RUNNING DATA ANALYTIC WORKLOADS IN THE CLOUD
WHO WE ARE

Jason Wang
Software Engineer

Mael Ropas
Sr. Sales Engineer

Vinithra Varadharajan
Sr. Engineering Manager

Eugene Fratkin
Director of Engineering
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 analytic database pipelines

WHY ARE YOU MOVING TO THE CLOUD

HOW TO MAP ON-PREM TO CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS + DIRECTOR

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 analytic database pipelines

WHY ARE YOU MOVING TO THE CLOUD

HOW TO MAP ON-PREM TO CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS + DIRECTOR

HANDS-ON LAB
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

Managed for you
Shift to the cloud

50% of the market by 2019

Managed for you

On-premise

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
Let's migrate!

Simplified (CIO) view:
Cloud is the same but saves money!
Lets migrate!

Simplified (CIO) view:

Hive cluster

Physical infra

Cloud infra

Hive cluster
Lets migrate!

Large on-prem cluster operated at high capacity will be cheaper than Cloud based solution...

Except...
When is it worth it?

Compute/storage needs are changing over time and/or hard to predict
When is it worth it?

IT lacks scale to make cluster maintenance cost effective
When is it worth it?

3. 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 analytic database pipelines

WHY ARE YOU MOVING TO THE CLOUD

HOW TO MAP ON-PREM TO CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS + DIRECTOR

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)
- Secure 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
Considerations for the cloud

- **Transient**
  - **Single Tenant**
    - 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
  - **Multi Tenant**

- **Long-lived**
  - **Impala ADB**
    - 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
Considerations for the cloud – 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
- Case where one can effectively leverage other advanced features provided by managed services (autoscaling, debugging)

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
Considerations for the cloud – good IaaS candidates

- Cluster has non-trivial lifecycle, such as HA cluster
- Case 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)
Considerations for the cloud – practical considerations when switching

● It may be difficult to switch from 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
● 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
Considerations for the cloud – 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
  - Less lock-in
  - 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 analytic database pipelines

WHY ARE YOU MOVING TO THE CLOUD

HOW TO MAP ON-PREM TO CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS + DIRECTOR

HANDS-ON LAB
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

- Cloud infrastructure
  - private subnet
    - User computers
    - Hive cluster
    - Active Directory
  - virtual private subnet
    - Hive cluster
    - Hue
    - CM
    - Oozie
    - Yarn
    - Hive
    - Sentry
    - Spark
  - HDFS

- Cloud infrastructure
  - storage
  - compute
  - IAM
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.
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 a 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
- EMRFS is AWS's version of S3guard only for EMR
Multi-cluster metadata

Hive cluster
- CM
- Spark
- HDFS

Impala cluster
- CM
- Impala
- HDFS

Cloud infrastructure
- storage

virtual private subnet
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

Cloud infrastructure

- Cloud infrastructure
  - IAM
  - Compute
  - Storage

Virtual private subnet

- User computers
- Active Directory
- Hive cluster
- HDFS
- Yarn
- Spark
- Oozie
- Sentry
- Hue

Private subnet
Bursting to the cloud

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 – 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 – data 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
Data security – object storage encryption

- 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 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

Hive cluster

virtual private subnet

Hive cluster

Hue

CM

Oozie

Yarn

Hive

Sentry

Spark

HDFS

IAM

compute

storage
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
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

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

WHY ARE YOU MOVING TO THE CLOUD

HOW TO MAP ON-PREM TO CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS + DIRECTOR

HANDS-ON LAB
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 Analytic DB

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 analytic database pipelines

WHY ARE YOU MOVING TO THE CLOUD

HOW TO MAP ON-PREM TO CLOUD

CLOUD ARCHITECTURE AND CONSIDERATIONS

CLOUDERA ALTUS + DIRECTOR

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
Part one: Data Exploration and Analytics on Sales Orders

Sales Orders
- customers
- departments
- order_items
- orders
- products
Part two: Data Pipeline - Prepare and clean the web logs

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

### Raw Logs

<table>
<thead>
<tr>
<th>IP Address</th>
<th>Timestamp</th>
<th>Department</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>79.133.215.123</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Water sports</td>
<td>Sunscreen</td>
</tr>
<tr>
<td>39.244.91.133</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Kids</td>
<td>Shorts</td>
</tr>
<tr>
<td>150.47.54.136</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Apparel</td>
<td>Jacket</td>
</tr>
<tr>
<td>217.89.36.129</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Outdoors</td>
<td>Umbrella</td>
</tr>
<tr>
<td>36.44.59.115</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Bedding</td>
<td>Pillow cases</td>
</tr>
<tr>
<td>39.244.91.133</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Kids</td>
<td>Shorts</td>
</tr>
<tr>
<td>150.47.54.136</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Apparel</td>
<td>Jacket</td>
</tr>
<tr>
<td>217.89.36.129</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Outdoors</td>
<td>Umbrella</td>
</tr>
<tr>
<td>36.44.59.115</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Bedding</td>
<td>Pillow cases</td>
</tr>
</tbody>
</table>

### Tokenized Logs

<table>
<thead>
<tr>
<th>IP Address</th>
<th>Timestamp</th>
<th>Department</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>79.133.215.123</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Water sports</td>
<td>Sunscreen</td>
</tr>
<tr>
<td>39.244.91.133</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Kids</td>
<td>Shorts</td>
</tr>
<tr>
<td>150.47.54.136</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Apparel</td>
<td>Jacket</td>
</tr>
<tr>
<td>217.89.36.129</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Outdoors</td>
<td>Umbrella</td>
</tr>
<tr>
<td>36.44.59.115</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Bedding</td>
<td>Pillow cases</td>
</tr>
<tr>
<td>39.244.91.133</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Kids</td>
<td>Shorts</td>
</tr>
<tr>
<td>150.47.54.136</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Apparel</td>
<td>Jacket</td>
</tr>
<tr>
<td>217.89.36.129</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Outdoors</td>
<td>Umbrella</td>
</tr>
<tr>
<td>36.44.59.115</td>
<td>14Jun2014 10:30 13:0400</td>
<td>Bedding</td>
<td>Pillow cases</td>
</tr>
</tbody>
</table>
Part three: Analytics on orders and site visits

SALES ORDERS

- 2 colors available
  - Drypoint GTX Jacket - Women's Plus
    - Sizes: $295.00
  - 3 colors available
    - Riber GTX Jacket - Men's
    - Sizes: $279.00
  - 4 colors available
    - Drypoint GTX Jacket - Men's
    - Sizes: $249.00

RAW LOGS

- 2 colors available
  - Drypoint GTX Jacket - Women's Plus
  - Sizes: $295.00
- 3 colors available
  - Riber GTX Jacket - Men's
  - Sizes: $279.00
- 4 colors available
  - Drypoint GTX Jacket - Men's
  - Sizes: $249.00

TOKENIZED LOGS

- IP Address: 79.133.215.123
  - Timestamp: 14/Jan/2014:10:30:13 -0400
  - Department: Footwear
  - Product: Football cleat
- IP Address: 39.234.91.133
  - Timestamp: 14/Jan/2014:10:30:13 -0400
  - Department: Water sports
  - Product: Sunscreen

DATA ENGINEERING

ANALYTIC DATABASE

ALTUS SDX
http://tiny.cloudera.com/stratalondon18
Handout #1: 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

For full details Google “Cloudera Data Engineering on AWS Best Practices”
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)

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

Cloudera Director
Batch Cluster
Batch Cluster
Batch Cluster

Shared HMS DB
Object Storage
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 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
Analytic Database on AWS - Best Practices

For full details Google “Cloudera Analytic Database on AWS Best Practices”
ADB 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)
ADB 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.
ADB Pattern #3: Persistent Cluster on Local Storage

Ideal for “lift and shift” 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 #1: Architecture and cost

Review cluster and S3 considerations in the handout
Handout #1a: Architecture and cost for an ADB 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 #1b: Architecture and cost for a DE cluster

Answers

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

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

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

4. Calculate total AWS cost
   \[ \text{$282.6 + $2,070 = $2,352.6} \]
Handout #2: Log into Altus

If you have an AWS account use Altus Direct (Handout #3)

If you don’t have an AWS account hold tight
Handout #3: Altus Direct

Walkthrough
Handout #4: Create the Clusters
15 minute break
Handout #5: Run Queries against ADB - Part 1

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

- Based on the total number of items sold, what are the top 5 best selling products?
Handout #6: 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 #7: Run Queries against ADB - Part 2

- Connect to Hue in the Analytic 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 #8: 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