



Deploying Cloudera in the Cloud

Wim Villano, Sales Engineer Cloudera

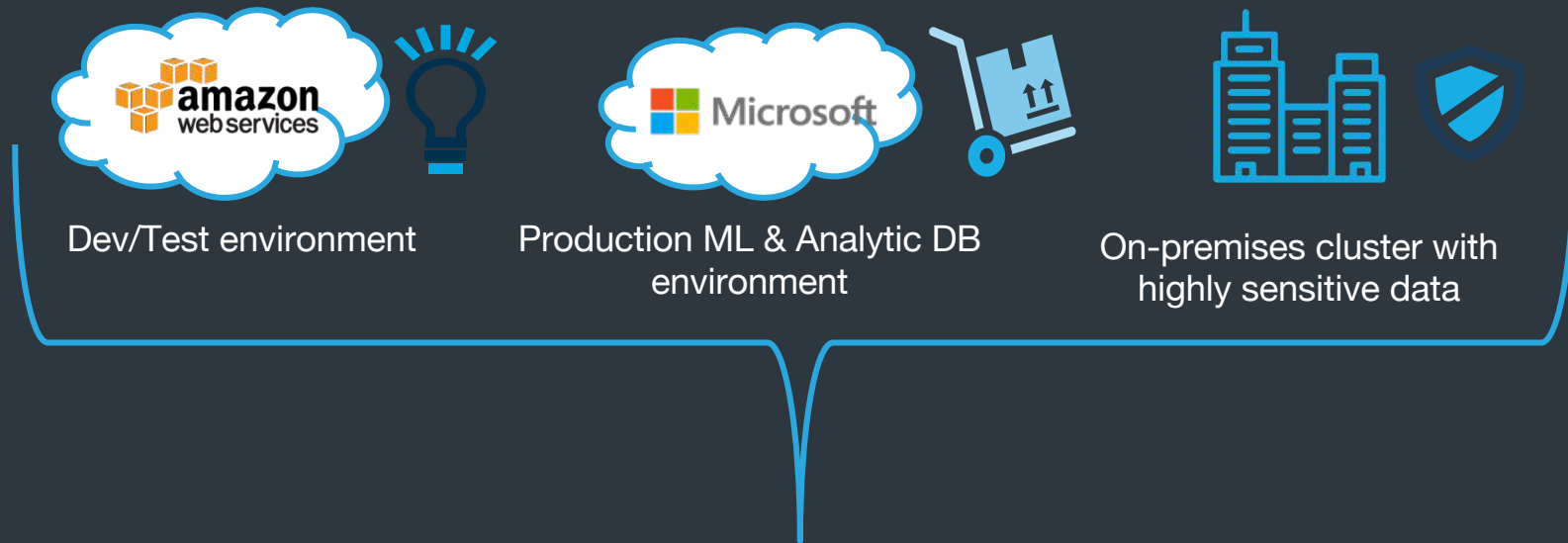
We believe

We believe your workloads
should run *wherever*
they gain the greatest advantage

cloudera

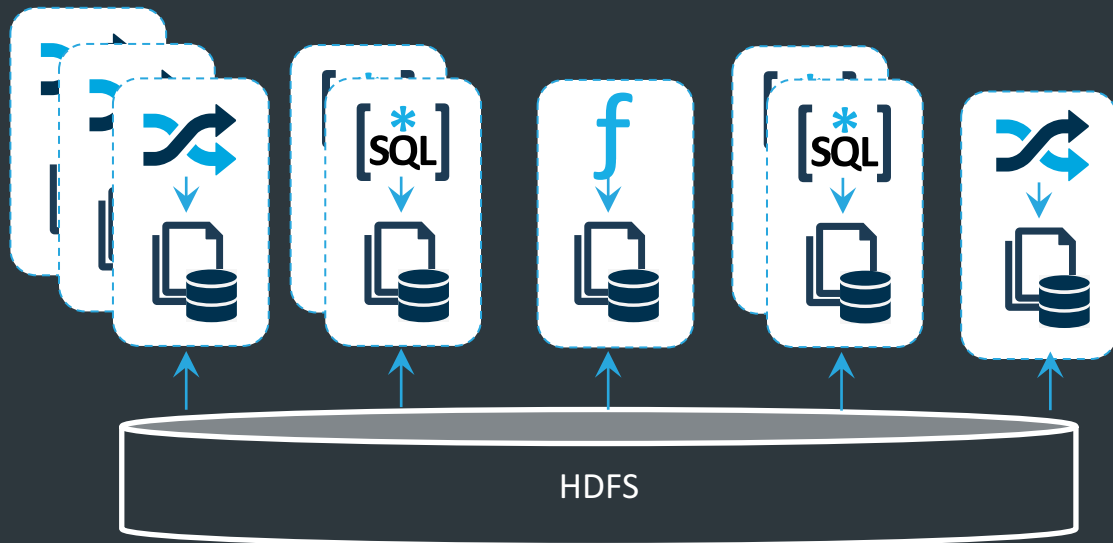


Hybrid/Multi-cloud architecture example

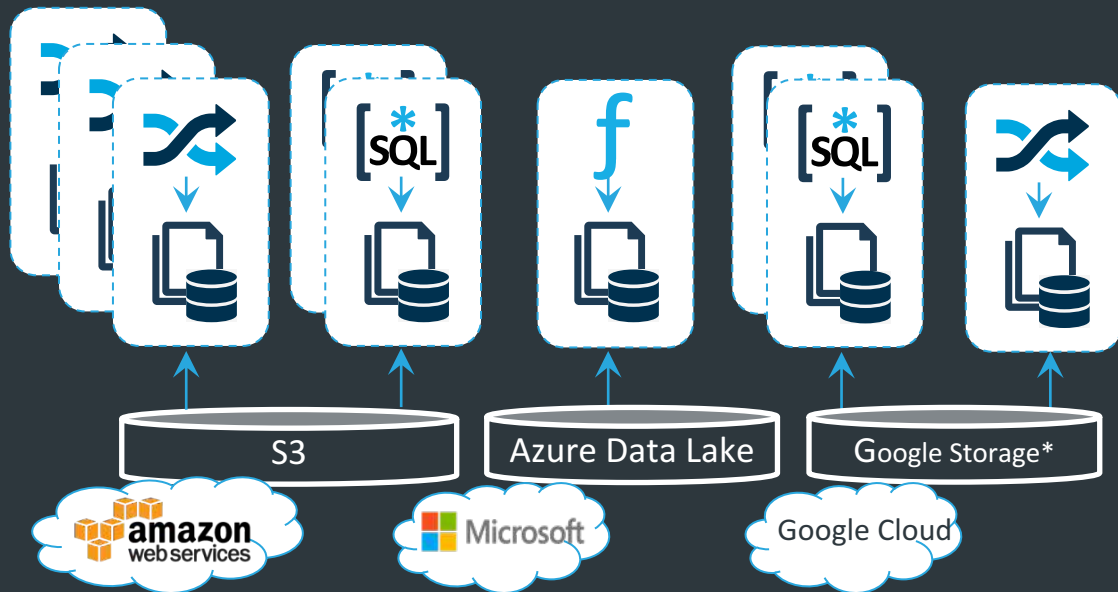


Applications and processes function the same in each environment

Traditional On-prem Workloads Share a Single Cluster



Cloud Leads to Separate Specialized Clusters



What is Cloudera Director?

A tool that customers can use to deploy & manage the lifecycle of Cloudera Enterprise clusters in the cloud

Spin up, grow and shrink, terminate
Unified view and management of cloud environments



Fast Cloud-native Deployments

Spin up, grow & shrink, terminate Cloudera clusters that read/write to object store

Easy Administration

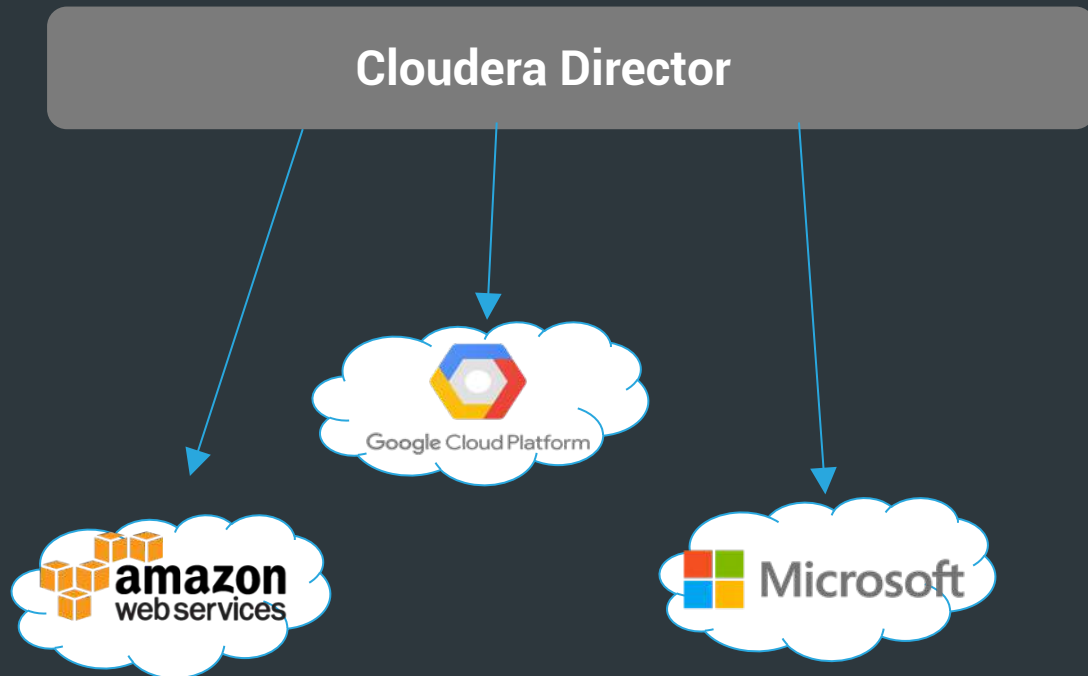
- Dynamic cluster lifecycle management
- Single pane of glass: multi-cluster view
- Create templates to run workloads in a pre-optimized manner

Flexible Deployments

- Multi-cloud: AWS, Azure, GCP
- Fast cluster deployments
- Scaling of CDH clusters
- Spot instance support
- Support for EBS volumes and various Azure storage formats

Enterprise-grade

- Integration across Cloudera Enterprise
- Management of CDH deployments at scale



Cloudera Altus™

Platform-as-a-service

Now it's easier than ever to process big data in the cloud.

Learn more at cloudera.com/altus

- Easy
- Agile
- Unified



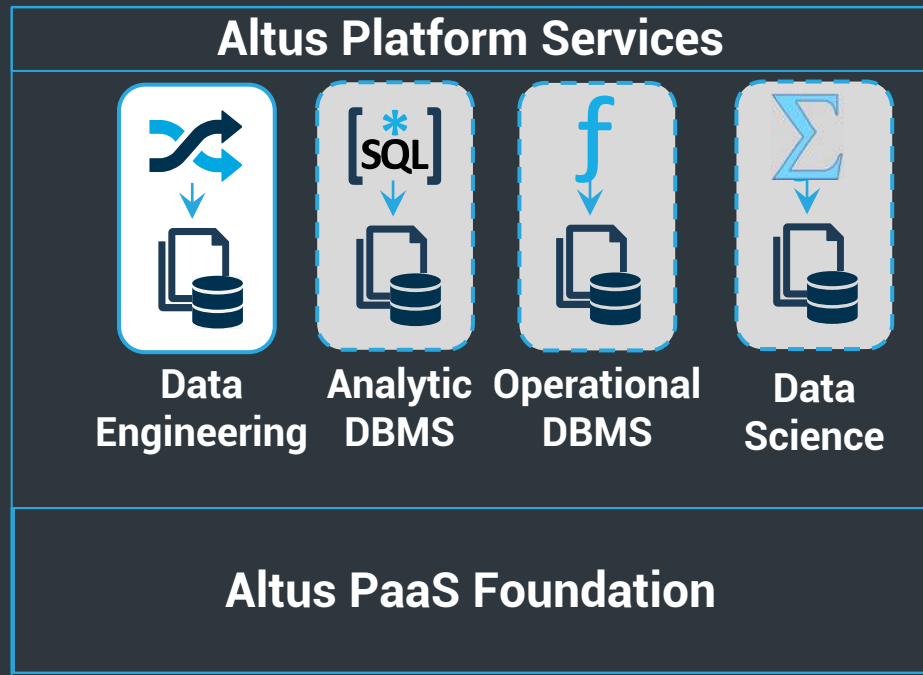
Everything you don't have to do

- Install any software to start working
- Install any hardware
- Worry about cluster configuration
- Upgrade/reconfigure clusters
- OS upgrades/patching
- Resource Management



Cloudera Altus is a PaaS for big data analytics

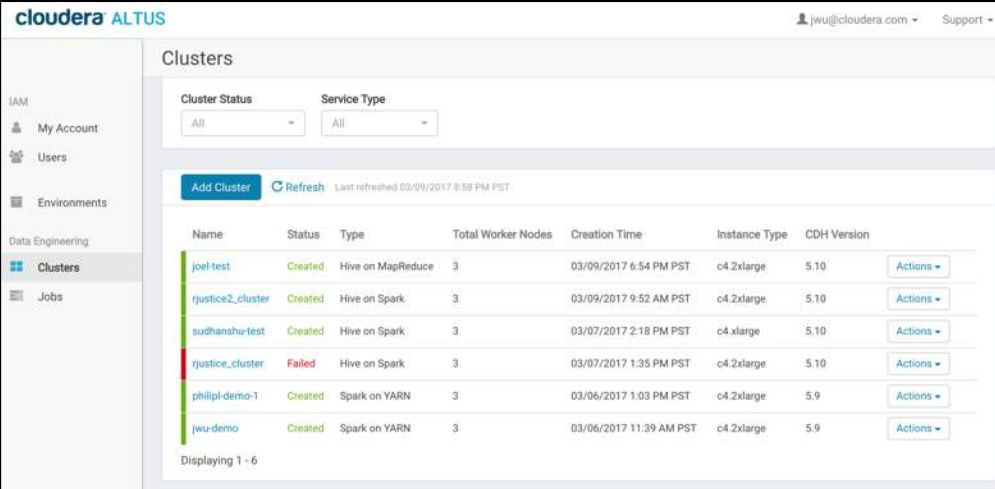
- Altus is umbrella brand for all of Cloudera's PaaS offerings.
- PaaS foundation acts as framework for building services.
- Altus for data engineers is first user-facing service.



Cloudera Altus for data engineering workloads

Managed service for workloads such as ETL, machine learning, and data processing:

- Managed transient CDH clusters in customer VPC
- Support for MR2, Hive, Spark, Hive-on-Spark
- Single pipeline batch processing with data on AWS S3 or Azure Data Lake



cloudera ALTUS

Cluster Status: All Service Type: All

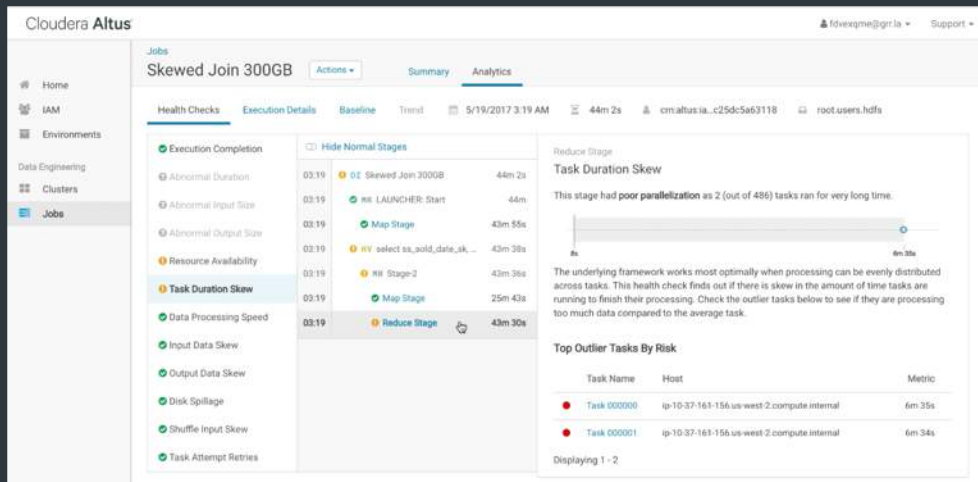
[Add Cluster](#) [Refresh](#) Last refreshed 03/09/2017 8:58 PM PST

Name	Status	Type	Total Worker Nodes	Creation Time	Instance Type	CDH Version	Actions
joel-test	Created	Hive on MapReduce	3	03/09/2017 6:54 PM PST	c4.2xlarge	5.10	Actions
rjustice2_cluster	Created	Hive on Spark	3	03/09/2017 9:52 AM PST	c4.2xlarge	5.10	Actions
sudhanshu-test	Created	Hive on Spark	3	03/07/2017 2:18 PM PST	c4.xlarge	5.10	Actions
rjustice_cluster	Failed	Hive on Spark	3	03/07/2017 1:35 PM PST	c4.2xlarge	5.10	Actions
philipi-demo-1	Created	Spark on YARN	3	03/06/2017 1:03 PM PST	c4.2xlarge	5.9	Actions
jwu-demo	Created	Spark on YARN	3	03/06/2017 11:39 AM PST	c4.2xlarge	5.9	Actions

Displaying 1 - 6

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



Altus feature overview

End-user focused

- Manages your cluster so you don't have to
- Job submission CLI/API
- Workload troubleshooting

Cloud-native

- Decouple storage and compute
- R/W Amazon S3/Azure ADLS
- Spin EC2 clusters up and down

Low cost

- Per-node/per-hour pricing
- Terminate clusters when not in use
- Spot with self-healing

Integrated Platform

- Same Cloudera platform on-premises and in the cloud
- Feed cleaned data into Impala clusters for BI analytics
- Share metadata across clusters

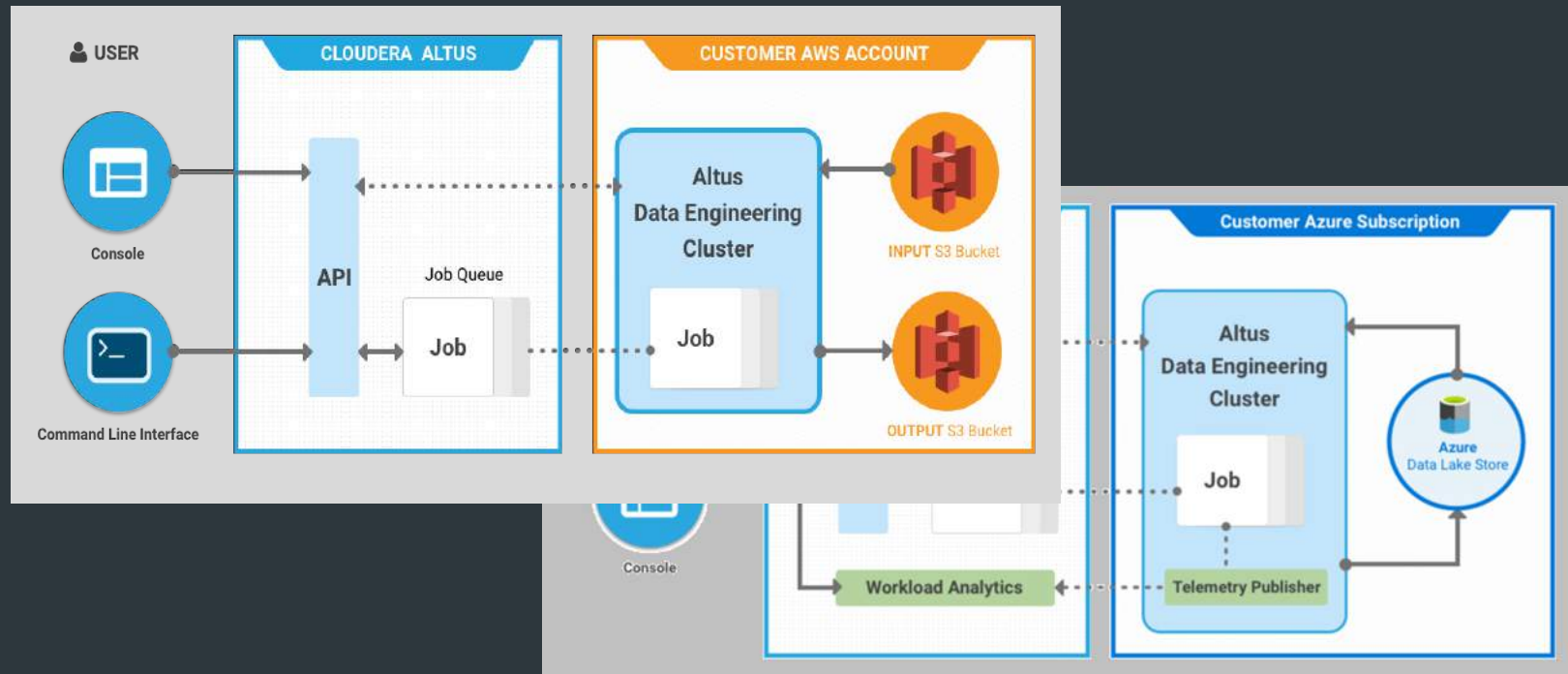
Secure

- Integrated with AWS security
- No Cloudera access to customer data
- Support for multi-AWS accounts
- Admin and end-user accounts

Easy to use

- Self-service for end-users
- Cloud console + familiar tools
- Cluster provisioning in minutes

Altus Service Architecture



Altus & Director – when to use which?

	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
3 rd Party Installations		x
Full Control of CM		x
Analytical / Operational - Impala, HBase, Search		x
Persistent (or Transient)		x
Grow / Shrink Cluster		x

Get started



Cloudera Altus

Convenient, easy-to-use, fully integrated PaaS for ETL and pipeline development



Cloudera Director

Tool for managing long-running Cloudera clusters in the cloud environment of your choice



Azure Marketplace

Provision and deploy Cloudera on the Azure Marketplace

Altus Demo

The Scenario

My Role: Data Analyst at DataCo – a Sports Retailer

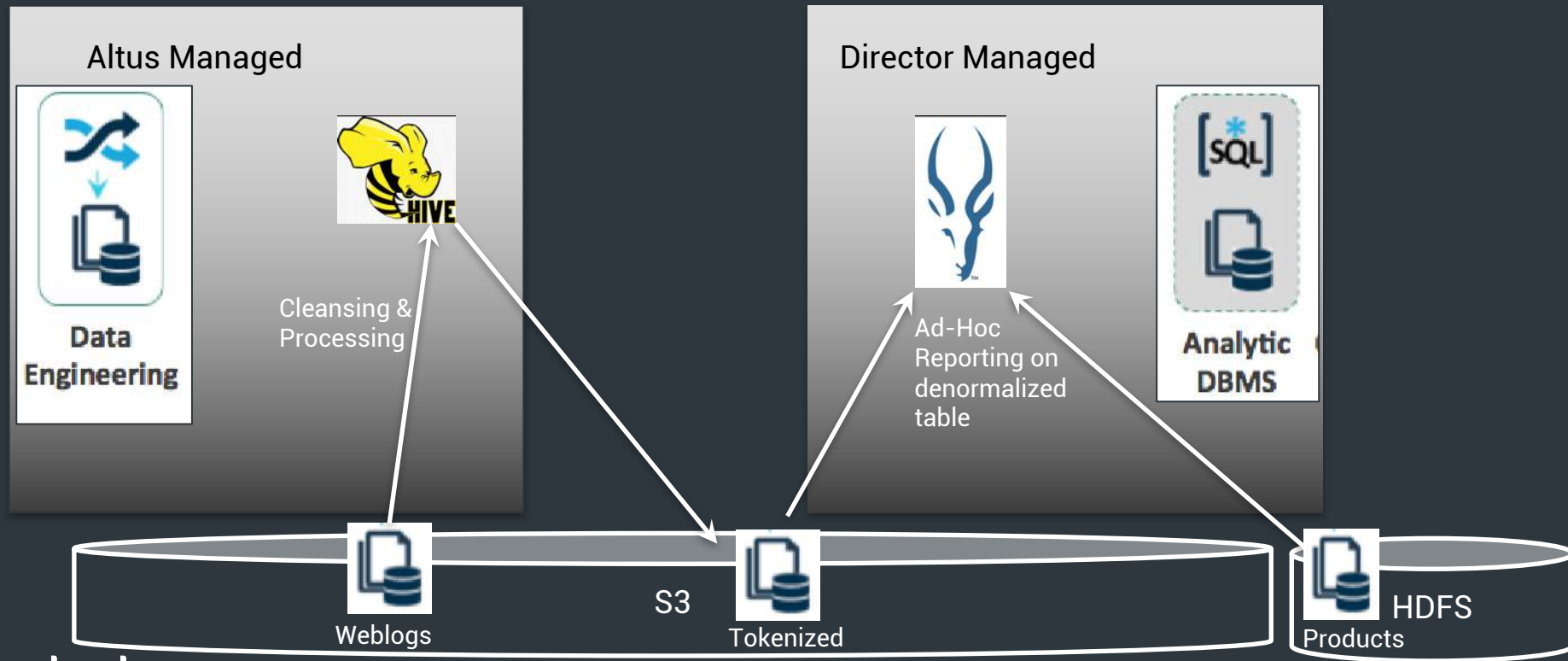
Business Issue: Experiencing lower than expected website sales. Why?

Technical Issues: Current cluster resources are fully consumed querying sales data
Limited budget for weblog analysis - not enough for expansion

Requirements: Just need a temporary platform to process weblogs once a day
Ability to join processed weblogs to order data

Cloudera Altus for data engineering workloads

Director for Impala Interactive Workloads



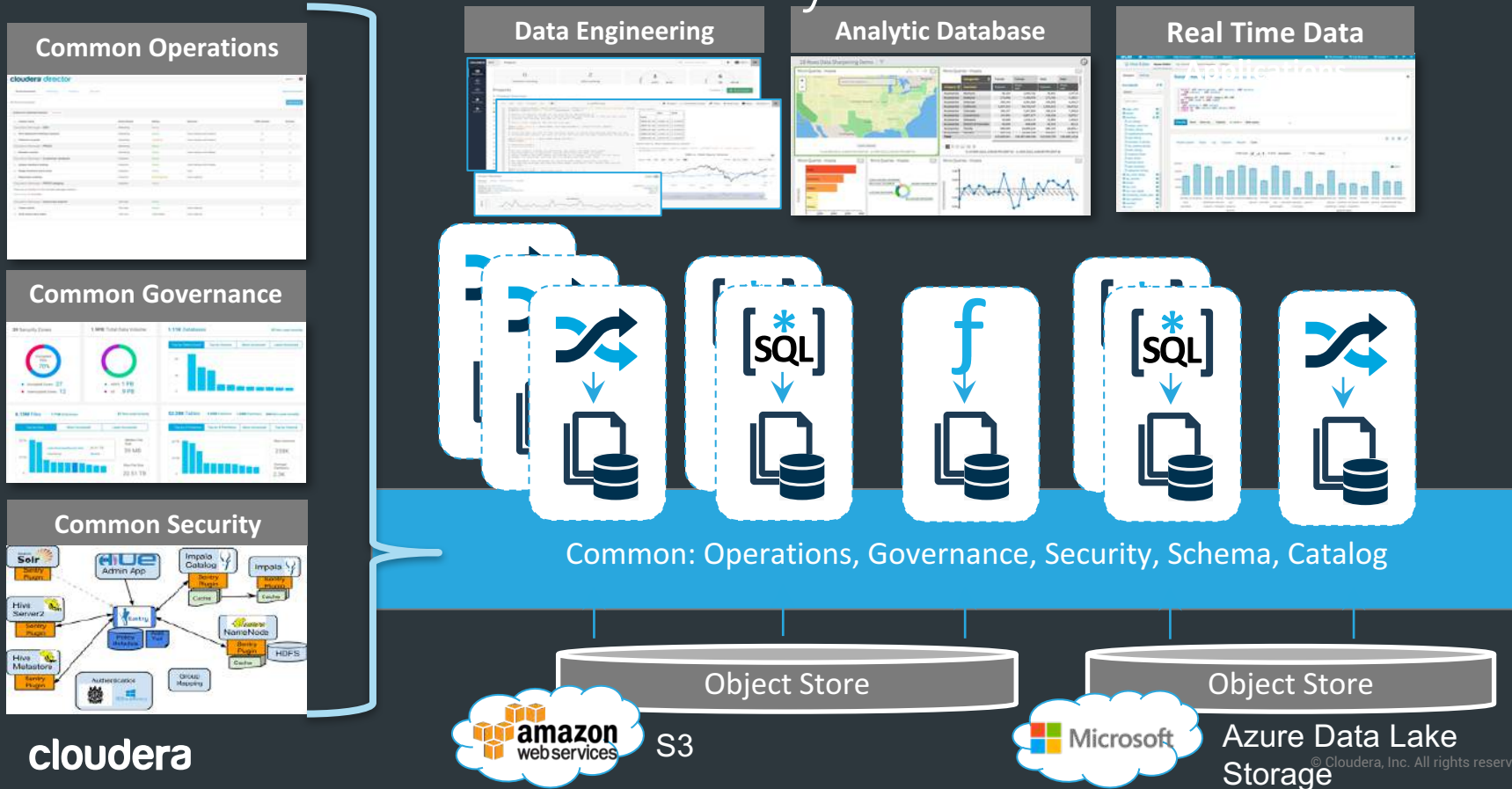
Thank you

- default
- Tables
- categories
- customers
- departments
- intermediate_access_logs
- order_items
- orders
- products
- tokenized_access_logs

```
1 --Products and URLs in one query
2 select product_name, url from (
3 select row_number() over(order by r.revenue desc) as r,
4 p.product_name, r.revenue
5 from products p inner join
6 (select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float)) as revenue
7 from order_items oi inner join orders o
8 on oi.order_item_order_id = o.order_id
9 where o.order_status <> 'CANCELED'
10 and o.order_status <> 'SUSPECTED_FRAUD'
11 group by order_item_product_id) r
12 on p.product_id = r.order_item_product_id
13 order by r.revenue desc)
14 as prod
15 inner join (
16 select row_number() over(order by count(*) desc) as r url as url count(*) as count from tokenized_access_logs
```

	product_name	url
1	Field & Stream Sportsman 16 Gun Fire Safe	/department/apparel/category/cleats/product/Perfect Fitness Perfect Rip Deck
2	Perfect Fitness Perfect Rip Deck	/department/apparel/category/featured shops/product/adidas Kids' RG III Mid Football Cleat
3	Diamondback Women's Serene Classic Comfort Bi	/department/golf/category/women's apparel/product/Nike Men's Dri-FIT Victory Golf Polo
4	Nike Men's Free 5.0+ Running Shoe	/department/apparel/category/men's footwear/product/Nike Men's CJ Elite 2 TD Football Cleat
5	Nike Men's Dri-FIT Victory Golf Polo	/department/fan shop/category/water sports/product/Pelican Sunstream 100 Kayak
6	Pelican Sunstream 100 Kayak	/department/fan shop/category/indoor/outdoor games/product/O'Brien Men's Neoprene Life Vest
7	O'Brien Men's Neoprene Life Vest	/department/fan shop/category/camping & hiking/product/Diamondback Women's Serene Classic Comfc
8	Nike Men's CJ Elite 2 TD Football Cleat	/department/fan shop/category/fishing/product/Field & Stream Sportsman 16 Gun Fire Safe
9	Under Armour Girls' Toddler Spine Surge Runni	/department/footwear/category/cardio equipment/product/Nike Men's Free 5.0+ Running Shoe
10	adidas Youth Germany Black/Red Away Match Soc	/department/footwear/category/fitness accessories/product/Under Armour Hustle Storm Medium Duffle E

Cloudera: Modern Platform for Data Management and Analytics



cloudera

© Cloudera, Inc. All rights reserved.