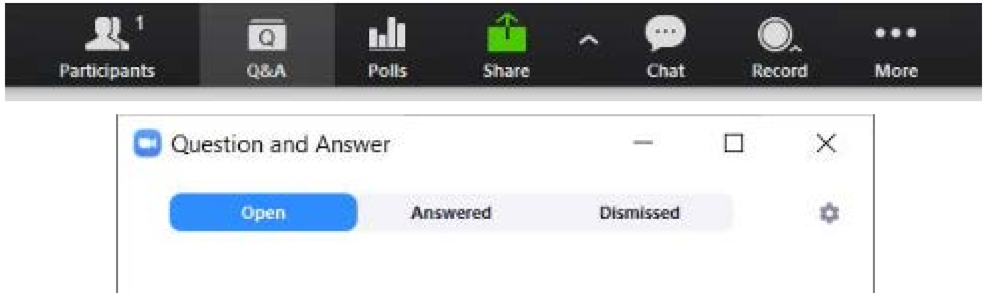
May 12th , 2020

Ephemeral clusters, an overview

Puneeth Natesha, Software Engineer, Informatica GCS(DEI)

Sampada Subnis, Software Engineer, Informatica GCS(DEI)

Housekeeping Tips



- Todays Webinar is scheduled to last 1 hour including Q&A
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to “All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available to view on our INFASupport YouTube channel and Success Portal. The link will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Success Portal

https://success.informatica.com

Learn. Adopt. Succeed.

Bootstrap product trial experience

Enriched Onboarding experience

FREE Product Learning Paths and weekly Expert sessions

Informatica Concierge with Chatbot integrations

Tailored training and content recommendations

Safe Harbor

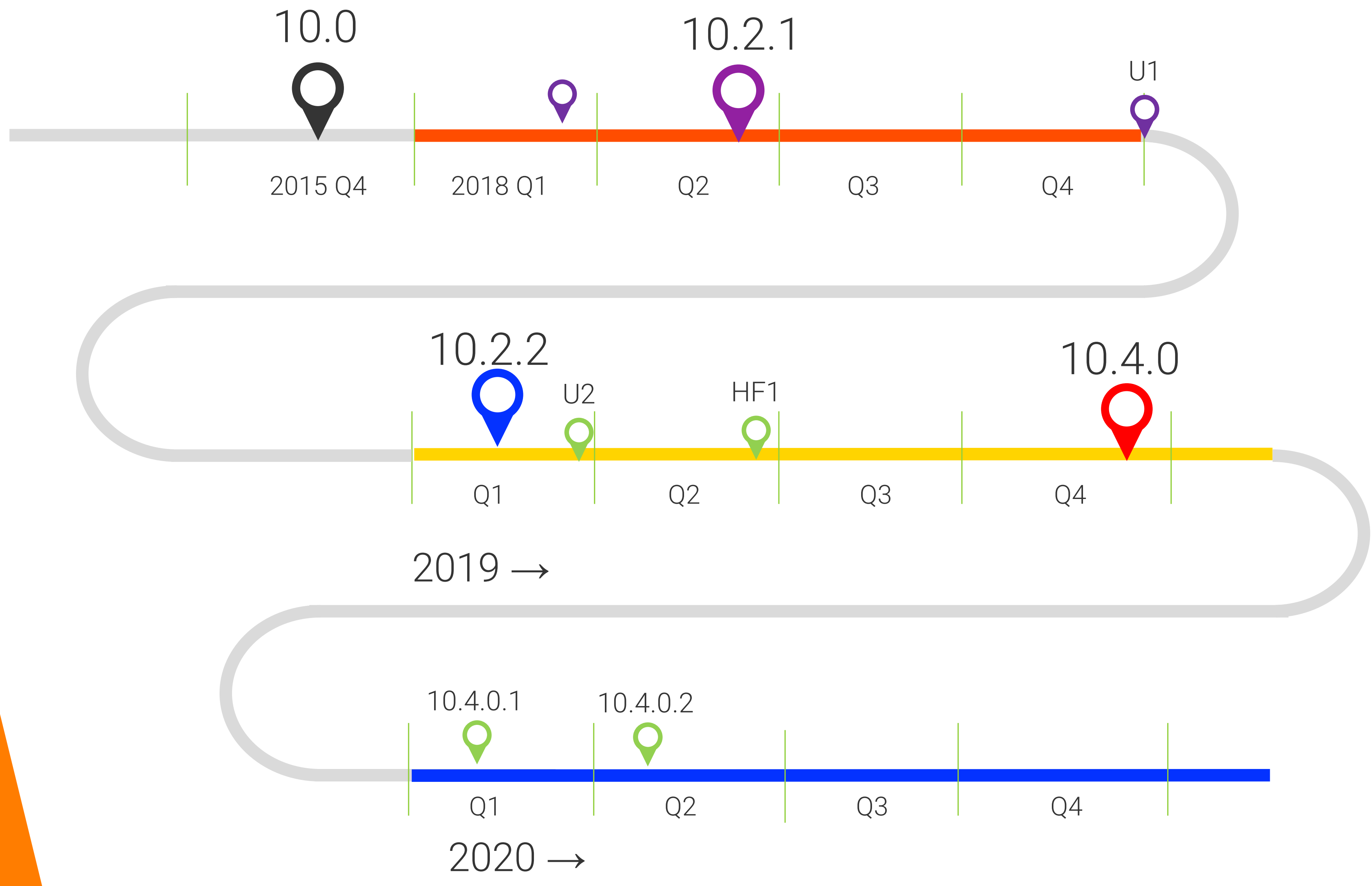
The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Agenda

- Ephemeral Cluster Support in BDM/DEI
- Pre requisites to create ephemeral cluster
- Cloud provisioning configuration
- Cluster Workflow Components
- Command line utilities
- Ephemeral Cluster Support in BDS/DES
- Demo
- Troubleshooting and self-service
- References
- Q&A

DEI Release cadence



Ephemeral Cluster Support in BDM/DEI

Informatica

Ephemeral cluster

- An ephemeral cluster is a cloud platform cluster that you create to run mappings and other tasks, and then terminate when tasks are complete.
- Create ephemeral clusters to save cloud platform resources.

An ephemeral can be used in both **DEI** & **DES** jobs

USE CASE:

- **DEI:** When the production job runs daily once or weekly run. During this time, cost of running cluster all the time over cloud.
- **DES:** Cost of running cluster all the time and hence if you want to shut down the cluster towards the end of the day and start the next morning and would like BDS mappings to start from the offset where it had left at the time of shutting down the cluster

Pre requisites to create ephemeral cluster

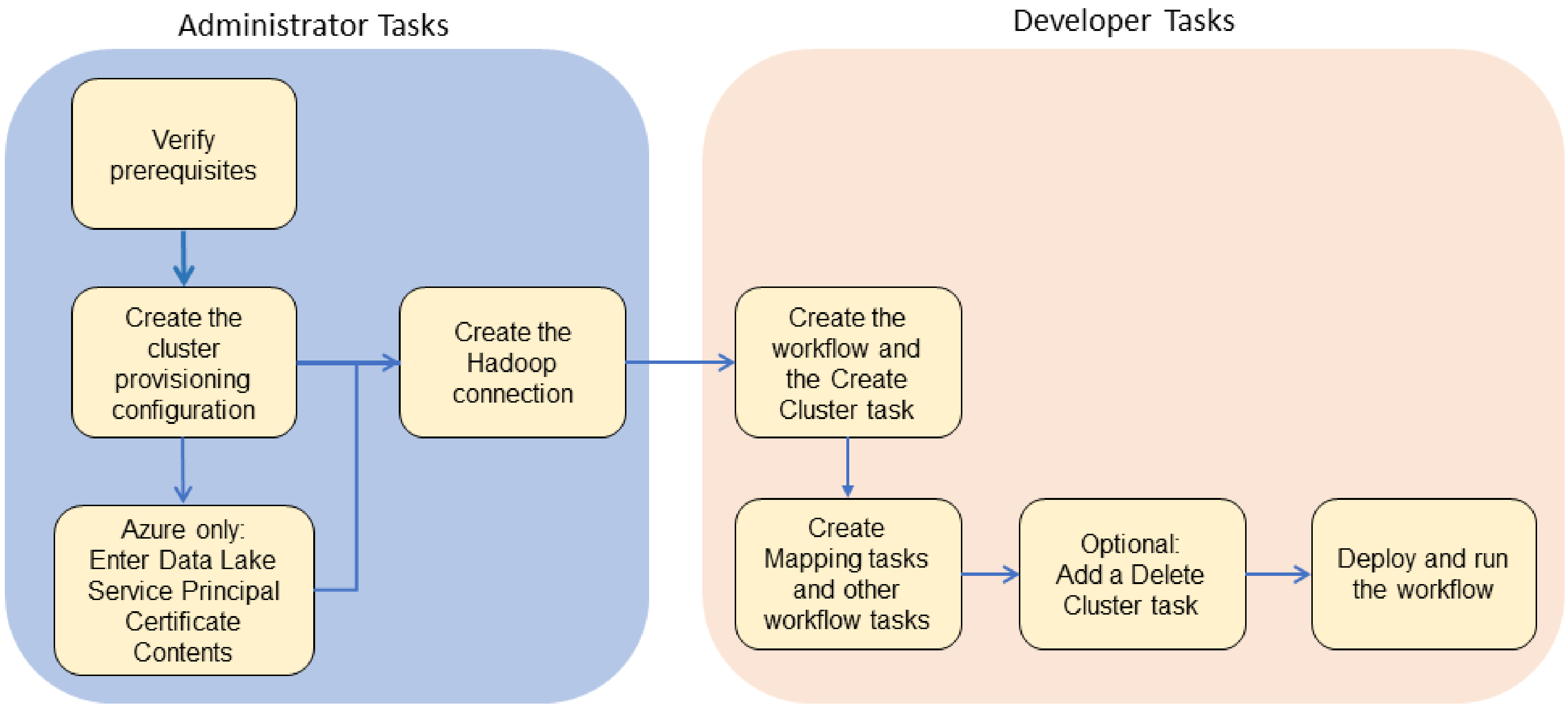
- Make sure you have purchased a license for BDM/DEI.

AWS cloud provisioning	Azure cloud provisioning	Databricks cloud provisioning
AWS Access Key ID	Subscription ID	Databricks domain
AWS Secret Access Key	Tenant ID	Databricks token ID
Region	Client ID	
EMR Role	Client Secret	
EC2 Instance Profile	Azure storage account name	
EC2 Subnet	Azure storage account key	
	Resource group, Virtual network resource group, Virtual Network, Subnet Name	

Cluster Workflow Process

Creation of a cluster workflow requires administrator and developer tasks.

The following image shows the process to create, configure, and run a cluster workflow:



Administrator Tasks

Developer Tasks

Verify prerequisites

Create the cluster provisioning configuration

Create the Hadoop connection

Azure only: Enter Data Lake Service Principal Certificate Contents

Create the workflow and the Create Cluster task

Create Mapping tasks and other workflow tasks

Optional: Add a Delete Cluster task

Deploy and run the workflow

Cloud provisioning configuration

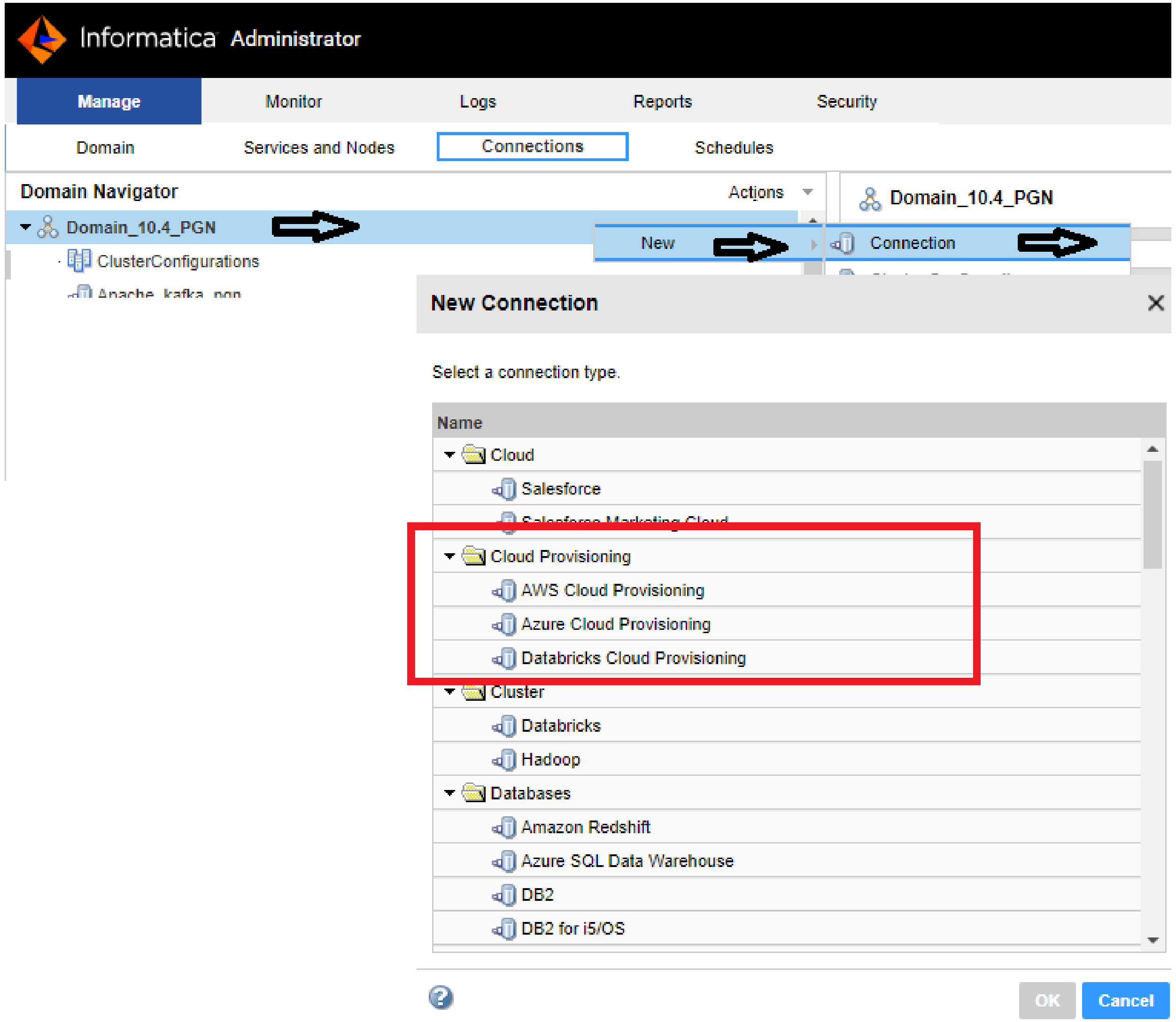
- The cloud provisioning configuration establishes a relationship between the Create Cluster task and the cluster connection that the workflows use to run mapping tasks.

- The Create Cluster task must include a reference to the cloud provisioning configuration. In turn, the cloud provisioning configuration points to the cluster connection that you create for use by the cluster workflow.

- The properties to populate depend on the Hadoop distribution you choose to build a cluster on. Choose one of the following connection types:

AWS Cloud Provisioning. Connects to an Amazon EMR cluster on Amazon Web Services.
Azure Cloud Provisioning. Connects to an HDInsight cluster on the Azure platform.
Databricks Cloud Provisioning. Connects to a Databricks cluster on the Azure Databricks platform.

Cloud provisioning configuration



New Connection

Select a connection type.

Name
- Cloud
  - Salesforce
  - Salesforce Marketing Cloud
- Cloud Provisioning
  - AWS Cloud Provisioning
  - Azure Cloud Provisioning
  - Databricks Cloud Provisioning
- Cluster
  - Databricks
  - Hadoop
- Databases
  - Amazon Redshift
  - Azure SQL Data Warehouse
  - DB2
  - DB2 for i5/OS

Cluster Workflows

You can run a workflow to create a cluster that runs Mapping and other tasks on a **cloud platform cluster.**

The cluster workflow uses other elements that enable communication between the Data Integration Service and the cloud platform, such as a **cloud provisioning configuration** and a **cluster connection**.

Create cluster workflows to create clusters to run on the **Amazon AWS** or **Microsoft Azure** cloud platforms in a Hadoop environment.

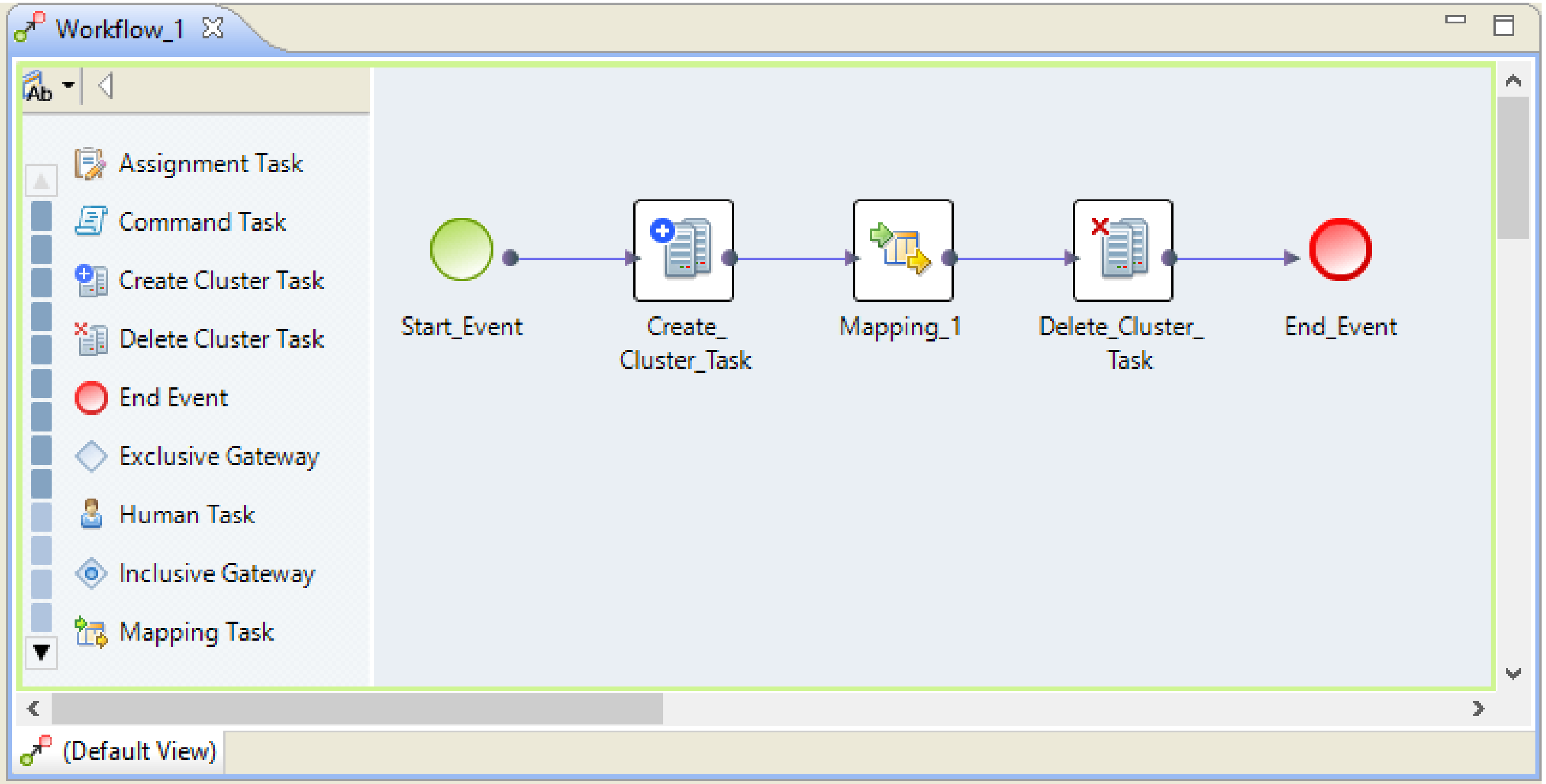
Create cluster workflows to create Databricks clusters to run in a Databricks environment.

On the Azure platform, you can create an ephemeral HDInsight cluster that accesses ADLS Gen2 resources. On the AWS platform, you can create an ephemeral Amazon EMR cluster to access S3, Redshift, and Snowflake resources.

Cluster Workflow Components

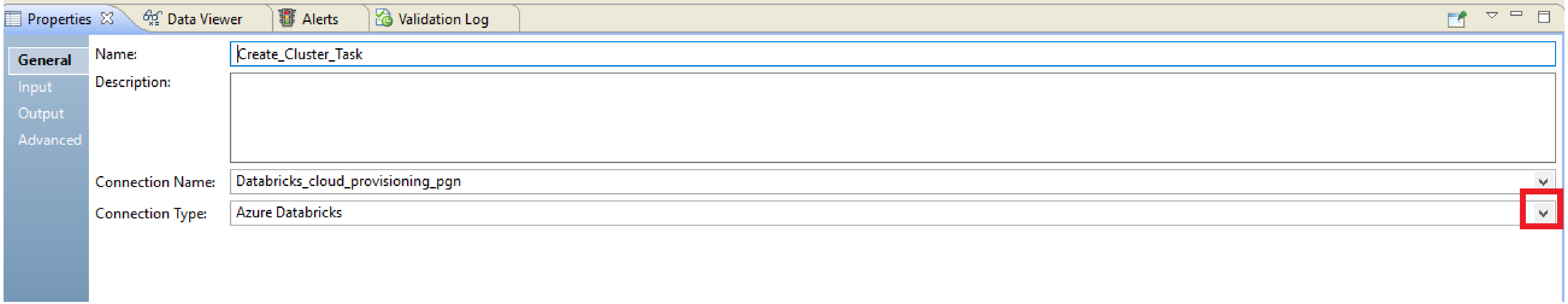
Creation of a cluster workflow requires administrator and developer tasks.

The following image shows the process to create, configure, and run a cluster workflow:



Create Cluster Task

In the create cluster task, assign the Cloud provisioning connection and select the connection type.



Name: Create_Cluster_Task

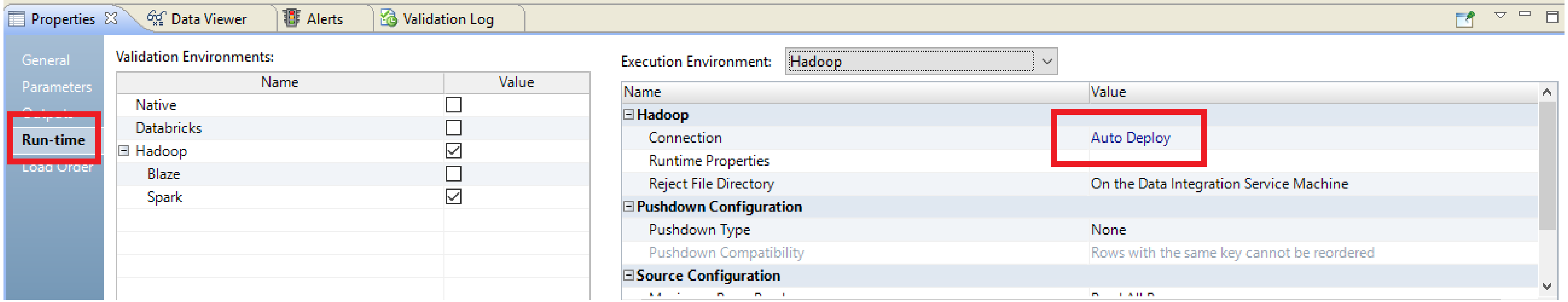
Description:

Connection Name: Databricks_cloud_provisioning_pgn

Connection Type: Azure Databricks

Mapping Task

During the mapping design, select the Run time Engine and the connection as **Auto Deploy**



Validation Environments:

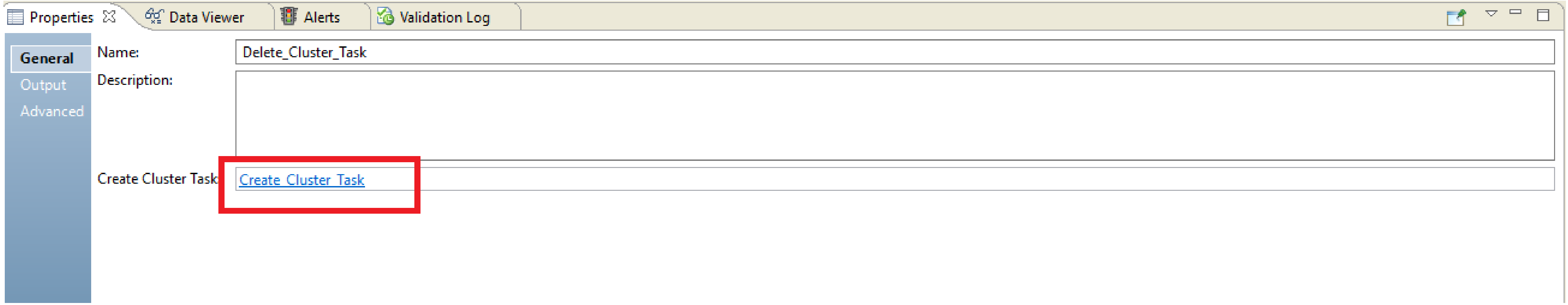
Name	Value
Native	□
Databricks	□
⊟ Hadoop	☑
Blaze	□
Spark	☑

Execution Environment: Hadoop

Name	Value
⊟ **Hadoop**	
Connection	Auto Deploy
Runtime Properties	
Reject File Directory	On the Data Integration Service Machine
⊟ **Pushdown Configuration**	
Pushdown Type	None
Pushdown Compatibility	Rows with the same key cannot be reordered
⊟ **Source Configuration**	

Delete Cluster Task

In the Delete cluster task, by default it select the create cluster task to delete.



Name: Delete_Cluster_Task

Description:

Create Cluster Task: Create_Cluster_Task

Command Line utilities

Informatica

- **How to list the ephemeral/transient cluster created using a cloud provisioning connection?**

Command: infacmd.sh ccps listclusters

```
Example: ./infacmd.sh ccps  -dn D_Galaxias -sn DIS -un Administrator –pd
*********  -cpcid Azure_CPC -sdn ISCBDM
```

- **How to delete the ephemeral/transient cluster ?**

Command: infacmd.sh ccps deleteClusters

```
Example: ./infacmd.sh ccps deleteClusters -dn Domain -sn DIS  -un Administrator -pd
*****-cpcid Azure_CPC -cids SampleCluster08__infa__1557834871483 -sdn ISCBDM
```

- **How to get the *-site.xmls to create the CCO of the ephemeral/transient cluster?**

Command: infacmd.sh ccps exportConfiguration

```
 Example: ./infacmd.sh cluster exportConfiguration -dn Domain -un Administrator -pd
** -sdn Native -cn
```

Ephemeral Cluster Support in BDS/DES

Informatica

Ephemeral Cluster Support – DES

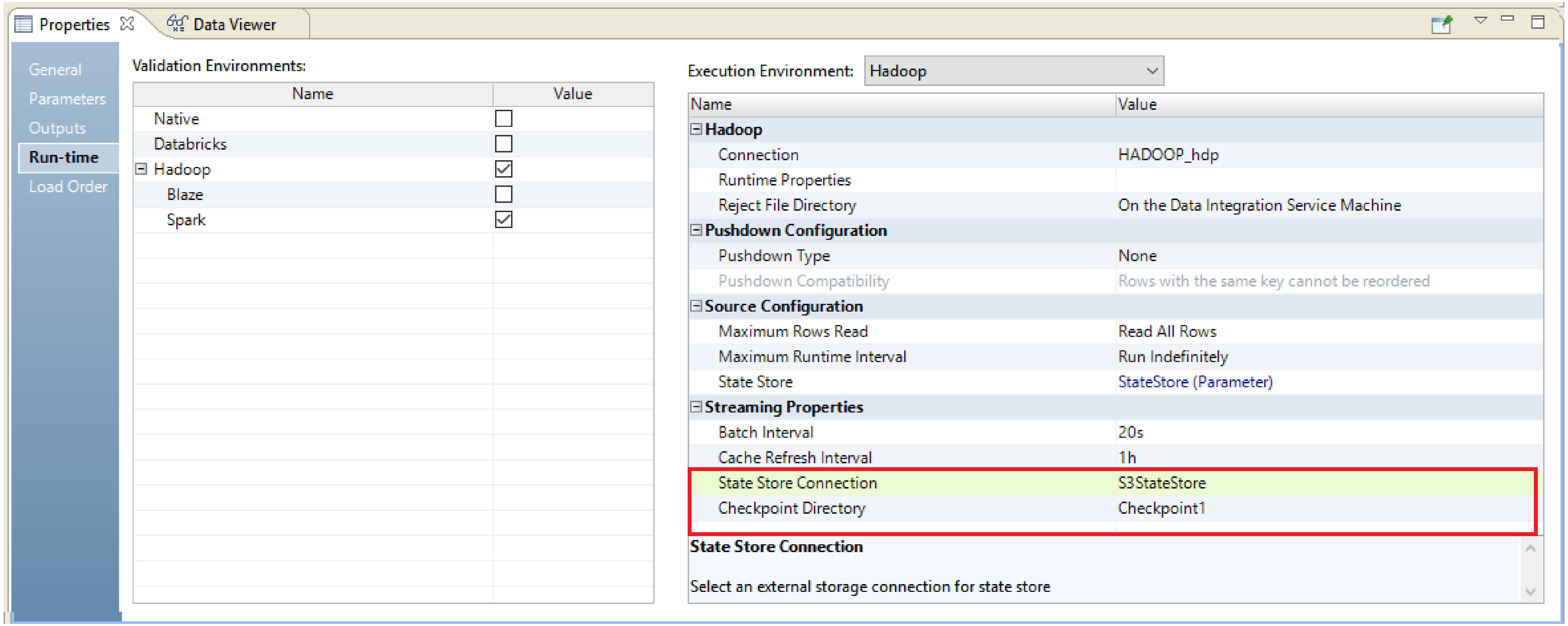
- As a streaming customer should be able to use Ephemeral cluster.

  - Resume mapping execution from where it had left, If cluster goes down and comes back.

- **Use case:** cost of running cloud based cluster all the time and hence customer wants to shut down the cluster towards the end of the day and start the next morning. They would like DES mappings to start from the offset where it had left at the time of shutting down the cluster.

- For supporting above use case User can use external storage for State Store and Checkpointing.
  - Supported External Storage
    - S3
    - ADLS

Continue …



Validation Environments:

Name	Value
Native	☐
Databricks	☐
⊟ Hadoop	☑
Blaze	☐
Spark	☑

Execution Environment: Hadoop

Name	Value
⊟ **Hadoop**	
Connection	HADOOP_hdp
Runtime Properties	
Reject File Directory	On the Data Integration Service Machine
⊟ **Pushdown Configuration**	
Pushdown Type	None
Pushdown Compatibility	Rows with the same key cannot be reordered
⊟ **Source Configuration**	
Maximum Rows Read	Read All Rows
Maximum Runtime Interval	Run Indefinitely
State Store	StateStore (Parameter)
⊟ **Streaming Properties**	
Batch Interval	20s
Cache Refresh Interval	1h
State Store Connection	S3StateStore
Checkpoint Directory	Checkpoint1

State Store Connection

Select an external storage connection for state store

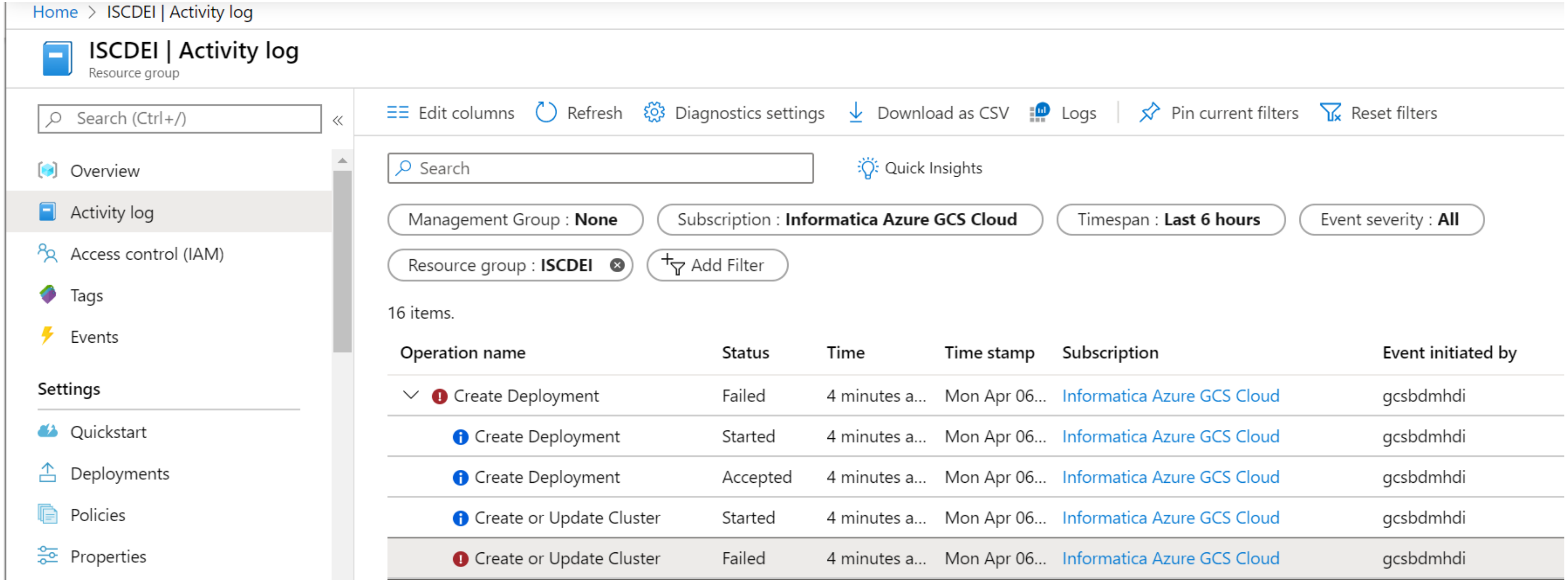
DEMO

Troubleshooting

Informatica

Logs to collect:

- 1. Cluster task log from $INFA_HOME/logs/<node>/services/DataIntegrationService/disLog/clustertask
- 2. Activity log from Azure portal: Click on the Resource Group -> Activity Log

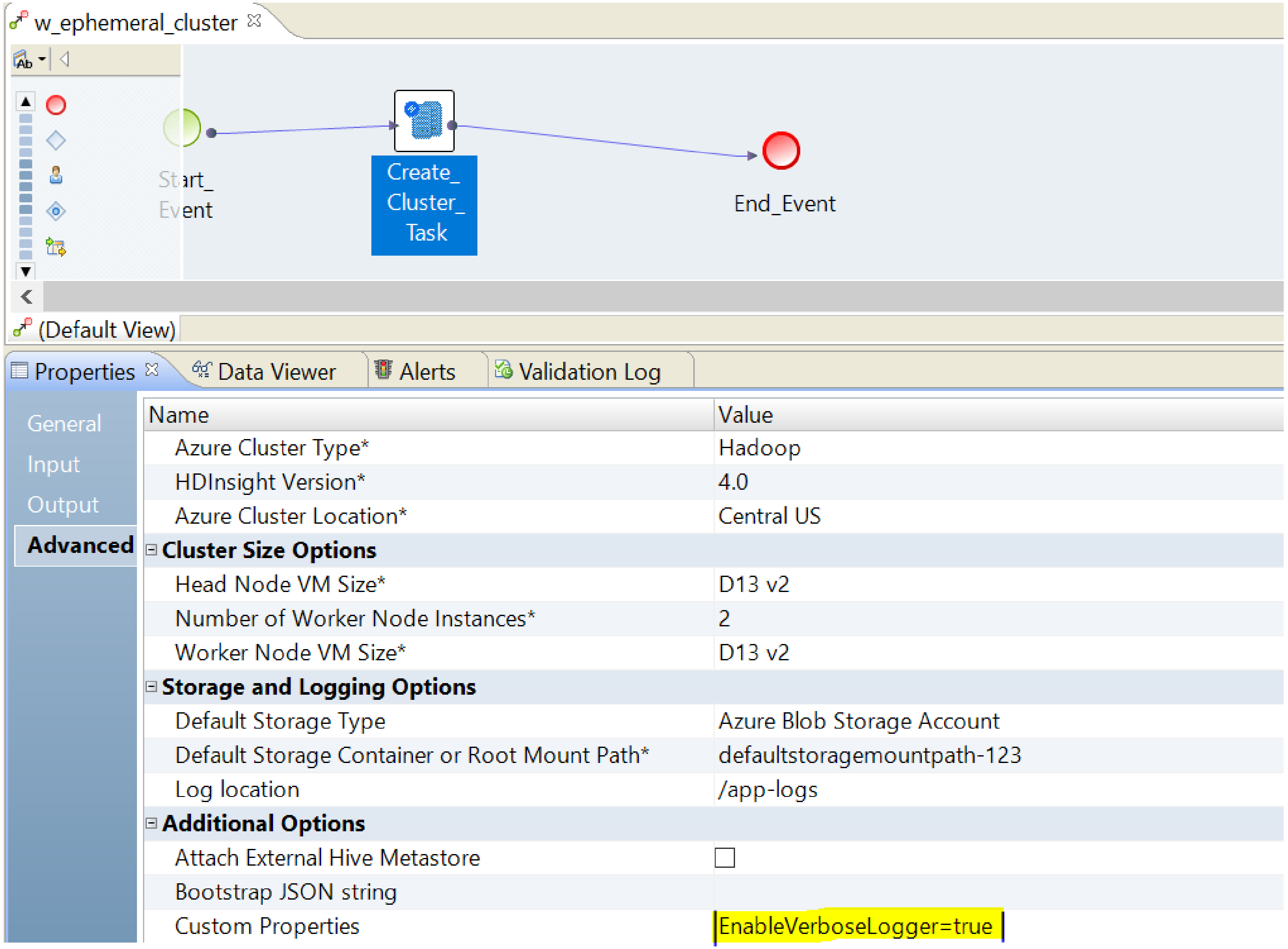


Management Group : **None** | Subscription : **Informatica Azure GCS Cloud** | Timespan : **Last 6 hours** | Event severity : **All** | Resource group : **ISCDEI**

16 items.

Operation name	Status	Time	Time stamp	Subscription	Event initiated by
Create Deployment	Failed	4 minutes a...	Mon Apr 06...	Informatica Azure GCS Cloud	gcsbdmhdi
Create Deployment	Started	4 minutes a...	Mon Apr 06...	Informatica Azure GCS Cloud	gcsbdmhdi
Create Deployment	Accepted	4 minutes a...	Mon Apr 06...	Informatica Azure GCS Cloud	gcsbdmhdi
Create or Update Cluster	Started	4 minutes a...	Mon Apr 06...	Informatica Azure GCS Cloud	gcsbdmhdi
Create or Update Cluster	Failed	4 minutes a...	Mon Apr 06...	Informatica Azure GCS Cloud	gcsbdmhdi

How to enable debug logging for create cluster task?



Name	Value
Azure Cluster Type*	Hadoop
HDInsight Version*	4.0
Azure Cluster Location*	Central US
⊟ **Cluster Size Options**	
Head Node VM Size*	D13 v2
Number of Worker Node Instances*	2
Worker Node VM Size*	D13 v2
⊟ **Storage and Logging Options**	
Default Storage Type	Azure Blob Storage Account
Default Storage Container or Root Mount Path*	defaultstoragemountpath-123
Log location	/app-logs
⊟ **Additional Options**	
Attach External Hive Metastore	□
Bootstrap JSON string	
Custom Properties	EnableVerboseLogger=true

Issue:1

- From cluster_task log we see below error message:

```
2020-04-05 21:48:19.300 <DTF-ThreadGroup-3-thread-7> SEVERE: Failed to
create the cluster due to the following error: [Failed to Create Azure
HDInsight Cluster with name [ephemeral_Cluster.azurehdinsight.net] due
to Cluster name specified is not
alphanumeric.]java.lang.RuntimeException: Failed to Create Azure
HDInsight Cluster with name [ephemeral_Cluster.azurehdinsight.net] due
to Cluster name specified is not alphanumeric.
```

To resolve:

Use only alphanumeric name as cluster name for the create cluster task

Issue:2

From cluster_task log we see below error message:

```
2020-03-31 17:56:00.276 <DTF-ThreadGroup-3-thread-10> SEVERE: Failed to
create the cluster due to the following error: [Failed to Create Azure
HDInsight Cluster with name [SampleCluster.azurehdinsight.net] due to
Status code 403, {"error":{"code":"AuthorizationFailed","message":"The
client '4b44ef9e-c162-4843-9771-8287eda6585c' with object id '4b44ef9e-
c162-4843-9771-8287eda6585c' does not have authorization to perform
action 'Microsoft.Resources/subscriptions/resourceGroups/resources/read'
over scope '/subscriptions/0da59e1d-ab2e-464e-8399-
e9d2620db07f/resourceGroups/ISCDEI' or the scope is invalid. If access
was recently granted, please refresh your credentials."}}]
```

To resolve:

Create a CONTRIBUTOR role and assign to the application you have created. We use application ID (Client ID) in the Cloud Provisioning Connection

Issue:3

- From mapping log we see below error message:

```
INFO: 20/04/06 20:24:35 INFO RetryInvocationHandler:
java.net.UnknownHostException: Invalid host name: local host is:
(unknown); destination host is: "hn1-
epheme.3cjfvk1rjohu3aimu4wtwpopoc.gx.internal.cloudapp.net":8050;
java.net.UnknownHostException; For more details see:
http://wiki.apache.org/hadoop/UnknownHost, while invoking
ApplicationClientProtocolPBClientImpl.getNewApplication over rm2
after 65 failover
```

To Resolve:

You need to have VPN configured between the VNET used by the cluster and Informatica server.

Or

Use Informatica Server which is running using the same VNET.

Issue:4

- From mapping log task log we see below error message:

```
Caused by: java.lang.RuntimeException: java.io.IOException:
java.net.ConnectException: Call From ISCDEIVM/10.0.0.4 to hn0-
sample.3cjfvk1rjohu3aimu4wtwpopoc.gx.internal.cloudapp.net:10020
failed on connection exception: java.net.ConnectException: Connection
refused; For more details see:
http://wiki.apache.org/hadoop/ConnectionRefused
```

To Resolve:

Make sure the port 10020 is open from Informatica Server machine.

Issue:5

- From the activity log we see below error message:

```
{\r\n \"code\": \"BadRequest\",\r\n \"message\": \"User input
 validation failed. Errors: The core-site config and the
 storageProfile contain same accounts.\"\r\n}

"
 }
```

To Resolve:

CR- BDM-29837 raised for this issue. The issue is resolved in the latest release of Informatica BDM/DEI.

Issue:6

- Mapping fails to run when configured with AUTO DEPLOY and OSP enabled

```
SEVERE: Data integration service failed to create DTM instancebecause
 of the following error:
 com.informatica.sdk.dtm.InvalidMappingException:
 [[DSCMN_10225] User [isuser] does not have permission [EXECUTE] on
 the following connections:[InternalHadoop_j_3L4NVQYIOIP1T]]
```

To Resolve:

CR BDM-31612 raised for this issue. We need to run the cluster workflow using Administrator user.

Useful Knowledgebase links &References

- https://kb.informatica.com/howto/6/Pages/23/593820.aspx

  Video KB:

  https://network.informatica.com/videos/3371

  https://www.youtube.com/watch?v=EbzfO70WP5Y&feature=youtu.be

- Cluster Workflows Overview
- Release Guide - Cluster Workflows and Ephemeral Clusters
- Implementing Informatica DEI with Ephemeral Clusters in a MS Azure Cloud Environment

Q&A

Informatica®

Thank You!

- Sampada Subnis
- Puneeth Natesha