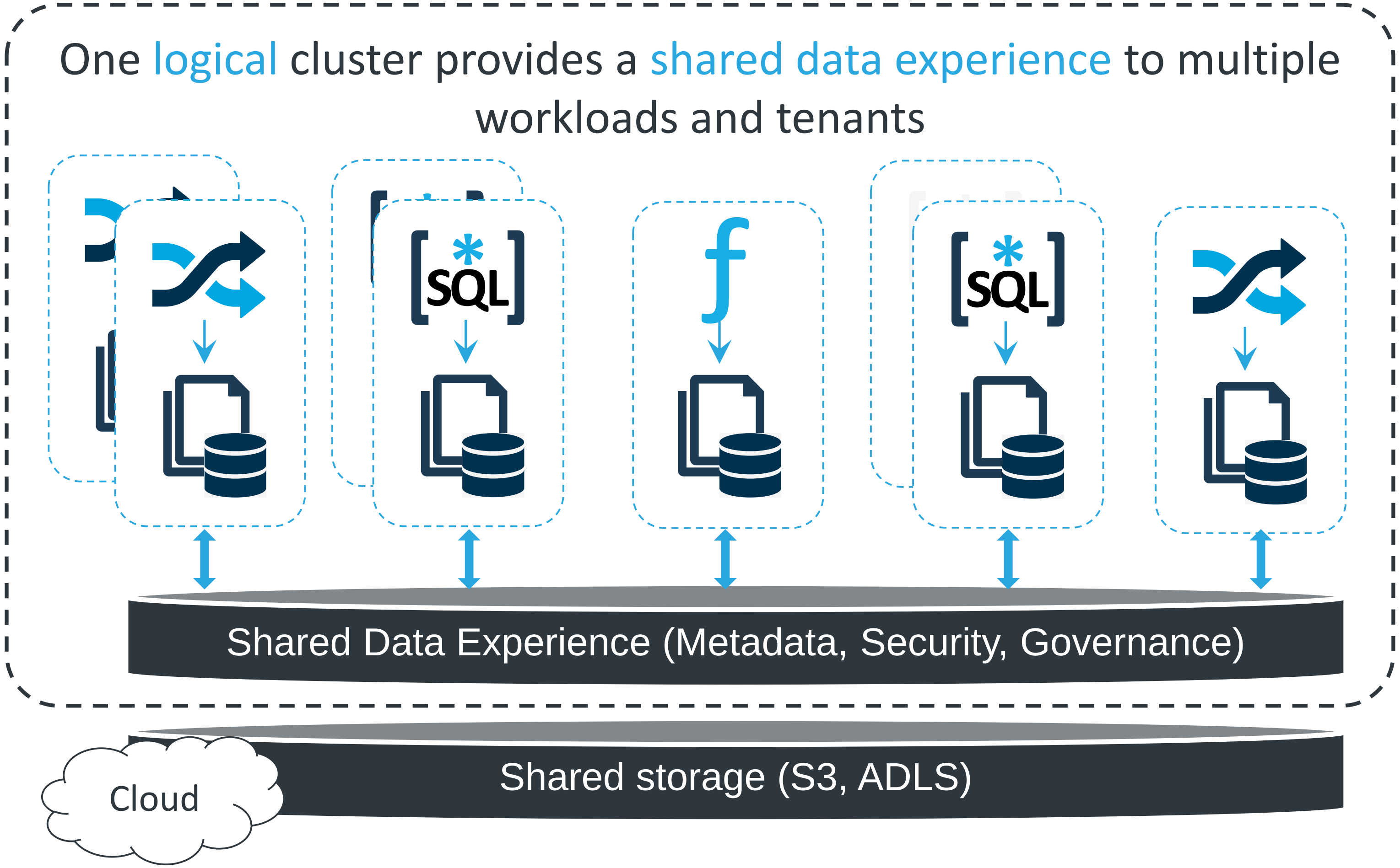
Traditional cloud deployments are strong where on-premises is weak, but at the expense of creating workload silos



◒	Moderate multi-function support
○	Weak on shared data experience
○	Weak operational model
◒	Moderate cost management
●	Strong on tenant isolation
●	Strong on workload elasticity
●	Strong on self service
●	Strong on speed of deployment

This is the experience of cloud house offerings

Only Cloud deployments with SDX optimize for all design goals

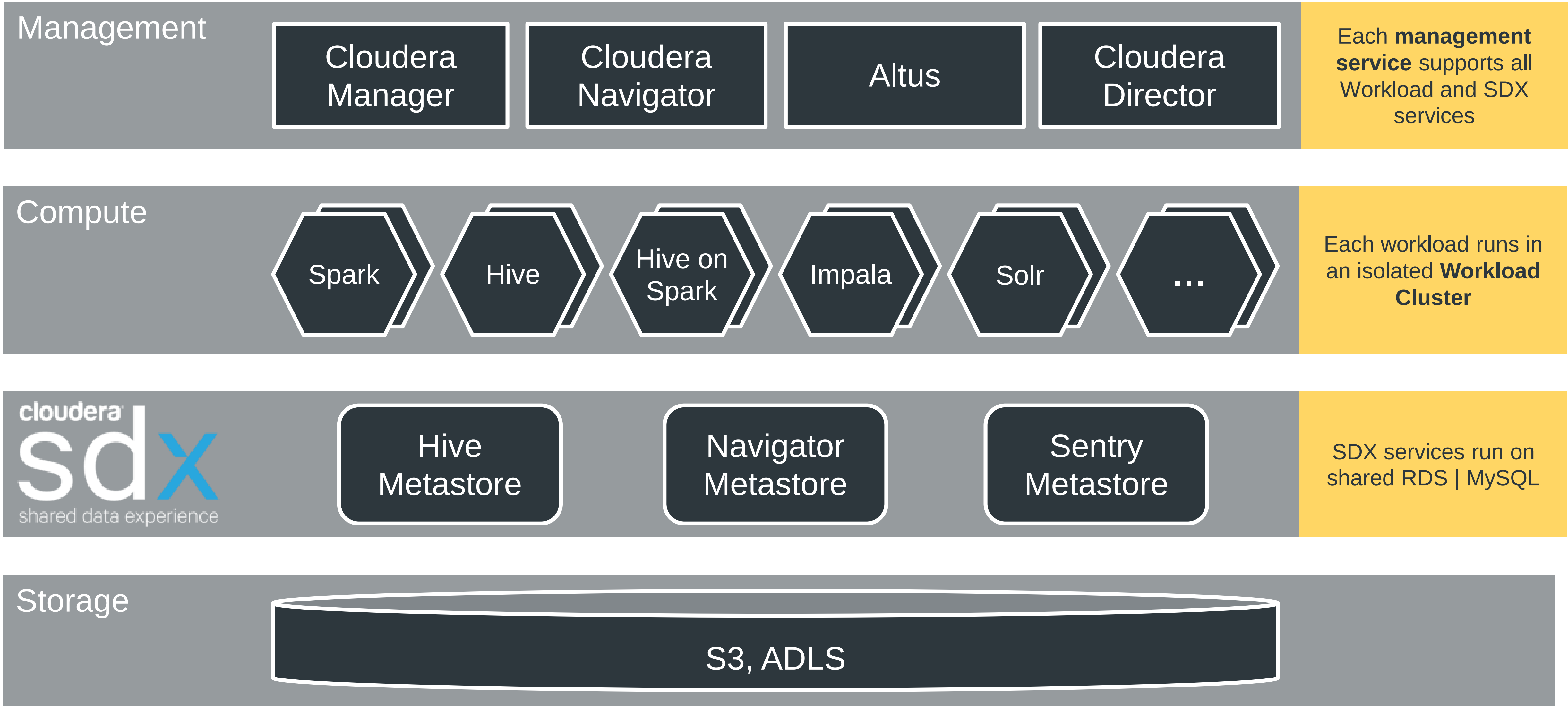

One logical cluster provides a shared data experience to multiple workloads and tenants
SQL
SQL
ƒ
Shared Data Experience (Metadata, Security, Governance)
Shared storage (S3, ADLS)
Cloud

●	Strong multi-function support
●	Strong shared data experience
●	Strong operational model
●	Strong on cost management
●	Strong on tenant isolation
●	Strong on workload elasticity
●	Strong on self service
●	Strong on speed of deployment

SDX makes it possible to transfer on-premises design wins to cloud

cloudera

Cloudera's Public Cloud Reference Architecture



Management

Cloudera Manager

Cloudera Navigator

Altus

Cloudera Director

Each **management service** supports all Workload and SDX services

Compute

Spark

Hive

Hive on Spark

Impala

Solr

...

Each workload runs in an isolated **Workload Cluster**

cloudera sdx shared data experience

Hive Metastore

Navigator Metastore

Sentry Metastore

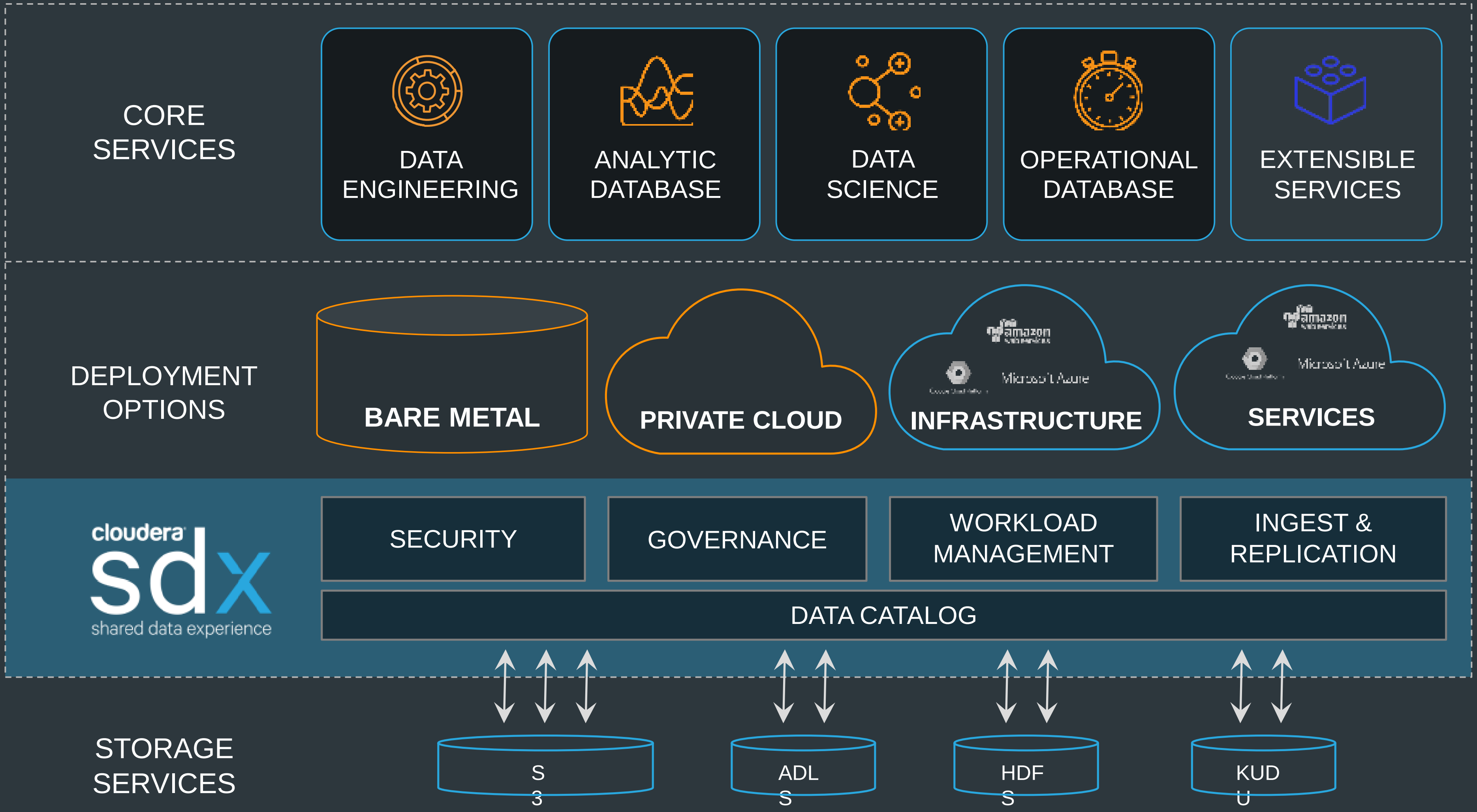
SDX services run on shared RDS | MySQL

Storage

S3, ADLS

Cloudera Enterprise

The modern platform for machine learning and analytics optimized for the cloud



CORE SERVICES

DATA ENGINEERING

ANALYTIC DATABASE

DATA SCIENCE

OPERATIONAL DATABASE

EXTENSIBLE SERVICES

DEPLOYMENT OPTIONS

BARE METAL

PRIVATE CLOUD

INFRASTRUCTURE

SERVICES

cloudera sdx shared data experience

SECURITY

GOVERNANCE

WORKLOAD MANAGEMENT

INGEST & REPLICATION

DATA CATALOG

STORAGE SERVICES

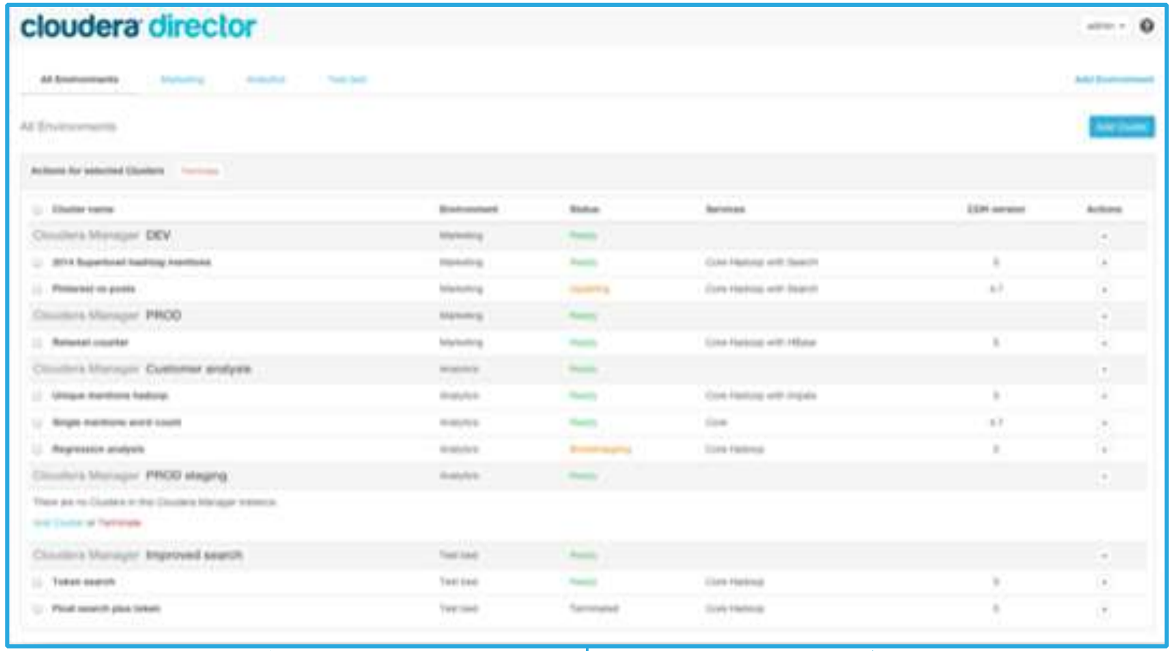
S
3

ADL
S

HDF
S

KUD
U

Cloudera Director for cluster lifecycle management



cloudera director

admin

All Environments | Marketing | Analytics | Test bed | Add Environment

All Environments

Add Cluster

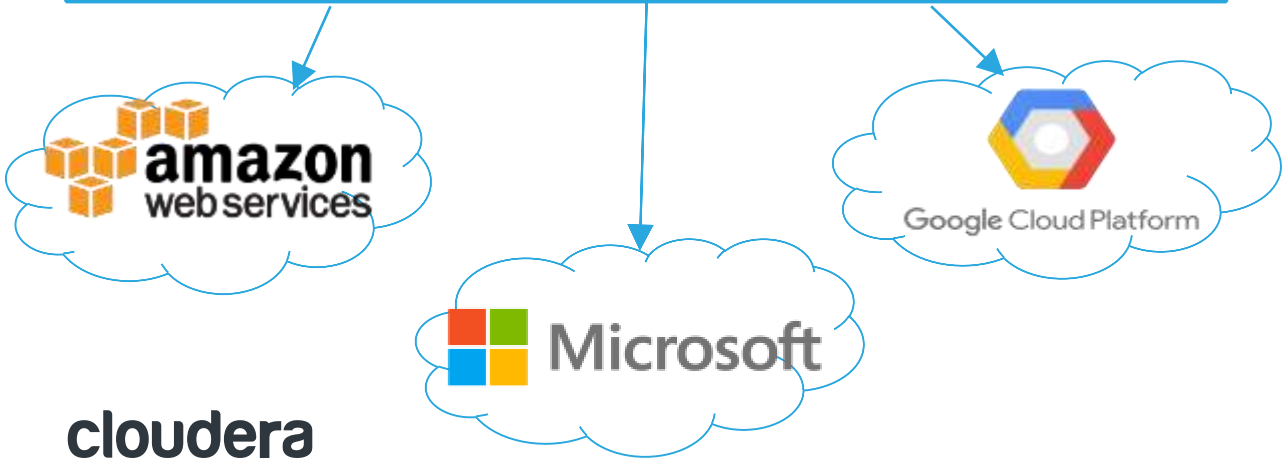
Actions for selected Clusters Terminate

□ Cluster name	Environment	Status	Services	CDH version	Actions
Cloudera Manager DEV	Marketing	Ready			▾
□ 2014 Superbowl hashtag mentions	Marketing	Ready	Core Hadoop with Search	5	▾
□ Pinterest re-posts	Marketing	Updating	Core Hadoop with Search	5.7	▾
Cloudera Manager PROD	Marketing	Ready			▾
□ Retweet counter	Marketing	Ready	Core Hadoop with HBase	5	▾
Cloudera Manager Customer analysis	Analytics	Ready			▾
□ Unique mentions hadoop	Analytics	Ready	Core Hadoop with Impala	5	▾
□ Single mentions word count	Analytics	Ready	Core	5.7	▾
□ Regression analysis	Analytics	Bootstrapping	Core Hadoop	5	▾
Cloudera Manager PROD staging	Analytics	Ready			▾

There are no Clusters in this Cloudera Manager instance.

Add Cluster or Terminate

□ Cluster name	Environment	Status	Services	CDH version	Actions
Cloudera Manager Improved search	Test bed	Ready			▾
□ Token search	Test bed	Ready	Core Hadoop	5	▾
□ Pivot search plus token	Test bed	Terminated	Core Hadoop	5	▾



Easy

- Single pane of glass for all cloud infrastructure
- Create templates to run applications in a pre-optimized manner

Flexible

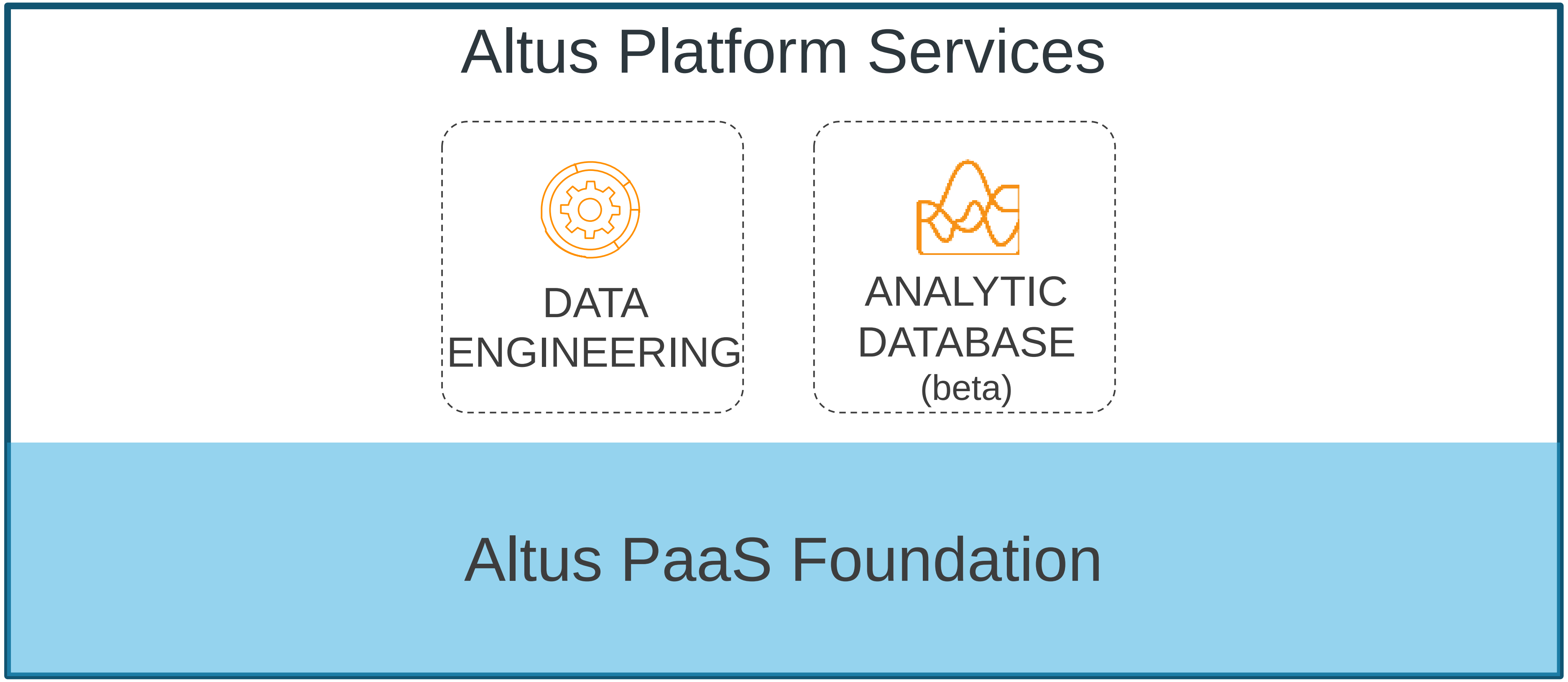
- Multi-cloud: AWS, Azure, GCP
- Hourly pricing with auto billing & metering
- Spot instance/block support

Enterprise-grade

- Integration across Cloudera Enterprise
- Management of CDH deployments at scale
- Deeply integrated with Cloudera Manager

Cloudera Altus

Multi-cloud foundation for building new cloud services



Altus Platform Services

DATA ENGINEERING

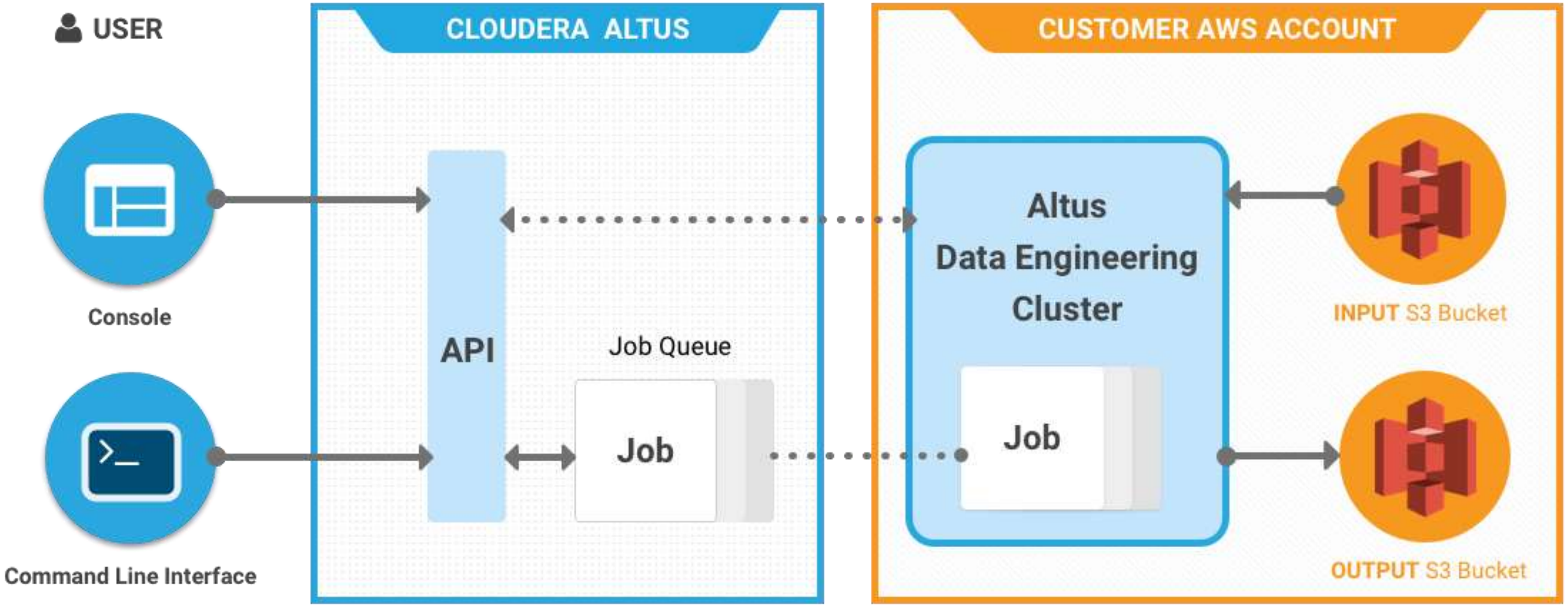
ANALYTIC DATABASE (beta)

Altus PaaS Foundation

Director vs Altus - when to use either?

	Altus	Director
Automated log saving	x	
Automated Cluster spin up / down (no extra coding)	x	
Data Engineering – Hive, Spark, HoS, MR	x	x
Production Job Driven	x	
Workload Analytics	x	
Cluster Duration	Purely Transient	Transient OR Persistent
Job Development / Exploration		x
3rd Party Installations		x
Full Control of CM		x
Analytical / Operational - Impala, HBase, Search		x
Persistent (or Transient)		x
Grow / Shrink Cluster		x

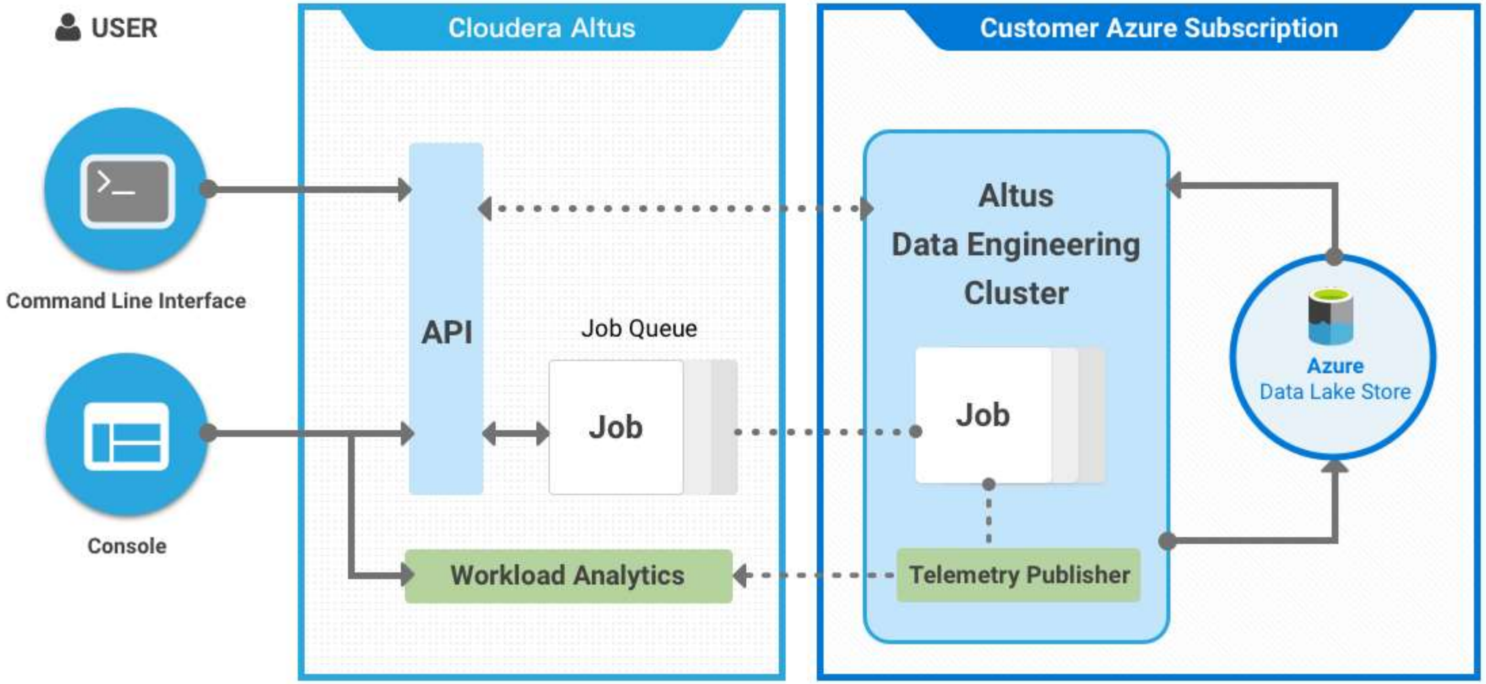
Altus Service Architecture (AWS)



- Runs in Cloudera’s secured and monitored environment
- Manages CDH clusters in customer cloud account
- Customer data does not pass to Cloudera (Workload Analytics requires opt-in log data transfer to Cloudera)

cloudera

Altus Service Architecture (Azure)

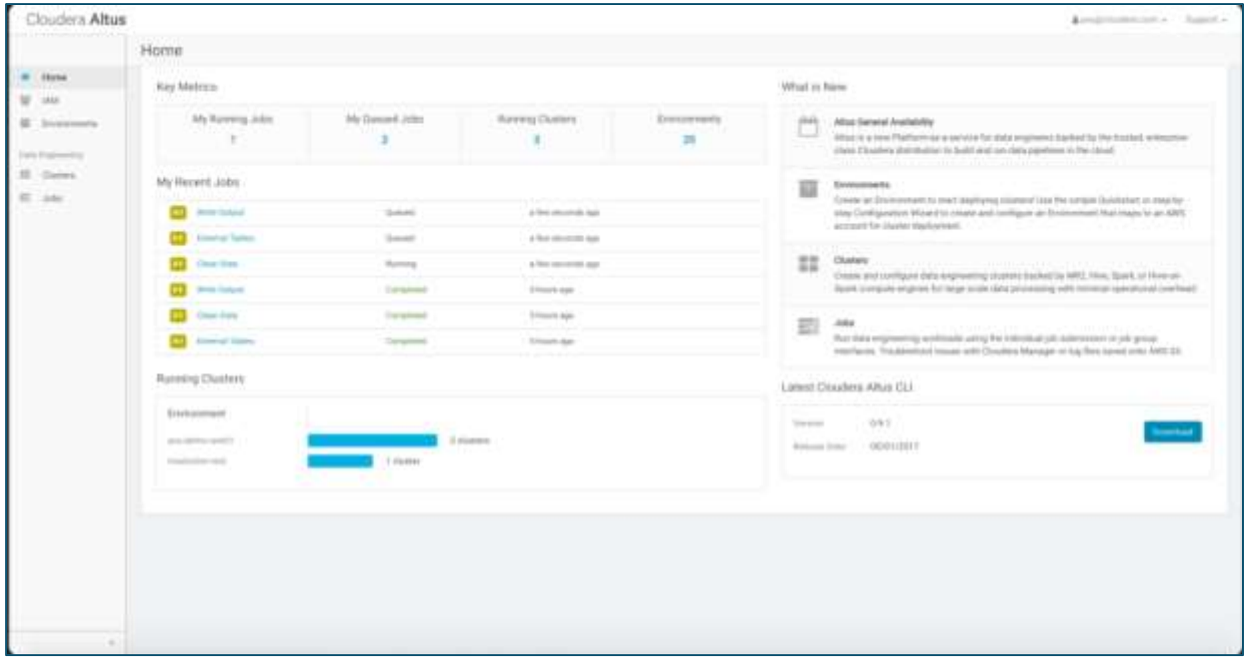


- Runs in Cloudera’s secured and monitored environment
- Manages CDH clusters in customer cloud account
- Customer data does not pass to Cloudera (Workload Analytics requires opt-in log data transfer to Cloudera)

Altus Data Engineering

for ETL, machine learning, and data processing

- Fast, easy job submission without the cluster management
- Built-in Workload Analytics for troubleshooting and optimization
- Lower costs with transient resources and pay-per-use pricing
- Full benefits of isolation + Shared Data Experience



Cloudera Altus

Home

Key Metrics

My Running Jobs	My Queued Jobs	Running Clusters	Environments
1	3	3	20

My Recent Jobs

Write Output	Queued	a few seconds ago
External Tables	Queued	a few seconds ago
Clean Data	Running	a few seconds ago
Write Output	Completed	3 hours ago
Clean Data	Completed	3 hours ago
External Tables	Completed	3 hours ago

Running Clusters

Environment	
[illegible]	2 clusters
[illegible]	1 cluster

What is New

Altus General Availability
Altus is a new Platform-as-a-service for data engineers backed by the trusted, enterprise-class Cloudera distribution to build and run data pipelines in the cloud.

Environments
Create an Environment to start deploying clusters! Use the simple Quickstart or step-by-step Configuration Wizard to create and configure an Environment that maps to an AWS account for cluster deployment.

Clusters
Create and configure data engineering clusters backed by MR2, Hive, Spark, or Hive-on-Spark compute engines for large scale data processing with minimal operational overhead.

Jobs
Run data engineering workloads using the individual job submission or job group interfaces. Troubleshoot issues with Cloudera Manager or log files saved onto AWS S3.

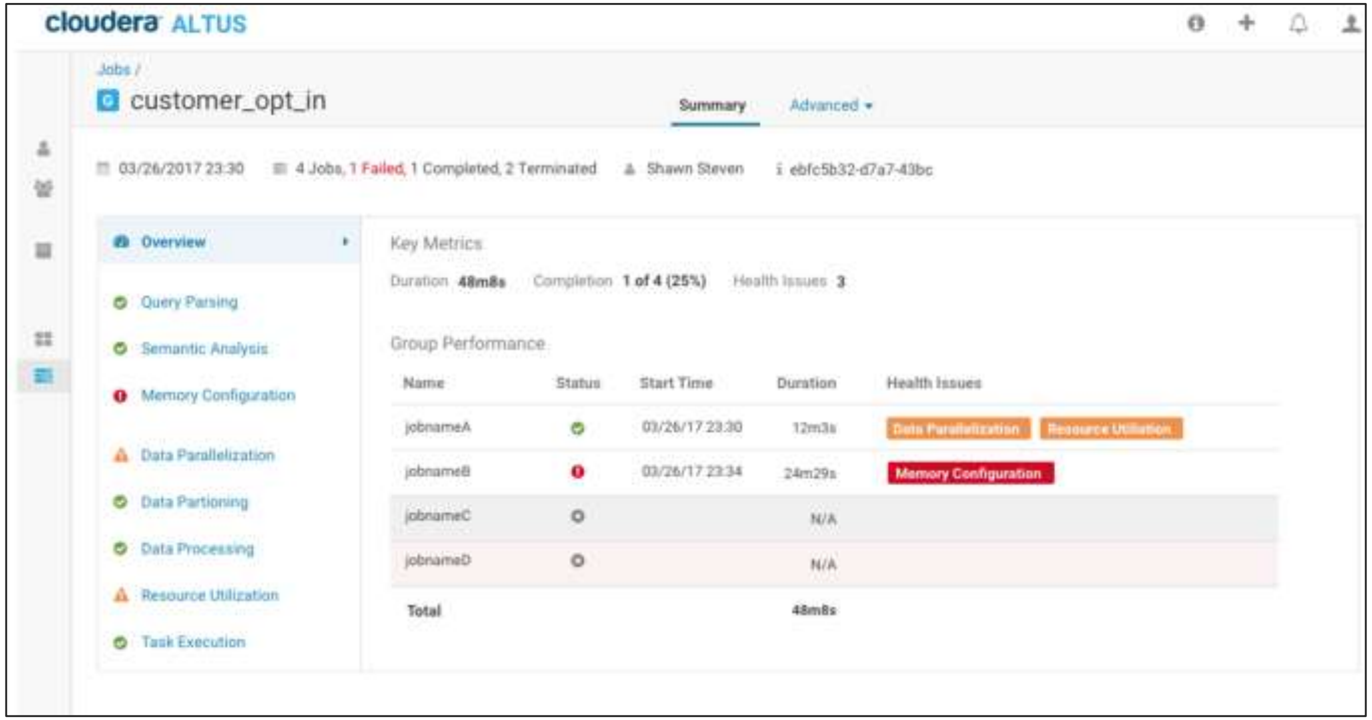
Latest Cloudera Altus CLI

Version: 0.9.1
Release Date: [illegible]

End-user focused with jobs as first-class objects

Workload troubleshooting and analytics

- Troubleshoot jobs after cluster termination through job log and configuration browsing
- Insight into causes of job failure
- Identification and root cause analysis of slow jobs



Jobs /

customer_opt_in

03/26/2017 23:30   4 Jobs, 1 Failed, 1 Completed, 2 Terminated   Shawn Steven   ebfc5b32-d7a7-43bc

- Overview
- ✓ Query Parsing
- ✓ Semantic Analysis
- ! Memory Configuration
- ⚠ Data Parallelization
- ✓ Data Partioning
- ✓ Data Processing
- ⚠ Resource Utilization
- ✓ Task Execution

Key Metrics

Duration **48m8s**   Completion **1 of 4 (25%)**   Health Issues **3**

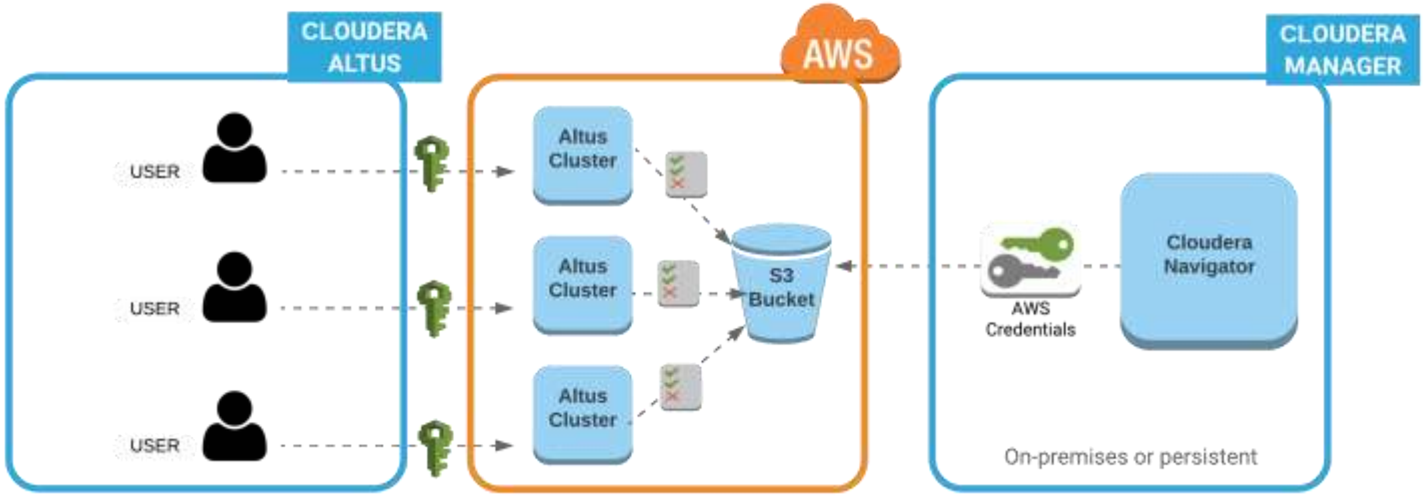
Group Performance

Name	Status	Start Time	Duration	Health Issues
jobnameA	✓	03/26/17 23:30	12m3s	Data Parallelization Resource Utilization
jobnameB	!	03/26/17 23:34	24m29s	Memory Configuration
jobnameC	○		N/A	
jobnameD	○		N/A	
Total			**48m8s**	

Capture metadata spanning multiple clusters

Persist metadata with Cloudera Navigator

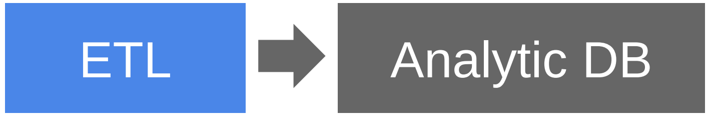
- Export metadata and lineage information from Altus clusters
- Insight into full data management pipeline including transient clusters



cloudera

Three immediate use cases for Altus Data Engineering

ETL FOR ANALYTIC DB

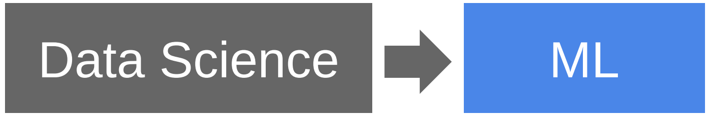


ETL

Analytic DB

Cloud-native batch preparation for Impala on IaaS or, soon, Altus Analytic DB.

BATCH MACHINE LEARNING

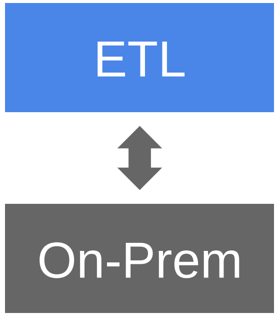


Data Science

ML

Scalable compute for massively-parallel batch machine learning training, scoring, or simulation.

ETL OFFLOAD



ETL

On-Prem

Offload batch processing jobs from overburdened on-premises clusters.

Altus Analytic Database

The first data warehouse cloud service to bring the warehouse to the data - delivering instant analytics to anyone

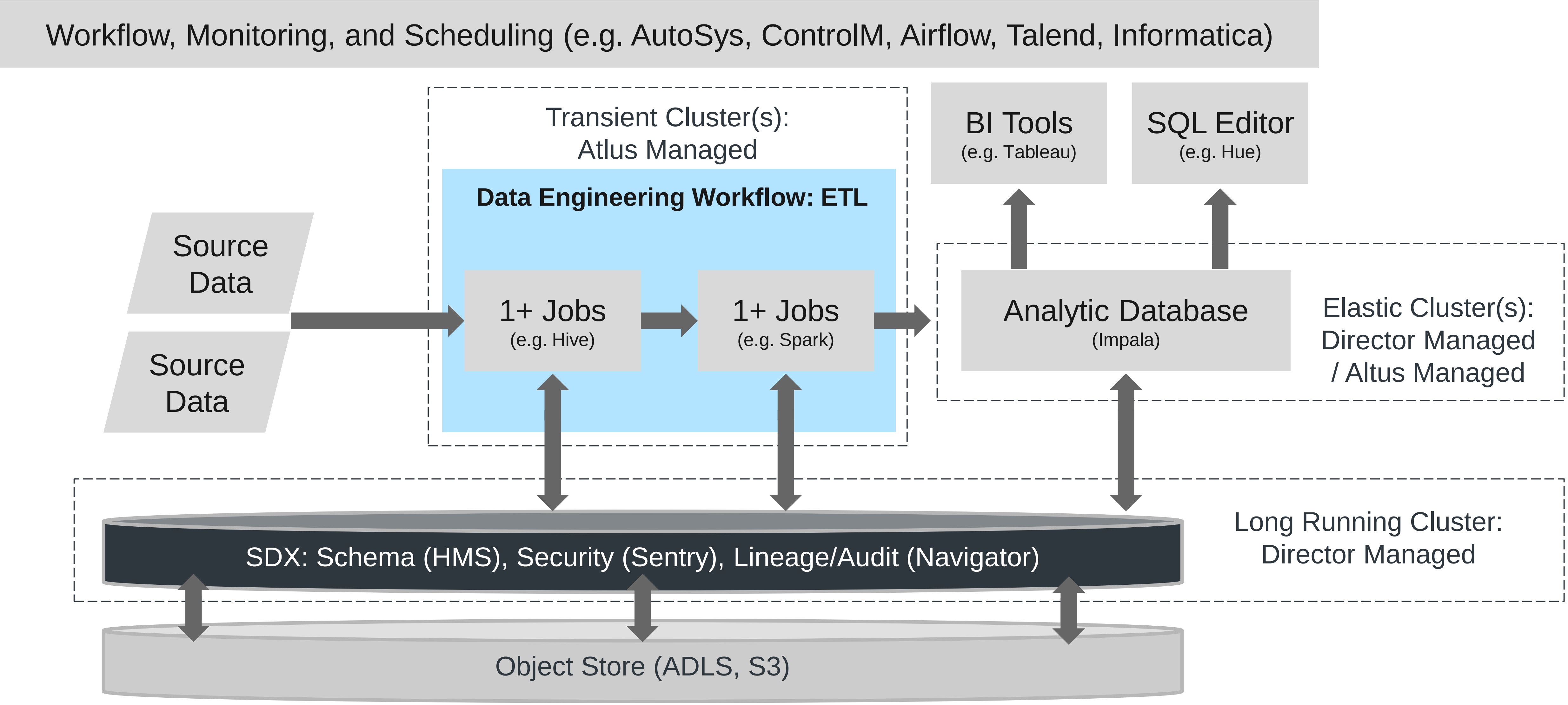
For business analysts:

- Query with predictable performance, at any time, without risking SLAs
- Bring limitless new users and use cases with instant self-service analytic access
- Data available for broad access (SQL, BI tools, Python, R, etc)

For IT:

- Easily and elastically provision isolated resources as and when they're needed
- Simple multi-tenant management including federated identity and consistent governance
- Eliminate data movement and copies across workloads

ETL to Cloud-Native Analytic Database



cloudera

The Scenario

My Role: Data Analyst at DataCo – a Sports Retailer

Business Issue: Experiencing lower than expected website sales. Why?

Technical Issues: I have a data warehouse on premise, which contains my sales
order data, but it is very old and slow and it is difficult to do ad hoc
queries on it.

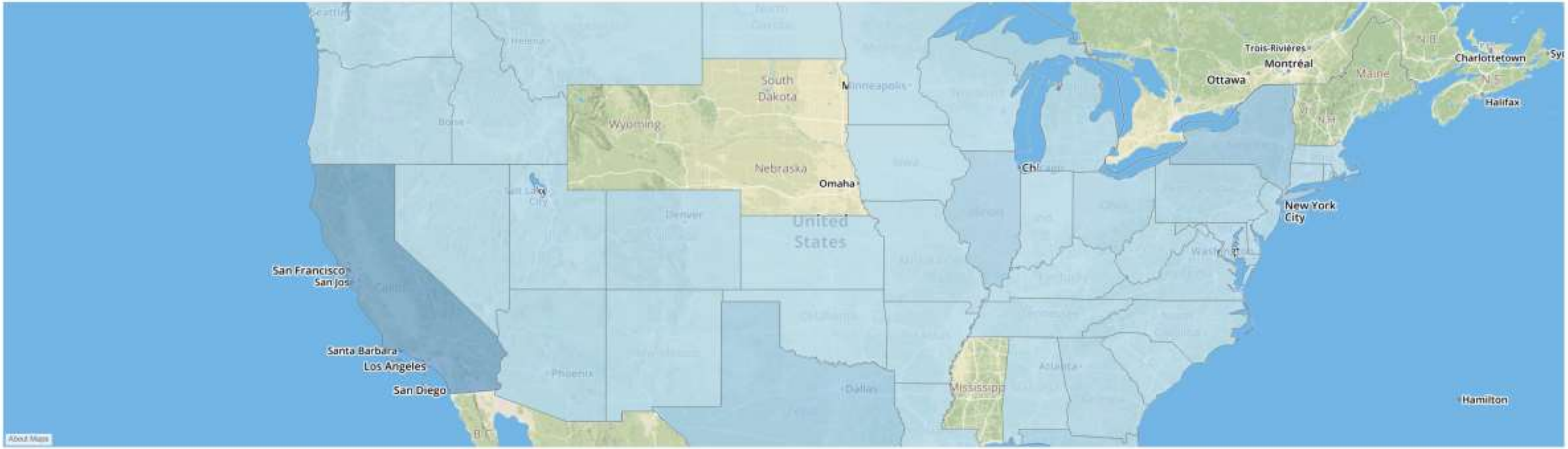
My clickstream data is too big to ingest into my data warehouse

Requirements: Need an Analytic Database to do ad hoc queries on order data

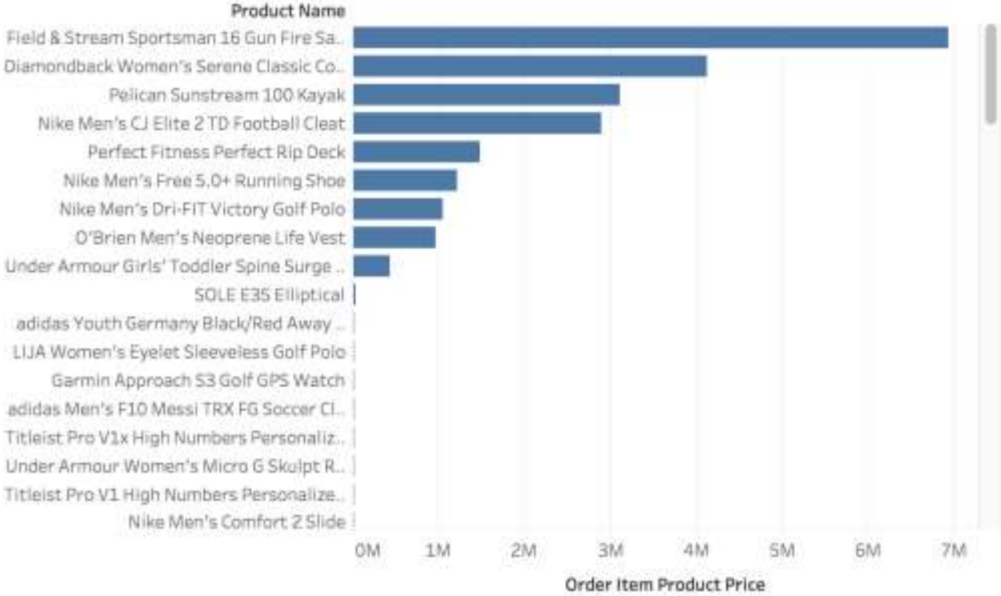
Need a temporary platform to process weblogs once a day

Ability to join processed weblogs to order data

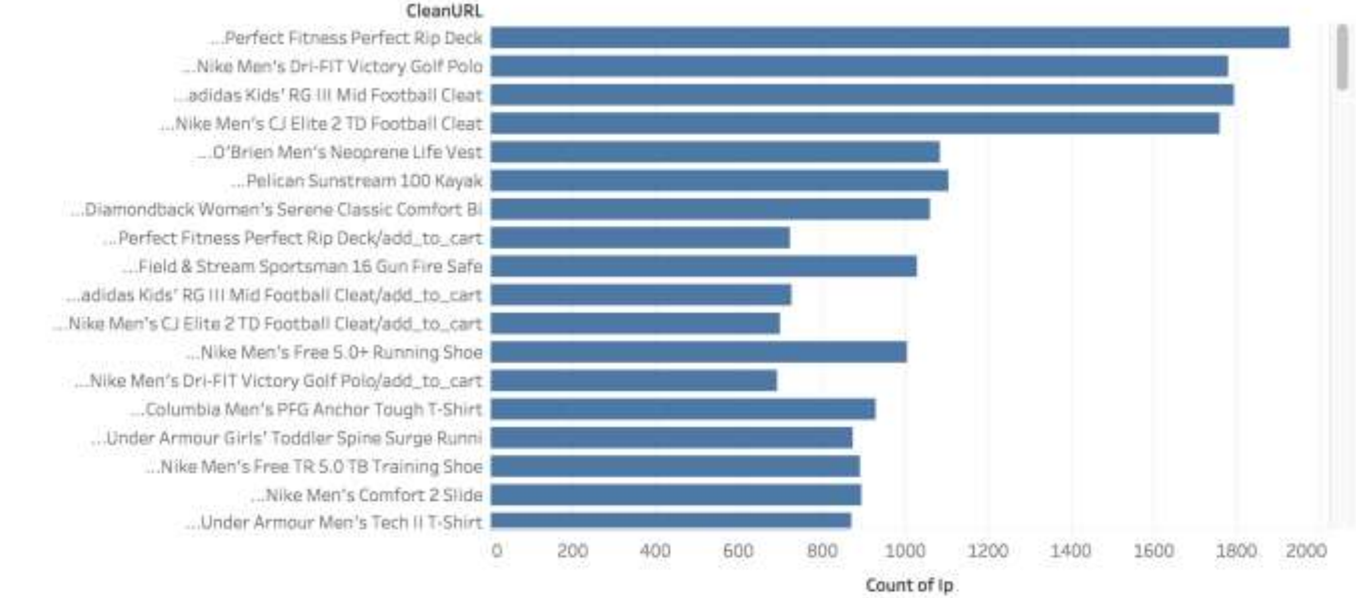
Sales by State



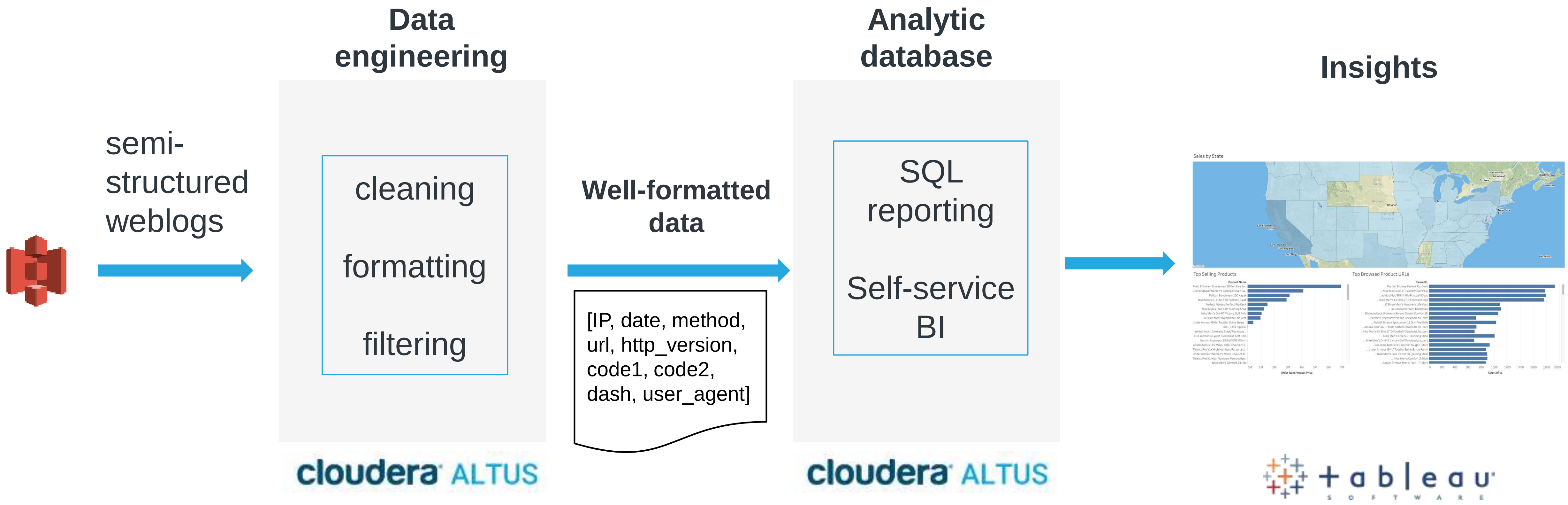
Top Selling Products



Top Browsed Product URLs



Demo - Retail clickstream analysis



semi-structured weblogs

Data engineering

cleaning

formatting

filtering

cloudera ALTUS

Well-formatted data

[IP, date, method, url, http_version, code1, code2, dash, user_agent]

Analytic database

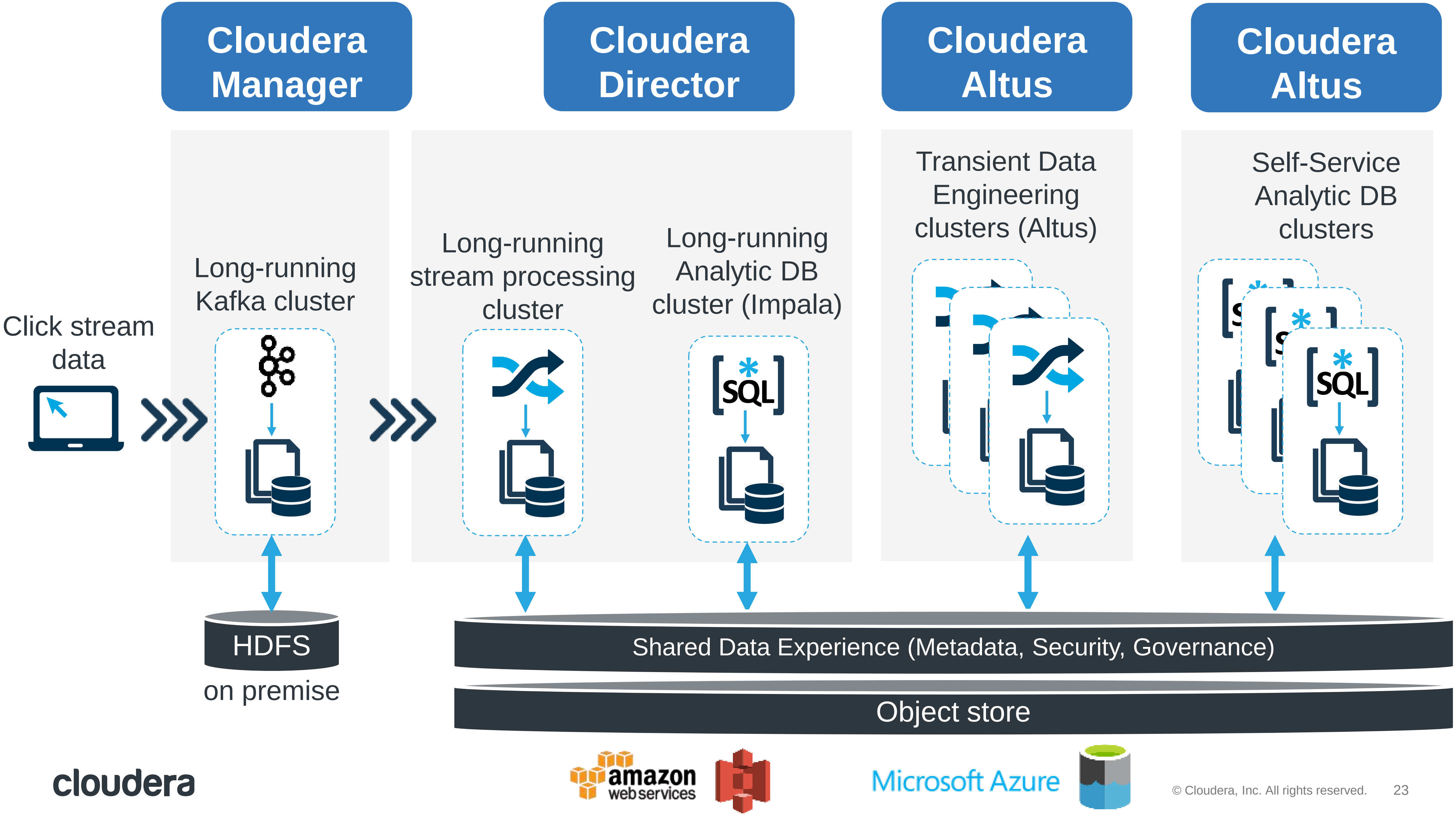
SQL reporting

Self-service BI

cloudera ALTUS

Insights

cloudera



Cloudera Manager

Cloudera Director

Cloudera Altus

Cloudera Altus

Click stream data

Long-running Kafka cluster

Long-running stream processing cluster

Long-running Analytic DB cluster (Impala)

Transient Data Engineering clusters (Altus)

Self-Service Analytic DB clusters

HDFS

on premise

Shared Data Experience (Metadata, Security, Governance)

Object store

cloudera

Q&A

Resources

Cloudera Altus

Cloudera Altus documentation

Cloudera Director

Cloudera Director documentation

Cloudera SDX

Try Cloudera Director on Microsoft Azure

Try Cloudera Director on AWS with AWS Quickstart

Cloudera Reference Architectures for public and private cloud deployments

cloudera