WHO WE ARE

Jason Wang  
Software Engineer

Brandon Freeman  
System Engineer

Stefan Salandy  
System Engineer

Suraj Acharya  
Software Engineer

Sudhanshu Arora  
Software Engineer
TODAY’S AGENDA

• We will introduce concepts related to cloud deployment
• Talk about architecture and some major considerations.
• Do hands-on exercises related running data engineering and Data Warehouse pipelines

WHY ARE YOU MOVING TO THE CLOUD

CONSIDERATIONS FOR MOVING TO THE CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS

HANDS-ON LAB
TODAY’S AGENDA

• We will introduce concepts related to cloud deployment
• Talk about architecture and some major considerations.
• Do hands-on exercises related running data engineering and Data Warehouse pipelines

WHY ARE YOU MOVING TO THE CLOUD

CONSIDERATIONS FOR MOVING TO THE CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS

HANDS-ON LAB
Shift To The Cloud

50% of the market by 2019

On-prem
- Application
- Data
- Runtime
- Middleware
- OS
- Virtualization
- Servers
- Storage
- Networking

Infrastructure as a Service
- Application
- Data
- Runtime
- Middleware
- OS
- Virtualization
- Servers
- Storage
- Networking

Platform as a Service
- Application
- Data
- Runtime
- Middleware
- OS
- Virtualization
- Servers
- Storage
- Networking

Software as a Service
- Application
- Data
- Runtime
- Middleware
- OS
- Virtualization
- Servers
- Storage
- Networking

Managed for you
Roll Your Own
Infrastructure Deployment Types

On-prem
- Application
- Data
- Runtime
- Middleware
- OS
- Virtualization
- Servers
- Storage
- Networking

Infrastructure as a Service
- Application
- Data
- Runtime
- Middleware
- OS
- Virtualization
- Servers
- Storage
- Networking

Platform as a Service
- Application
- Data
- Runtime
- Middleware
- OS
- Virtualization
- Servers
- Storage
- Networking

Software as a Service
- Application
- Data
- Runtime
- Middleware
- OS
- Virtualization
- Servers
- Storage
- Networking
Migrating To The Cloud

Simplified (CIO) view:
Cloud is the same but we save money!
Migrating To The Cloud

On-Prem

Cloud

**Don't lift and shift**
**Cost can add up quickly**
When Is It Worth It?

Compute/storage needs are changing over time and/or hard to predict.
When Is It Worth It?

2 IT lacks scale to make cluster maintenance cost effective
When Is It Worth It?

IT lacks scale to integrate many new components
When Is It Worth It?

IT lacks scale to manage component lifecycles (upgrades, patches, etc)
TODAY’S AGENDA

• We will introduce concepts related to cloud deployment
• Talk about architecture and some major considerations.
• Do hands-on exercises related running data engineering and Data Warehouse pipelines

WHY ARE YOU MOVING TO THE CLOUD

CONSIDERATIONS FOR MOVING TO THE CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS

HANDS-ON LAB
Considerations for the cloud
There could be many dimensions along which to analyze use case

- Tenancy (single / multi user)
- Lifecycle (transient / long-lived / HA)
- Security requirements and network access
- Multiple services sharing resources / resource contention
- Level of configuration tuning
- Etc.

Based on that On-prem, IaaS, PaaS, or SaaS may make sense
Tenancy & Lifecycle

**Transient**
- Hive ETL
  - Ran by headless server
  - Does not need authorization
  - Does not need to exist after processing
  - Only needs to preserve metadata
  - Stores data in object storage

**Long-lived**
- Impala DW
  - Multiple people and user personas
  - Cluster needs security
  - Cluster will need to be available, but can tolerate scheduled downtime
  - Stores data in object storage

**Highly Available**
- Kafka streams
  - All processed with same user
  - Has stringent performance requirements
  - Cannot take downtime
  - Requires extensive monitoring
  - Stores objects within the cluster or in object storage
Good PaaS candidates

- Clusters that need to be provisioned on-demand or another simple lifecycle
- Clusters where configuration can be derived from basic cluster characteristics (size, service type), and do not need to be highly optimized
- Cases where one can effectively leverage other advanced features provided by managed services (autoscaling, debugging)

- Ran by headless server
- Does not need authorization
- Does not need to exist after processing
- Only needs to preserve metadata
- Stores data in object storage
Good IaaS candidates

- Cluster has non-trivial lifecycle, such as an HA cluster
- Cases where monitoring needs to be integrated with sophisticated monitoring and metering system
- Configurations need to be highly optimized for given use case – several services need to run on one cluster with non-trivial resource contention
- Very restrictive security environment into which cluster needs to deploy (can’t talk to the managed service)

- All processed with same user
- Has stringent performance requirements
- Cannot take downtime
- Requires extensive monitoring
- Stores objects within the cluster or in object storage

Kafka streams
Switching to the cloud

● It is worth considering moving in stages – one use case at a time
  ○ Lift and shift
  ○ Data into S3 for a more cloud native deployment
  ○ Metadata shareable between clusters
  ○ Full cloud native, on demand deployment

● It may be difficult to switch from on on-prem deployment model to a cloud deployment model
  ○ IT needs to map access and infrastructure to a new deployment model
  ○ Developers need to adjust all pipelines to cloud native architecture
Public cloud lock-in

- Choice of public cloud can be even more consequential than choice of on-prem vendors
- Some companies maintain multi-cloud infrastructure – several public clouds or on-prem and public
  - Pros: less lock-in
  - Cons: multi-front effort
- Can maintain migration plan “from the cloud”
  - Make sure there is understood data migration plan
  - Make sure there is understood metadata migration plan
  - Rely on non-proprietary technologies for core processing
TODAY'S AGENDA

- We will introduce concepts related to cloud deployment
- Talk about architecture and some major considerations.
- Do hands-on exercises related running data engineering and Data Warehouse pipelines
High Level View
On-Prem Deployment

Technical View

Hive Cluster
- Hue
- Oozie
- CM
- Yarn
- Hive
- Sentry
- Spark
- HDFS

Private subnet

Storage

Physical infra

Physical infra

Active Directory

BDR

SSSD

users
Properties and considerations of cloud deployment

- Decoupled storage and compute
- Multi-cluster metadata
- Bursting to the cloud
- Data security
- User management

For clarity, we may specifically refer to AWS component names; unless pointed out otherwise the same applies to Azure, GCP.
Decoupled compute and storage

Private subnet

- User computers
- Hive cluster
- Active Directory

Cloud infrastructure

- Storage
- Compute
- IAM

Virtual private subnet

Hive cluster

- Hue
- CM
- Oozie
- Yarn
- Hive
- Sentry
- Spark

HDFS

© Cloudera, Inc. All rights reserved.
Decoupled compute and storage – local and attached volumes

- Except for specific instance types cloud VM’s have minimal local disk
- Local disk may, by default, contain sensitive information – e.g. logs – so you should not forget that it exists
- Most of the time virtual disk needs to be attached for each instance, for things like Spark memory spills and logs. Even when no data is meant to persist there.
Decoupled compute and storage – same HDFS new role

- Local HDFS can make sense for latency sensitive data computations, but it should not be used to persist data.
- Probability of VM failure is almost an order of magnitude higher than that of well maintained physical server.
- If spot instances are used, you must build framework to be able to tolerate catastrophic failures.
- HDFS, and its level of redundancy, is not meant to deal with volatility of public cloud.
- Not all instances need to be parts of HDFS.
Decoupled compute and storage – S3 vs volumes

- S3 is elastic
- S3 is much more reliable
- Attached volume provide lower latency access, even with HDFS
- S3 has higher latency, but can come close in throughput
Decoupled compute and storage – S3 consistency

- AWS S3 is not strongly consistent, unlike Azure ADLS, HDFS
- Objects written to S3, may not be instantly seen by everyone
- While object is present, meta-data is lagging – cannot list
Decoupled compute and storage – S3 consistency

S3guard introduces strong consistency to S3

- DynamoDB serves as the local metadata source
- If objects from S3 get deleted without using S3guard, it needs to be invalidated
- S3guard is a version of EMRFS from AWS’s for EMR
Multi-cluster metadata

Hive cluster
- CM
- Hive
- Spark
- HDFS

Impala cluster
- CM
- Impala
- HDFS

Cloud infrastructure
- storage
Multi-cluster metadata

- One of the main advantages of running in the cloud is the ability to separate workloads
  - Each cluster is optimized for single workflow
  - Cluster lifecycle matches workflow’s demands
- Persistent storage (e.g. S3, ADLS) is a necessary part, but not sufficient; we need answers for metadata:
  - Schema
  - Authorization
  - Lineage / data governance
- All of them need to be persistent, with lifecycle decoupled from that of the cluster
Multi-cluster metadata

- If we are sharing resources we need to address:
  - Performance impact
  - Concurrency
  - Other design limitations
Multi-cluster metadata – Latency

- Impact of Schema and Authorization latency
  - Most of the Apache stack is being optimized for cloud use cases, but was not designed this way originally
  - Depending on the location and load on these components latency may increase significantly
  - Certain queries are extremely metadata intensive and should be done with care
    - e.g. Invalidate entire metadata for Impala
    - e.g. Declare DDL over large number of objects
Multi-cluster metadata – Concurrency

- Concurrent write access by multiple components can leave metadata in inconsistent state
  - While locking exists in HMS, it is not meant to be multi-cluster
  - As a result, it is responsibility of the user to minimize the risk of concurrent writes
Multi-cluster metadata – other design limitations

- Managed: relationship between table and data lifecycle
  - SQL syntax explicitly differentiates between managed and unmanaged tables
- Local/Global: visibility of data across clusters
  - Notion of global tables does not explicitly exist, which can cause issues.
- Example:
  - Cluster is created to run a daily workflow, which creates table in short-lived HDFS
  - Metadata is shared via global HMS
  - Table and its content disappears when cluster is shutdown, but HMS maintains a record of it as a managed table
  - Subsequent execution will fail since declaration of the table will collide with declaration of non-existing table
Multi-cluster metadata – other design limitations

- HMS was designed to be local to the cluster
  - No strict boundary existed for file access
  - Some variables were stored in local process
- If central HMS exists in different subnet with limited data access (HDFS, S3) will fail
Bursting to the cloud

- User computers
- Hive cluster
- Active Directory

Cloud infrastructure:
- Storage
- Compute
- IAM

Virtual private subnet:
- Private subnet

Hive cluster:
- Hue
- CM
- Oozie
- Yarn
- Hive
- Sentry
- Spark
- HDFS
Bursting to the cloud - Why?

Concept:

- Local cluster is at max capacity
- We need to perform critical computational tasks
- We will create a cluster on the cloud and run the tasks!
Bursting to the cloud – Raw Data Movement

- Data could be periodically sync’d using BDR (e.g. with Cloudera Manager)
- Data can take very long time to copy
- Is BDR data fresh enough for bursting purposes?
- Is it feasible given data size?
Bursting to the cloud – Metadata Movement

- Metadata should also be replicated during BDR
- Metadata recreation can take hours, which is an obstacle not only for bursting, but for any on-demand use
Security
Data Security - Overview
Data Security – Object Storage Encryption Options

- Three server side types of S3 encryption
  - S3-SSE (server side). Built into ADLS.
  - S3-KMS (server side, specific key)
  - S3-C (customer provided)
- Type of encryption needs to be specified per object read/write
- To read data in a bucket encrypted with KMS key
  - Right to read from bucket
  - Right to use KMS key
Data Security – Object Storage Encryption

- Default bucket encryption – policy specified on bucket
  - Was always there in Azure
  - Available in AWS since late 2017
- Will do the work for you
  - Check that instance has bucket access
  - Check which KMS key to use
  - Check that instance has KMS key access
Data access and security – object storage encryption

- One can set per user Credentials
  - As Boto Environment variables on EC2
    - Credentials resolved:
      1. Session credentials
      2. Local Environment
      3. Global Environment
      4. Instance profile
  - Not recommended – hard to manage
    - Accessible to user
    - Need to rotate
- Prefered way is to use Role based access control
Data Security – Wire Encryption

- Easy outside of cluster / Hard inside of cluster
  - Communication to S3
    - Secured using HTTPS
  - Communication between UI components
    - TLS
  - Communication between Impala workers
    - Kerberos for establishing connection
    - TLS for data
  - HDFS RPC uses its own framework
    - Supports Kerberos for Authentication
    - TLS for data
- Other services communication
  - E.g. HMS, YARN client-server relationships
  - Does not use TLS (AES based), uses Kerberos (3DES based)
  - No hardware acceleration – significantly slower data movement
Identity Management

Cloud infrastructure

- Compute
- Storage
- IAM

Private subnet

- User computers
- Hive cluster
- Active Directory

Virtual private subnet

- Hive cluster
- Hue
- CM
- Oozie
- Yarn
- Hive
- Spark
- Sentry
- HDFS

IAM
Identity management

- On-demand cluster deployment needs to connect cloud identity to corporate identity
- There are three main types of identities that need to be reconciled
  - Corporate User Directory
  - Cluster OS users
  - IAM roles / users
Identity management – LDAP on-prem and cloud

- Need to have both LDAP and KDC (Kerberos Key Distribution Center)
  - Can be maintained separately
- Two common choices for user management
  - Active Directory
  - FreeIPA
- Both contain LDAP and KDC
  - Will synchronize users
Identity management – LDAP on-prem and cloud

- In case of AD content can be synchronized using Active Directory Federation Service (ADFS)
- Requires two directional network access
  - May be a problem for some corporate infosec
  - Most of the time non-starter with third party providers
Identity management – LDAP on-prem and cloud

- For third party integration the preferred way is to use SAML
  - No secrets that can be leaked – just a signed assertion
  - Does not require inbound connection to on-prem
  - Could be exposed via API endpoint, not whitelisting AD/FreeIPA
Identity management – OS users

- OS users need to be synchronized with user directory
- Standard tool is SSSD
- Needs to be automated to create on-demand provisioning
  - Or done by managed services
Identity management – HDFS users

- HDFS users can come from several places
  - Directly form LDAP
    - Not efficient; lacks some features
  - Can maintain isolated set of users
    - Does not help
  - Can refer to OS users
    - Unix system lookup of user-group membership
Identity management – CM, Hue, Impala

- Other components support various authentication means
  - LDAP (CM, Hue, Impala)
  - KDC (Hue, Impala)
  - SAML (CM)
  - Internal user directory (Hue)
  - Customer service (CM)
- LDAP is the common denominator for most

Diagram:
- Cloud user Directory
- LDAP
- KDC
- SAML
- Corporate User Directory (on-prem)
- SSSD
- Instance
- OS Users
- Impala cluster
- Hue
- Impala
- HDFS
- CM
Identity management – IAM

- Depending on the public cloud can be easy or hard
  - Azure uses AD so can just federate
  - AWS need to use third party tools
- Unfortunately, this synchronization is still difficult to tie to data (object store) access, as we discussed
- IAM required for S3 bucket access
Questions?
TODAY’S AGENDA

WHY ARE YOU MOVING TO THE CLOUD

CONSIDERATIONS FOR MOVING TO THE CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS

HANDS-ON LAB

• We will introduce concepts related to cloud deployment
• Talk about architecture and some major considerations.
• Do hands-on exercises related running data engineering and Data Warehouse pipelines
ALTUS DIRECTOR

MULTI CLOUD

IaaS

CLOUDERA ALTUS

MULTI CLOUD

PaaS

Data Warehouse

DATA ENGINEERING

DATA CATALOG

SECURITY

GOVERNANCE

WORKLOAD MANAGEMENT

INGEST & REPLICATION

CDH

AWS S3

Microsoft ADLS
Migrate data to the cloud. Migrate services to run on the cloud. Evaluate performance without investment in re-engineering

Given data and metadata are cloud based, opportunistically migrate specific use cases to cloud native frameworks
Altus Data Engineering

What is it?
• Short-lived
• Single tenant
• Hive, Spark, or MapReduce Cluster

What is it used for?
• ETL jobs and data pipelining
• Batch processing
• With data living in S3 or ADLS
• Provides fast and easy job submission without cluster management
Altus Data Warehouse

What is it?
- Long-lived
- Multi tenant
- Impala Cluster

What is it used for?
- Data warehousing
- Analytics
- With data living in S3 or ADLS
- Provides fast and easy analytics without cluster management
Altus Shared Data Experience (SDX)

What is it?

- Cloud native shared metadata store
- With metadata living Cloudera managed cloud storage

What is it used for?

- **Shared cataloging** to define and preserve structure and business context of data
- Provides **unified security** across transient and recurring workloads
- Enables **consistent governance** across all data to increase compliance
TODAY’S AGENDA

- We will introduce concepts related to cloud deployment
- Talk about architecture and some major considerations.
- Do hands-on exercises related running data engineering and Data Warehouse pipelines

WHY ARE YOU MOVING TO THE CLOUD

CONSIDERATIONS FOR MOVING TO THE CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS

HANDS-ON LAB
Setting the scene
Solving a business need

- We work for an outdoor clothing retail company and website sales are struggling

- We need to figure out whether sales orders correlate with website visits and what steps we can take to improve sales

- We’ll architect and price out clusters, set up a data pipeline, run analytics on the data, and then troubleshoot common problems
Already set up: raw data ingestion

Sales Orders
- customers
- departments
- order_items
- orders
- products

Raw Logs
Part one: Data Exploration and Analytics on Sales Orders

Sales Orders
- customers
- departments
- order_items
- orders
- products

Data Warehouse

ALTUS SDX
Part two: Data Pipeline - Prepare and clean the web logs

Sales Orders
- customers
- departments
- order_items
- orders
- products

Raw Logs

Tokenized logs

Sales Orders
- customers
- departments
- order_items
- orders
- products

Raw Logs

Tokenized logs
Part three: Analytics on orders and site visits

### Sales Orders
- **customers**
- **departments**
- **order_items**
- **orders**
- **products**

### Raw Logs
- **IP Address**
- **Timestamp**
- **Department**
- **Product**

<table>
<thead>
<tr>
<th>IP Address</th>
<th>Timestamp</th>
<th>Department</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>39.244.91.133</td>
<td>14/Jan/2014:10:30:13 -0400</td>
<td>Kids</td>
<td>Shorts</td>
</tr>
<tr>
<td>150.47.54.136</td>
<td>14/Jan/2014:10:30:13 -0400</td>
<td>Apparel</td>
<td>Jacket</td>
</tr>
<tr>
<td>217.89.36.129</td>
<td>14/Jan/2014:10:30:13 -0400</td>
<td>Outdoors</td>
<td>Umbrella</td>
</tr>
<tr>
<td>36.44.59.115</td>
<td>14/Jan/2014:10:30:13 -0400</td>
<td>Bedding</td>
<td>Pillow cases</td>
</tr>
</tbody>
</table>
- **Tokenized logs**

### ALTUS SDX
- **DATA ENGINEERING**
- **Data Warehouse**
https://tinyurl.com/stratanyyc18

Starting back at 3:20pm EDT
Handout #1: Log into Altus
Handout #2: Create the Clusters
15 minute break
Handout #3: Architecture and cost

Compute (EC2) and storage (S3) will usually be the biggest costs (and, depending on workload, data transfer as well)

There are a lot of things to consider when designing a cluster architecture, three big ones for people used to on-prem deployments are:

- Transient vs. Persistent
- How to manage data (S3 vs HDFS)
- Automate, automate, automate

See the handout for more things to consider
Data Engineering on AWS - Best Practices

**Default Choice**

**Transient Batch (most flexible)**
- Spin up clusters as needed.
- On-demand/spot instances
- Usage-based pricing
- Sized for workload
- Cluster per tenant/user

**Persistent Batch (most control)**
- Persistent cluster(s) for frequent ETL.
- Reserved instances
- Node-based pricing
- Grow/shrink
- Cluster per tenant group

**Persistent Batch on HDFS (fastest)**
- Top performance for frequent ETL.
- Reserved instances
- Node-based pricing
- Grow/shrink
- Shared across tenant groups

Diagram:
- Batch Cluster
- Persistent Cluster
- HDFS
- Object Storage
**DE Pattern #1: Transient Cluster on Object Storage**

Ideal for async, infrequent, and irregular jobs. Often set up as an automatic pipeline where cluster startup times don’t matter.

**Pros**
- Tailor clusters to jobs
- Run workloads in isolation
- Use Spot instances to lower cost
- Create / terminate clusters as needed

**Cons**
- Incur cluster startup cost (time and money, time will probably matter less)
- Spot instances can be terminated (need to automate handling this)
DE Pattern #2: Persistent Cluster on Object Storage

Ideal for homogeneous jobs that run well on the same cluster and jobs that run frequently or continuously.

Pros
- No cluster startup cost (time and money)
- Grow and shrink the cluster on demand

Cons
- Potentially paying for unused compute (can shrink the cluster to lower the base cost)
- Less workload isolation
DE Pattern #3: Persistent Cluster on Local Storage

Ideal for “lift and shift” jobs, jobs that require maximum performance. Or when you’re first getting started in the cloud.

Pros
- Good first step in cloud migration
- Most performant
- No cluster startup cost (time and money)

Cons
- Potentially paying for unused compute
- Clusters are less elastic (unlike S3 it’s difficult to resize HDFS)
- Less workload isolation / cluster flexibility
Data Warehouse on AWS - Best Practices
DW Pattern #1: Transient Cluster on Object Storage

Ideal for async, unpredictable, and irregular workloads. Or exploratory workloads and workloads still under development.

Pros
- Quick iteration
- Tailor instances and jobs to workloads
- Run workloads in isolation
- Use Spot instances to lower cost

Cons
- Incur cluster startup cost (time and money)
- Spot instances are not always ideal (subject to more interruptions)
DW Pattern #2: Persistent Cluster on Object Storage

Ideal for frequent, flexible, and changeable workloads.

Pros
- Predictable results
- Full multi-tenant isolation
- No cluster startup cost (time and money)

Cons
- Per node performance is lower with S3 than with HDFS. You can compensate for this by using more, cheaper nodes, for more throughput.
DW Pattern #3: Persistent Cluster on Local Storage

Ideal for “lift and shift” jobs, jobs that require maximum performance. Or when you’re first getting started in the cloud.

Pros
- Good first step in cloud migration
- Most performant for tight SLAs
- No cluster startup cost (time and money)

Cons
- Clusters are less elastic (unlike S3 it’s difficult to resize HDFS)
- Less workload isolation
- Less instance type and cluster flexibility
Handout #3: Architecture and cost

Review cluster and S3 considerations in the handout
Handout #3a: Architecture and cost for a Data Warehouse cluster
Handout #3b: Architecture and cost for a DE cluster

Assignment

1. Calculate per hour cost for one node (m4.2xlarge)

2. Calculate per month cost for the whole cluster (at 4 hours / day)

3. Calculate per month S3 cost (assuming all 90 days of data is already in S3)

4. Calculate total AWS cost
Handout #3b: Architecture and cost for a DE cluster

Answers

1. Calculate per hour cost for one node (m4.2xlarge)
   $0.4 [ec2] + $0.071 [ebs] = $0.471

2. Calculate per month cost for the whole cluster (at 4 hours / day)
   $0.471 [1 node / hour] * 5 [nodes / cluster] * (4 * 30) = $282.6

3. Calculate per month S3 cost (assuming all 90 days of data is already in S3)
   1TB * 90 = 90,000GB * $0.023 = $2,070

4. Calculate total AWS cost
   $282.6 + $2,070 = $2,352.6
Handout #4: Run Queries against Data Warehouse - Part 1

- Review the Data in S3 using AWS cloud-altus-demo account
- Connect to your Data Warehouse cluster and create tables
- Explore the data: run your first queries

- Based on the total number of items sold, what are the top 5 best selling products?
Handout #5: Run Queries against DE

• Review the web logs in S3 using AWS cloud-altus-demo account
• Execute a batch job to parse the logs and produce a tokenized table
• Look at S3 and see what has been created

• Optional exercise: run a second batch job to produce a denormalized products table
Handout #6: Run Queries against Data Warehouse - Part 2

- Use the SQL Editor in the Data Warehouse Cluster
- See what new tables are available for Data Analysts
- Explore the tokenized logs table

- Run a complex query: compare the sales information with the web site visits and see if there is anything wrong. Use the provided SQL query.
Handout #7: Troubleshoot failures

- Create a Hive DE job expected to fail
- Use Workload Analytics to understand the error

- (Optional) People with their own account connect using dataeng@altus-demo.com
- Look for the failed SparkPi job
- Use Workload Analytics to understand the error
WRAP-UP
Key benefits of PaaS

Spin up working environments ad hoc

Bring your own data and tools

Adjust resources on-demand

Pay for your actual consumption of resources
THANK YOU