March 9, 2021

Cloudera Data Platform Integration with DEI

Thirumurugan Swaminathan
Informatica Subject Matter Expert

Informatica™

Housekeeping Tips

- Today’s Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- 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.

Informatica™

Feature Rich Success Portal

Bootstrap trial and POC Customers

Enriched Customer Onboarding experience

Product Learning Paths and Weekly Expert Sessions

Informatica Concierge with Chatbot integrations

Tailored training and content recommendations

Informatica®

More Information

Success Portal

https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com

University

https://www.informatica.com/in/services-and-training/informatica-university.html

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

- Introduction
- Changes - CDP vs CDH vs HDP
- CDP Support with Informatica DEI
- Preparing Informatica DEI for CDP Integration
- Best Practices
- Troubleshooting
- Q&A

Introduction

- Cloudera Data Platform (CDP) is an Integrated Analytics and Data management Platform from Cloudera.
- Cloudera Runtime is the core open-source software distribution within CDP.
- CDP platform has two form factors:
  - CDP Public Cloud
  - CDP Private Cloud
- Cloudera Runtime Cluster is the technology backbone of both the form factors.

Changes - CDP Runtime Cluster vs CDH vs HDP

Component	CDP	CDH	HDP
Cluster Manager	CLOUDERA Manager	Cloudera Manager	Ambari
Authorization Manager	Apache Ranger	Sentry	Apache Ranger
Hive Managed Tables - Default Storage Format	Apache orc ORC - Bucketed	TEXTFILE	HDP 3.x ORC - Bucketed HDP 2.6.x TEXTFILE
Hive Managed Tables - Default Nature	Transactional	Non-Transactional	HDP 3.x Transactional HDP 2.6.x Non-Transactional
Hive Service - Supported Query Execution Framework(s)	Tez	MapReduce/Spark	HDP 3.x Tez HDP 2.6.x MapReduce/Tez/Spark

Informatica™

CDP Support with Informatica DEI

- CDP Support is with Informatica DEI is available from 10.4.1.1 onwards.

Deployment Type	Supported version	Informatica version
Public Cloud	7.2	From 10.4.1.2 onwards
Private Cloud	7.1.x	From 10.4.1.1 onwards

Execution Engine	10.4.1.1	10.4.1.2	10.4.1.3 (onwards)
Spark	S	S	S
Blaze	NS	NS	S

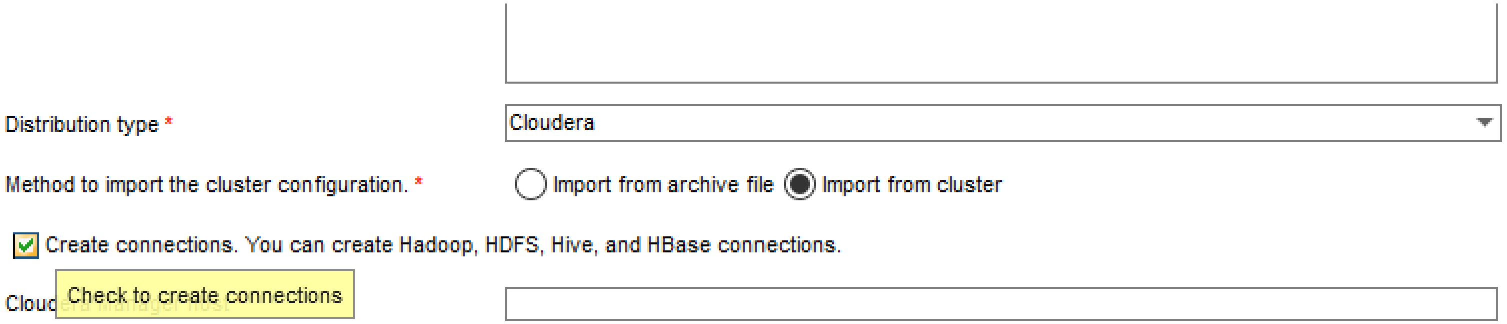
- Public Cloud Deployments in both AWS and Azure Ecosystems are supported.

Preparing Informatica DEI for CDP Integration

- CDP Integration variants in Informatica DEI
  - Onboard New CDP Runtime cluster
  - After CDH -> CDP upgrade
  - After HDP -> CDP upgrade

CDP Integration Configurations – New Cluster

- Create a new *'Cluster Configuration Object'* (CCO) for CDP cluster.
  - Use *'Import from Cluster'* (or) *'Import from Archive'* option in Admin console .
- Use 'Create Connections' option in CCO for creating Hadoop, Hive, HDFS, HBase connections.



Distribution type * Cloudera

Method to import the cluster configuration. * ○ Import from archive file ● Import from cluster

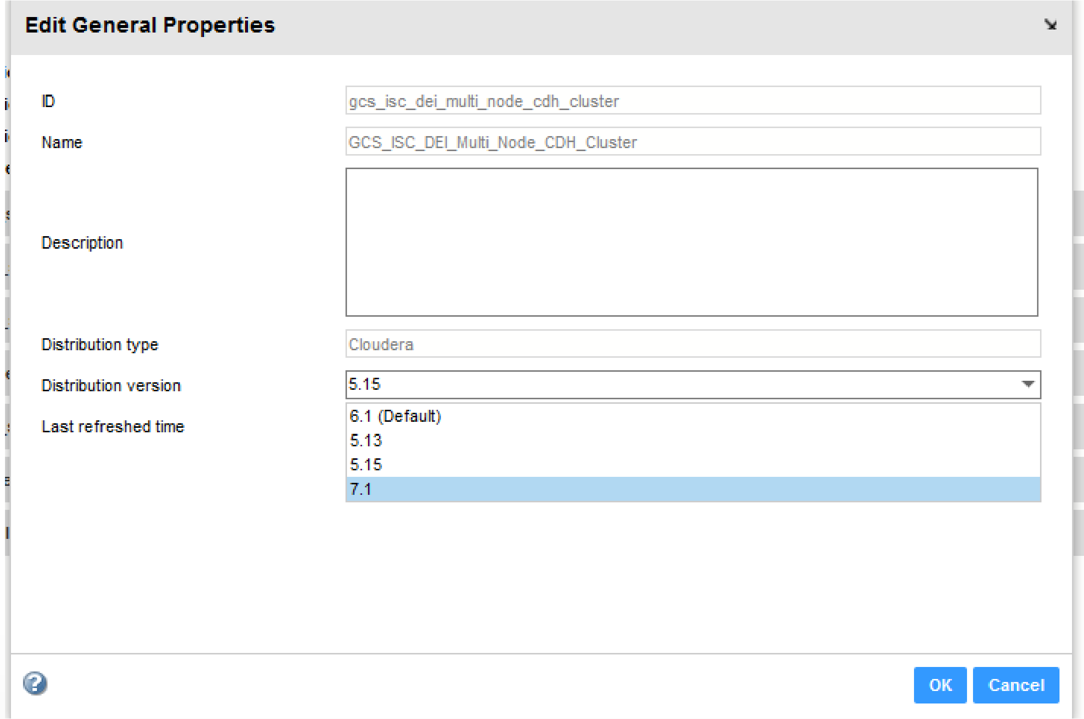
☑ Create connections. You can create Hadoop, HDFS, Hive, and HBase connections.

Cloudera Manager host

Check to create connections

DEI Integration Configurations - CDH to CDP

- After upgrading CDH cluster to CDP, perform following in Informatica Domain:
  - Refresh the existing *'Cluster Configuration Object'* (CCO), earlier created for CDH cluster
    - Use *'Import from Cluster'* (or) *'Import from Archive'* option in Admin console.
  - Ensure that Distribution version for the CCO is set to 7.1



Edit General Properties

ID: gcs_isc_dei_multi_node_cdh_cluster

Name: GCS_ISC_DEI_Multi_Node_CDH_Cluster

Description:

Distribution type: Cloudera

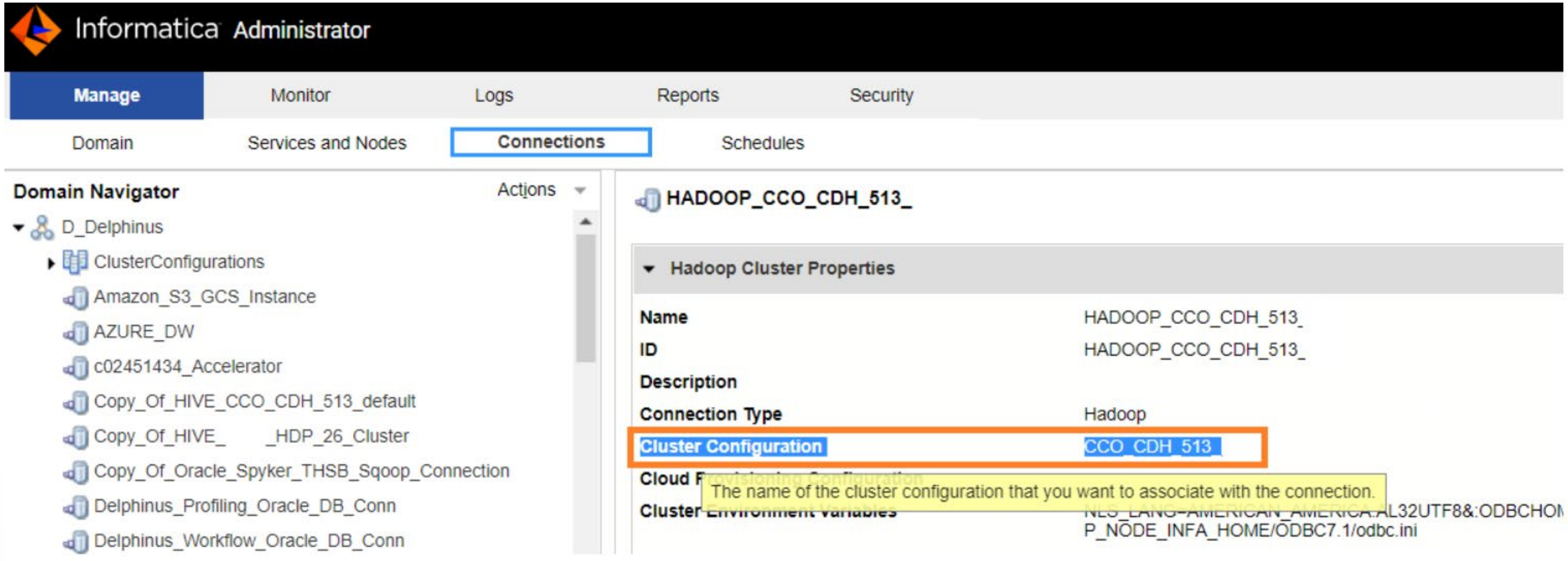
Distribution version: 5.15

6.1 (Default)
5.13
5.15
7.1

Last refreshed time:

CDP Integration Configurations - HDP to CDP

- After upgrading HDP cluster to CDP, perform the following steps in Informatica Domain:
  - Create a new *'Cluster Configuration Object'* (CCO) for CDP cluster.
    - Use *'Import from Cluster'* (or) *'Import from Archive'* option in Admin console .
  - Associate the existing Hadoop, Hive, HDFS, HBase connections of earlier HDP to the new CCO of CDP.



- D_Delphinus
  - ClusterConfigurations
  - Amazon_S3_GCS_Instance
  - AZURE_DW
  - c02451434_Accelerator
  - Copy_Of_HIVE_CCO_CDH_513_default
  - Copy_Of_HIVE_ _HDP_26_Cluster
  - Copy_Of_Oracle_Spyker_THSB_Sqoop_Connection
  - Delphinus_Profiling_Oracle_DB_Conn
  - Delphinus_Workflow_Oracle_DB_Conn

HADOOP_CCO_CDH_513_

Hadoop Cluster Properties

Name	HADOOP_CCO_CDH_513_
ID	HADOOP_CCO_CDH_513_
Description	
Connection Type	Hadoop
Cluster Configuration	CCO_CDH_513_
Cloud Provisioning Configuration	
Cluster Environment Variables	NLS_LANG=AMERICAN_AMERICA.AL32UTF8&:ODBCHOM P_NODE_INFA_HOME/ODBC7.1/odbc.ini

The name of the cluster configuration that you want to associate with the connection.

CDP Integration Configurations – HDP to CDP – Contd.

- If required, to automate the CCO update for the multiple connections, use below 'infacmd' commands:

  infacmd cluster listAssociatedConnections

  infacmd isp UpdateConnection

- *'infacmd cluster listAssociatedConnections'* - to get all the ‘Hive’/’HDFS’/’Hadoop’ connections, associated with a given HDP cluster’s CCO.

```
${infa_domain_home}/isp/bin/infacmd.sh cluster listAssociatedConnections -dn ${infa_domain_name} -un
${infacmd_user_name} -pd ${infacmd_user_password} -sdn ${infacmd_user_security_domain} -cn
${hdp_cco_name}
```

- Run ‘*infacmd isp UpdateConnection*’ for each of required connection to update CCO information.

```
${infa_domain_home}/isp/bin/infacmd.sh isp updateConnection -dn ${infa_domain_name} -un
${infacmd_user_name} -sdn ${infacmd_user_security_domain} -cn ${hive_or_hdfs_hadoop_conn_name} -o
"clusterConfigId='${new_cdp_cco_id}'"
```

For more information - infacmd cluster listAssociatedConnections , infacmd isp updateConnection , infacmd isp ListConnectionOptions (to view the connection attributes)

CDP Integration Configurations - General

- Remove the custom property - *'SparkSqoopDisSideInvocation=false'* - from the DIS, if present.
- When the cluster is using Kerberos Authentication, perform following:
  - Get *'krb5.conf'* and a 'user keytab' file for running jobs in CDP Runtime cluster from Admin Team.
  - Copy both *'krb5.conf'* & 'user keytab' files into Informatica Node(s) server machine(s).
  - Configure Keytab and User SPN details in the DIS & verify Kerberos connectivity - KB 523726.
  - Ensure that impersonation entries are configured DIS SPN user in *'core-site.xml'* of CDP cluster - KB 561374:
    - `property - hadoop.proxyuser.<DIS_Principal_user>.hosts`
      - `value =  <informatica_server_hostname(s)_&_Hadoop_data_node(s)>`
    - `property - hadoop.proxyuser.<DIS_Principal_user>.groups`
      - `value = <csv_list_of_groups_associated_with_Hadoop_conn_imp_users>`

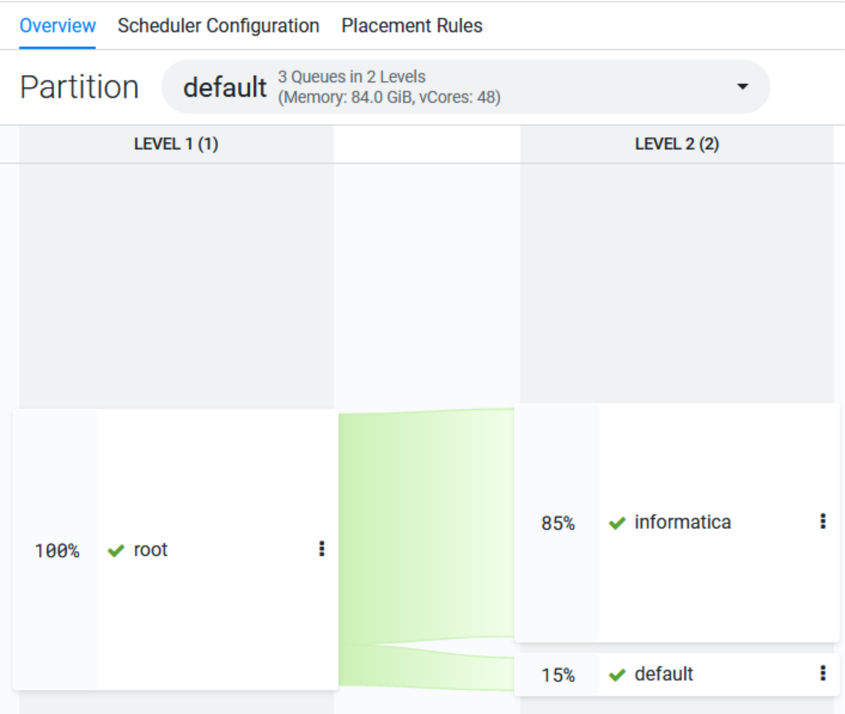
For more information - DEI - Integration Guide > CDP Integration Tasks

Best Practices – CDP Integration

- Recommended to use Spark as Execution engine for Informatica Mappings run in CDP Cluster.
- Use dedicated YARN Queue for Informatica Mappings in CDP Cluster.
- Enable Spark Dynamic Allocation for Informatica Spark Mappings.
- Setup *'Spark History Server'* in CDP cluster and integrate with Informatica DEI.

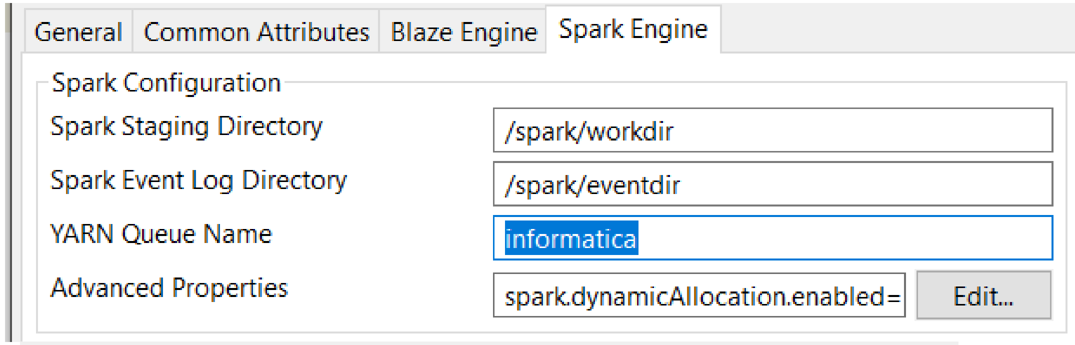
Best Practices – Using dedicated YARN Queue

- Create YARN Queue of desired capacity in CDP Runtime cluster for Informatica Mappings.
- From Informatica DEI, configure YARN Queue information in Hadoop and Sqoop connections:
  - Configure YARN Queue for Spark Jobs from Informatica DEI (KB 531634)
  - Configure YARN Queue for Sqoop Jobs from Informatica DEI (KB 531659)
  - Configure YARN Queue for Blaze Engine of Informatica DEI (KB 531589) (Pre-emption should be disabled)



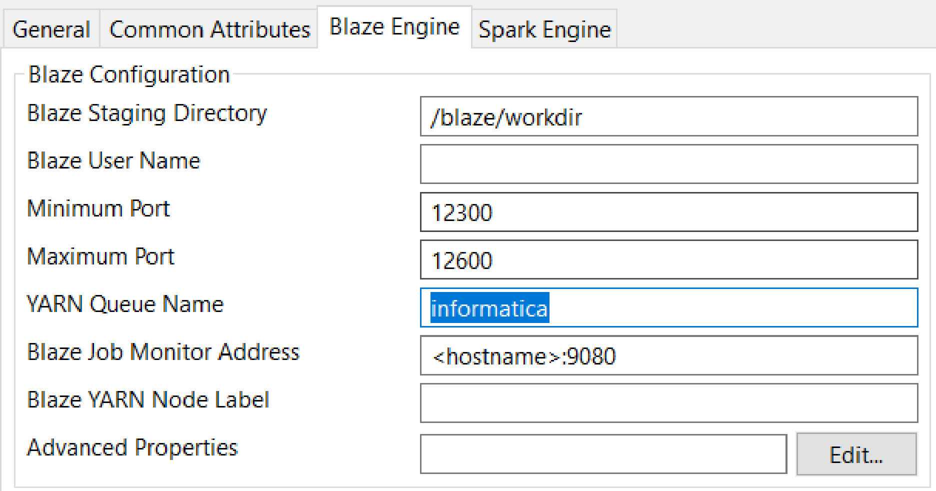
Partition default 3 Queues in 2 Levels (Memory: 84.0 GiB, vCores: 48)

LEVEL 1 (1)	LEVEL 2 (2)
100% ✓ root	85% ✓ informatica
	15% ✓ default



Spark Configuration

Spark Staging Directory	/spark/workdir
Spark Event Log Directory	/spark/eventdir
YARN Queue Name	informatica
Advanced Properties	spark.dynamicAllocation.enabled=



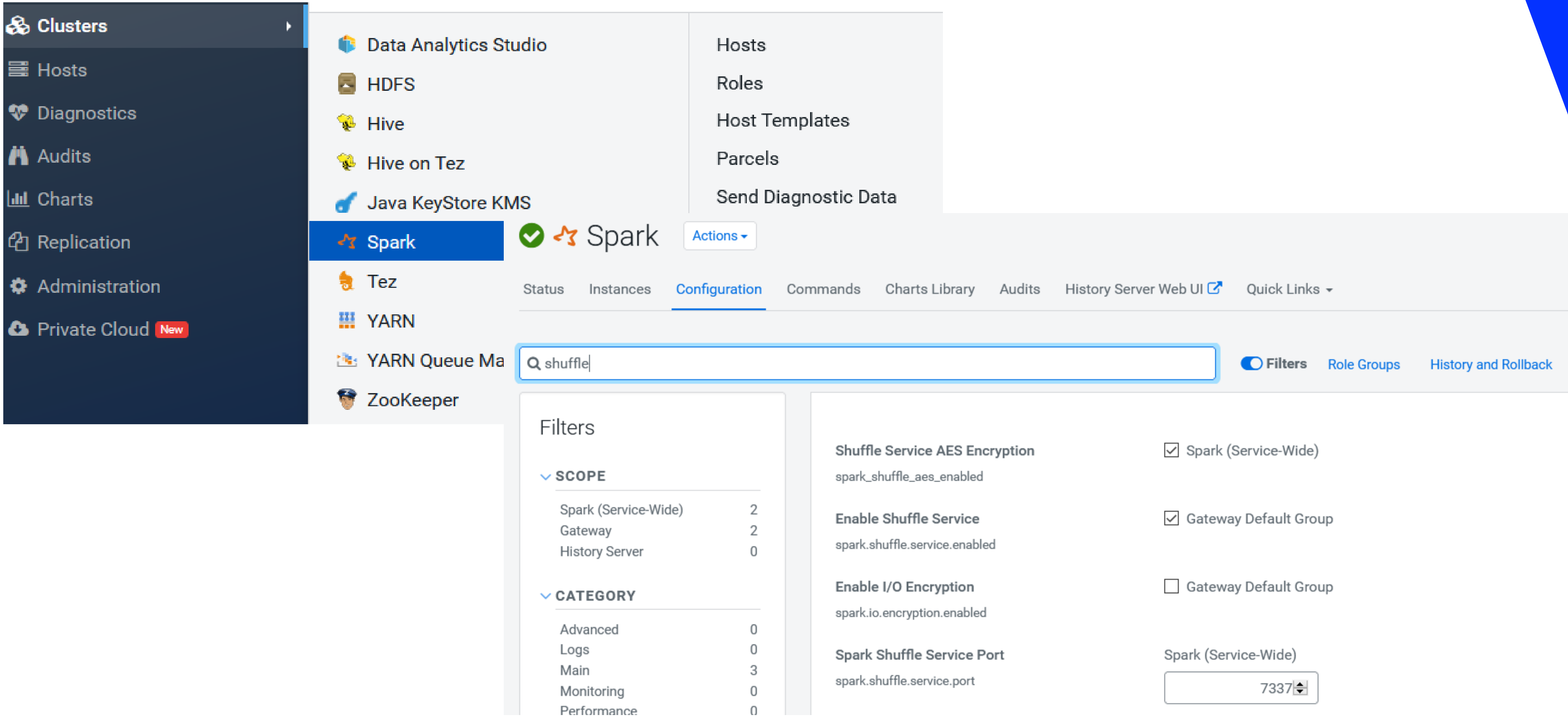
Blaze Configuration

Blaze Staging Directory	/blaze/workdir
Blaze User Name	
Minimum Port	12300
Maximum Port	12600
YARN Queue Name	informatica
Blaze Job Monitor Address	<hostname>:9080
Blaze YARN Node Label	
Advanced Properties	

Best Practices – Using Spark Dynamic Allocation

- By default, Informatica Spark mappings run in CDP Runtime cluster follows *'Spark Static Allocation'* approach with Executors.
- Recommended to enable *'Spark Dynamic Allocation'* method.
- For using *'Spark Dynamic Allocation'*, i.e., to automatically scale up/down the Spark executors, depending on the jobs:
  - Enable *'Spark Shuffle Service'* in the Hadoop cluster.
  - In *'Spark Engine > Advanced Properties'* section of Hadoop Connection, configure following properties:
    - **spark.dynamicAllocation.enabled=true**
    - **spark.shuffle.service.enabled= true**
    - **spark.shuffle.service.port=7337** (default for Spark Shuffle)
- Use *'spark.dynamicAllocation.maxExecutors'* to limit the maximum number of executors launched by Informatica Spark mapping, if needed.
- More Information - KB 516652

Best Practices – Using Spark Dynamic Allocation



Filters

SCOPE	
Spark (Service-Wide)	2
Gateway	2
History Server	0

CATEGORY	
Advanced	0
Logs	0
Main	3
Monitoring	0
Performance	0

Shuffle Service AES Encryption
spark_shuffle_aes_enabled
☑ Spark (Service-Wide)

Enable Shuffle Service
spark.shuffle.service.enabled
☑ Gateway Default Group

Enable I/O Encryption
spark.io.encryption.enabled
☐ Gateway Default Group

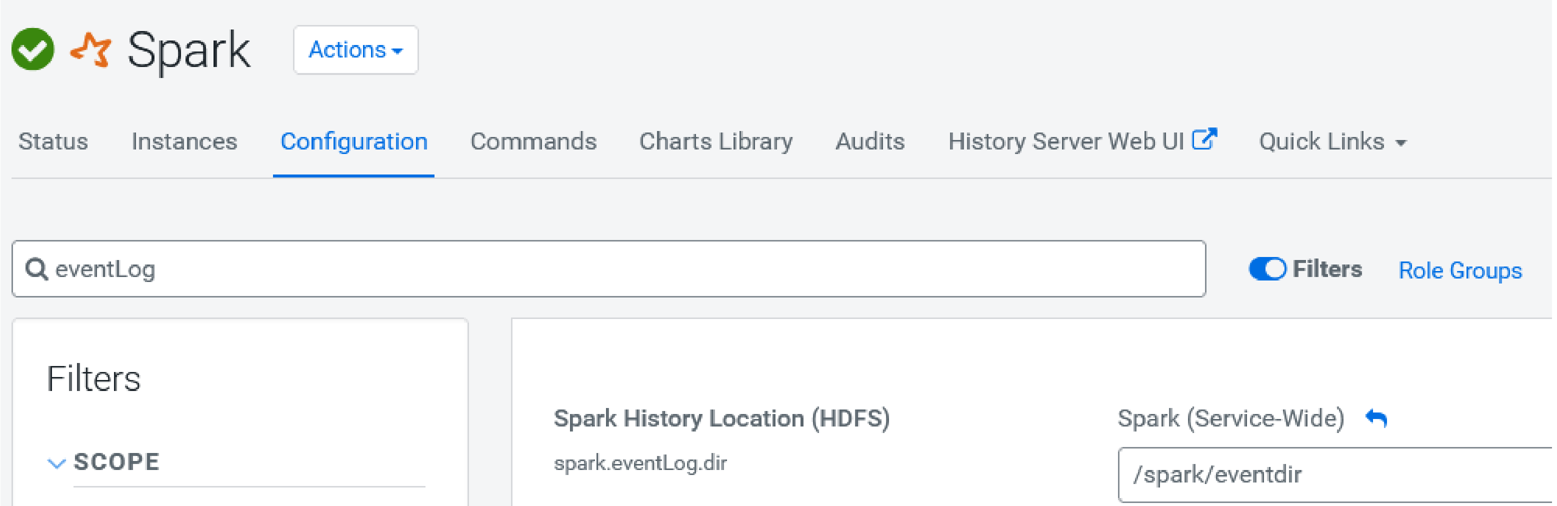
Spark Shuffle Service Port
spark.shuffle.service.port
Spark (Service-Wide)
7337

Informatica

Best Practices – Integrating Spark History Server

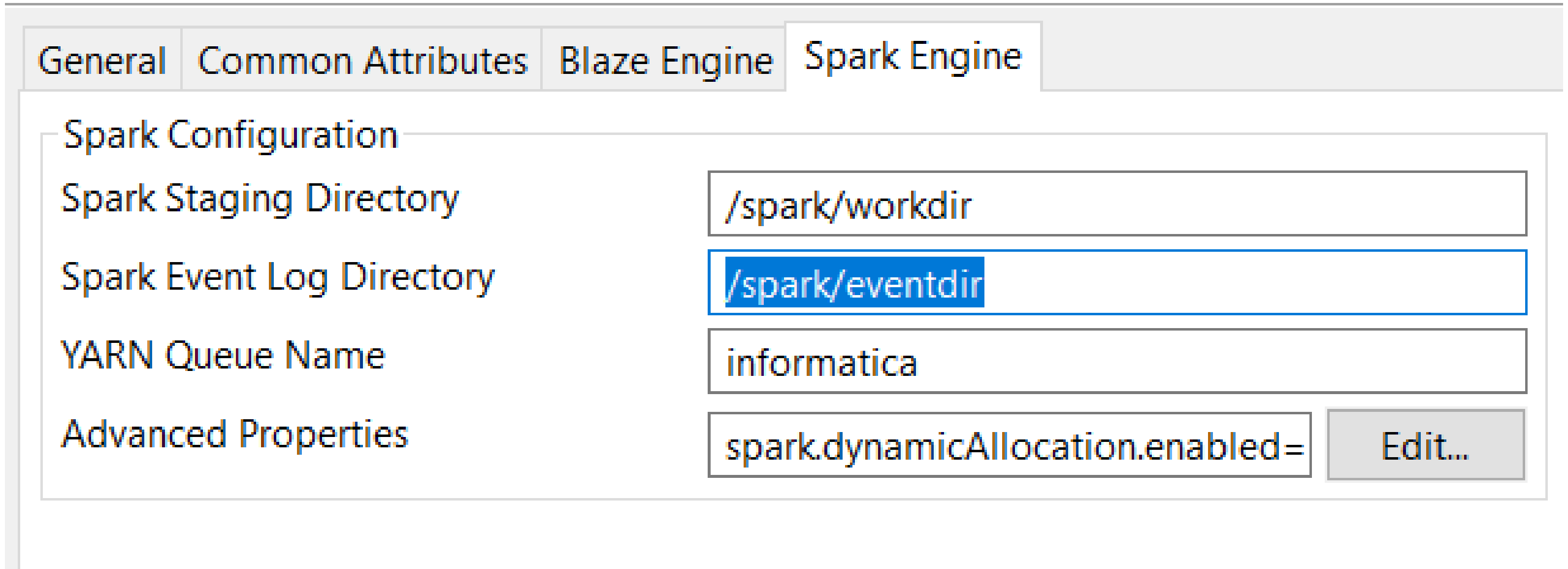
- When a Informatica Spark mapping is in ‘Running' state, execution details & performance metrics
  - Jobs, Stages, DAGs, Number of executors launched , Memory utilized, Volume of Shuffle read/written and so on..
  - Can be accessed through *‘Application Master Tracking URL’* in YARN RM.
  - However, once the mapping execution finishes, those details would be not be accessible.
- For viewing Historical execution and performance details for Informatica Spark mappings :
  - Ensure Spark History Server is setup and active in CDP cluster.
  - Add following property in *'Spark Engine > Advanced Properties'* section of Hadoop connection:
    - **spark.yarn.historyServer.address=http://[spark_history_server_host]:[spark_history_server_port]** *(Default Port: 18088)*
  - Event Log location used in Information Hadoop connection should be same as the *'spark.eventLog.dir'* in Spark History server.
  - 'write' permissions on the HDFS *‘Event Log Location’* for Impersonation user of Informatica Spark mappings.

Best Practices – Integrating Spark History Server



Spark History Location (HDFS)
spark.eventLog.dir

Spark (Service-Wide): /spark/eventdir



Spark Configuration

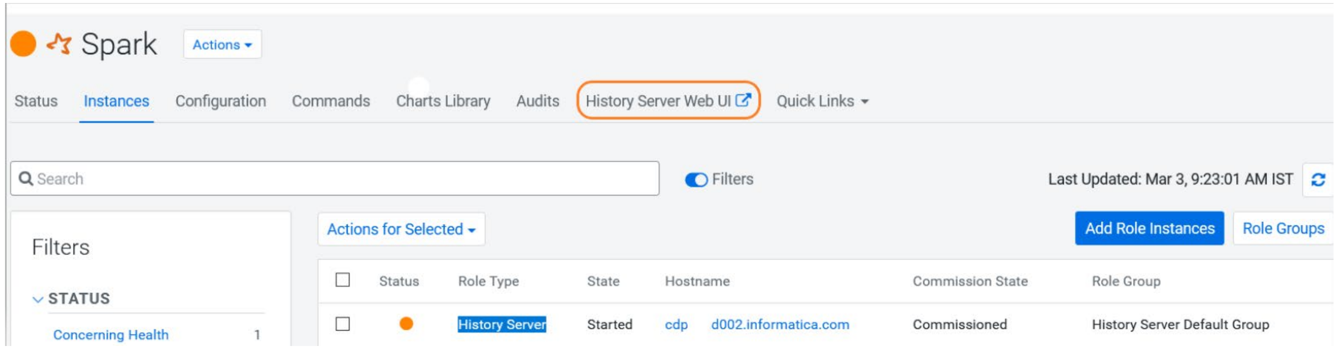
Spark Staging Directory: /spark/workdir

Spark Event Log Directory: /spark/eventdir

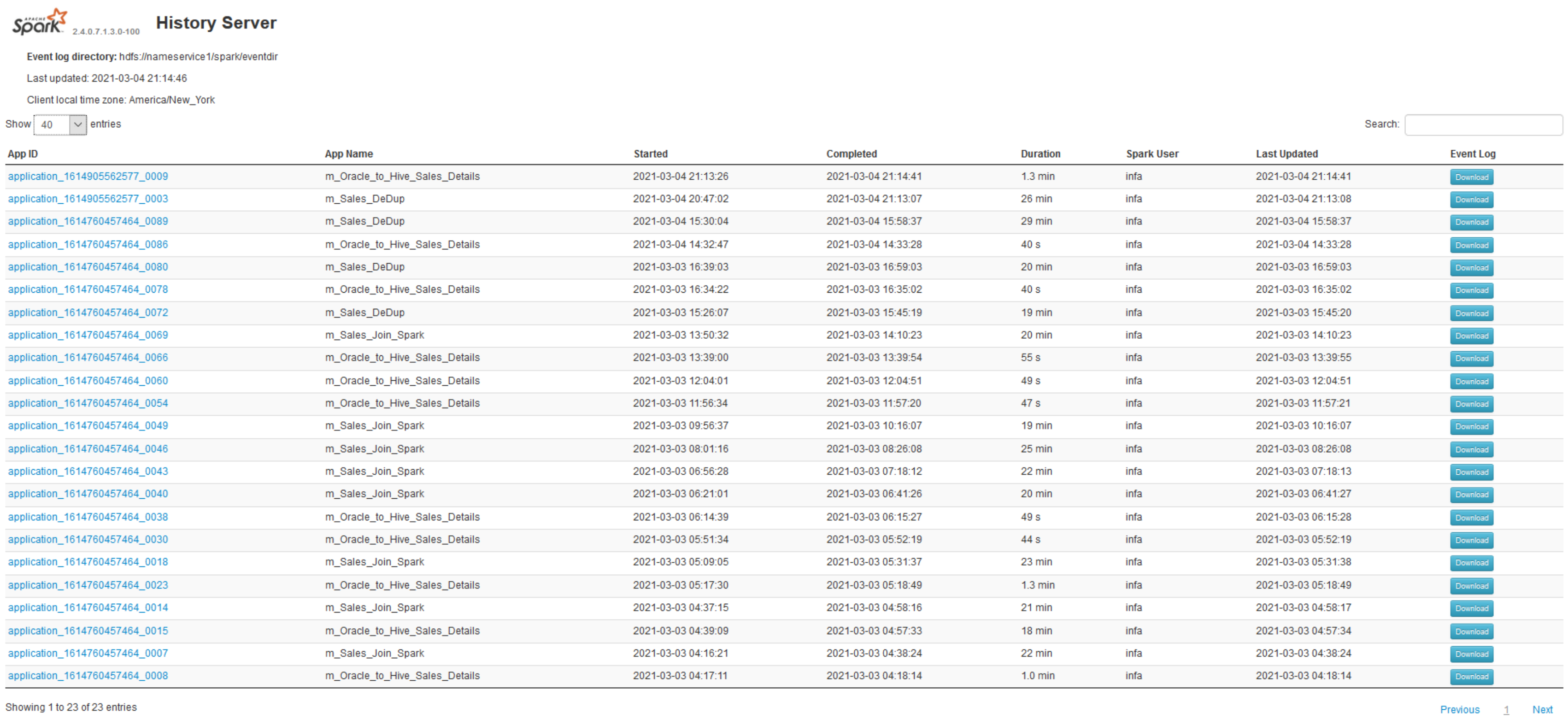
YARN Queue Name: informatica

Advanced Properties: spark.dynamicAllocation.enabled=

Best Practices – Integrating Spark History Server



□	Status	Role Type	State	Hostname		Commission State	Role Group
□	●	History Server	Started	cdp	d002.informatica.com	Commissioned	History Server Default Group



History Server

Event log directory: hdfs://nameservice1/spark/eventdir

Last updated: 2021-03-04 21:14:46

Client local time zone: America/New_York

Show 40 entries

App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
application_1614905562577_0009	m_Oracle_to_Hive_Sales_Details	2021-03-04 21:13:26	2021-03-04 21:14:41	1.3 min	infa	2021-03-04 21:14:41	Download
application_1614905562577_0003	m_Sales_DeDup	2021-03-04 20:47:02	2021-03-04 21:13:07	26 min	infa	2021-03-04 21:13:08	Download
application_1614760457464_0089	m_Sales_DeDup	2021-03-04 15:30:04	2021-03-04 15:58:37	29 min	infa	2021-03-04 15:58:37	Download
application_1614760457464_0086	m_Oracle_to_Hive_Sales_Details	2021-03-04 14:32:47	2021-03-04 14:33:28	40 s	infa	2021-03-04 14:33:28	Download
application_1614760457464_0080	m_Sales_DeDup	2021-03-03 16:39:03	2021-03-03 16:59:03	20 min	infa	2021-03-03 16:59:03	Download
application_1614760457464_0078	m_Oracle_to_Hive_Sales_Details	2021-03-03 16:34:22	2021-03-03 16:35:02	40 s	infa	2021-03-03 16:35:02	Download
application_1614760457464_0072	m_Sales_DeDup	2021-03-03 15:26:07	2021-03-03 15:45:19	19 min	infa	2021-03-03 15:45:20	Download
application_1614760457464_0069	m_Sales_Join_Spark	2021-03-03 13:50:32	2021-03-03 14:10:23	20 min	infa	2021-03-03 14:10:23	Download
application_1614760457464_0066	m_Oracle_to_Hive_Sales_Details	2021-03-03 13:39:00	2021-03-03 13:39:54	55 s	infa	2021-03-03 13:39:55	Download
application_1614760457464_0060	m_Oracle_to_Hive_Sales_Details	2021-03-03 12:04:01	2021-03-03 12:04:51	49 s	infa	2021-03-03 12:04:51	Download
application_1614760457464_0054	m_Oracle_to_Hive_Sales_Details	2021-03-03 11:56:34	2021-03-03 11:57:20	47 s	infa	2021-03-03 11:57:21	Download
application_1614760457464_0049	m_Sales_Join_Spark	2021-03-03 09:56:37	2021-03-03 10:16:07	19 min	infa	2021-03-03 10:16:07	Download
application_1614760457464_0046	m_Sales_Join_Spark	2021-03-03 08:01:16	2021-03-03 08:26:08	25 min	infa	2021-03-03 08:26:08	Download
application_1614760457464_0043	m_Sales_Join_Spark	2021-03-03 06:56:28	2021-03-03 07:18:12	22 min	infa	2021-03-03 07:18:13	Download
application_1614760457464_0040	m_Sales_Join_Spark	2021-03-03 06:21:01	2021-03-03 06:41:26	20 min	infa	2021-03-03 06:41:27	Download
application_1614760457464_0038	m_Oracle_to_Hive_Sales_Details	2021-03-03 06:14:39	2021-03-03 06:15:27	49 s	infa	2021-03-03 06:15:28	Download
application_1614760457464_0030	m_Oracle_to_Hive_Sales_Details	2021-03-03 05:51:34	2021-03-03 05:52:19	44 s	infa	2021-03-03 05:52:19	Download
application_1614760457464_0018	m_Sales_Join_Spark	2021-03-03 05:09:05	2021-03-03 05:31:37	23 min	infa	2021-03-03 05:31:38	Download
application_1614760457464_0023	m_Oracle_to_Hive_Sales_Details	2021-03-03 05:17:30	2021-03-03 05:18:49	1.3 min	infa	2021-03-03 05:18:49	Download
application_1614760457464_0014	m_Sales_Join_Spark	2021-03-03 04:37:15	2021-03-03 04:58:16	21 min	infa	2021-03-03 04:58:17	Download
application_1614760457464_0015	m_Oracle_to_Hive_Sales_Details	2021-03-03 04:39:09	2021-03-03 04:57:33	18 min	infa	2021-03-03 04:57:34	Download
application_1614760457464_0007	m_Sales_Join_Spark	2021-03-03 04:16:21	2021-03-03 04:38:24	22 min	infa	2021-03-03 04:38:24	Download
application_1614760457464_0008	m_Oracle_to_Hive_Sales_Details	2021-03-03 04:17:11	2021-03-03 04:18:14	1.0 min	infa	2021-03-03 04:18:14	Download

Showing 1 to 23 of 23 entries

Informatica™

Demo

Troubleshooting – Spark Configuration and Performance

- Verbose Init Logs & Spark Yarn Application Logs
  - Verbose Initialization - KB 532357
  - Hadoop YARN Application Logs - KB 524731
- Thread Dumps from Spark History Server UI (while Spark mapping is in *'Running'* state).
- Spark Driver/Executor Container Size (KB 526094)
- Join Broadcast settings (KB 531936, KB 565352)

More Information

- Cloudera Data Platform
  - Cloudera Public Cloud Base
  - Cloudera Private Cloud Base
  - CDP Runtime Cluster - Adding YARN Queues
- Spark Performance Tuning
  - Spark Performance Tuning and Sizing Guide
  - YouTube - Informatica - Spark History Server Integration
- Informatica DEI - YARN Queue Configuration
  - Configure YARN Queue for Spark Jobs from Informatica DEI (KB 531634)
  - Configure YARN Queue for Sqoop Jobs from Informatica DEI (KB 531659)
  - Configure YARN Queue for Blaze Engine of Informatica DEI (KB 531589)
- Sqoop Mappings and Connections
  - Sqoop Performance Tuning Guide
  - HOW TO: Configure Sqoop for Oracle Databases in Informatica Developer (KB 500711)

Q & A

Thank You