Cisco Workload Automation Hive Adapter Guide

Version 6.3

First Published: August, 2015
Last Updated: October 10, 2017

Cisco Systems, Inc. www.cisco.com

THE SPECIFICATIONS AND INFORMATION REGARDING THE PRODUCTS IN THIS MANUAL ARE SUBJECT TO CHANGE WITHOUT NOTICE. ALL STATEMENTS, INFORMATION, AND RECOMMENDATIONS IN THIS MANUAL ARE BELIEVED TO BE ACCURATE BUT ARE PRESENTED WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. USERS MUST TAKE FULL RESPONSIBILITY FOR THEIR APPLICATION OF ANY PRODUCTS.

THE SOFTWARE LICENSE AND LIMITED WARRANTY FOR THE ACCOMPANYING PRODUCT ARE SET FORTH IN THE INFORMATION PACKET THAT SHIPPED WITH THE PRODUCT AND ARE INCORPORATED HEREIN BY THIS REFERENCE. IF YOU ARE UNABLE TO LOCATE THE SOFTWARE LICENSE OR LIMITED WARRANTY, CONTACT YOUR CISCO REPRESENTATIVE FOR A COPY.

NOTWITHSTANDING ANY OTHER WARRANTY HEREIN, ALL DOCUMENT FILES AND SOFTWARE OF THESE SUPPLIERS ARE PROVIDED “AS IS” WITH ALL FAULTS. CISCO AND THE ABOVE-NAMED SUPPLIERS DISCLAIM ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING, WITHOUT LIMITATION, THOSE OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE.

IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THIS MANUAL, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Any Internet Protocol (IP) addresses and phone numbers used in this document are not intended to be actual addresses and phone numbers. Any examples, command display output, network topology diagrams, and other figures included in the document are shown for illustrative purposes only. Any use of actual IP addresses or phone numbers in illustrative content is unintentional and coincidental.

All printed copies and duplicate soft copies are considered un-Controlled copies and the original on-line version should be referred to for latest version.

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco website at www.cisco.com/go/offices.

Preface

This guide describes the installation, configuration, and usage of the Hive Adapter with Cisco Workload Automation (CWA).

Audience

This guide is for administrators who install and configure the Hive Adapter for use with Cisco Workload Automation, and who troubleshoot CWA installation and requirements issues.

Related Documentation

See the *Cisco Workload Automation Documentation Overview* for your release on cisco.com at:

> http://www.cisco.com/c/en/us/support/cloud-systems-management/tidal-enterprise-scheduler/products-documentation-roadmaps-list.html

...for a list of all Cisco Workload Automation guides.

Note: We sometimes update the documentation after original publication. Therefore, you should also review the documentation on Cisco.com for any updates.

Obtaining Documentation and Submitting a Service Request

For information on obtaining documentation, submitting a service request, and gathering additional information, see What's New in Cisco Product Documentation at:

http://www.cisco.com/en/US/docs/general/whatsnew/whatsnew.html.

Subscribe to What's New in Cisco Product Documentation, which lists all new and revised Cisco technical documentation, as an RSS feed and deliver content directly to your desktop using a reader application. The RSS feeds are a free service.

Document Change History

The table below provides the revision history for the *Cisco Workload AuTomation Hive Adapter Guide*.

Version Number	Issue Date	Reason for Change
6.1.0	October 2012	New Cisco version.
6.2.1	June 2014	Available in online Help only.

Document Change History

Version Number	Issue Date	Reason for Change
6.2.1 SP2	June 2015	Configuration provided in the *Cisco Workload Automation Installation Guide*; usage provided in online Help only.
6.2.1 SP3	May 2016	Consolidated all Hive Adapter documentation into one document.
6.3 Beta	June 2016	Rebranded “Cisco Tidal Enterprise Scheduler (TES)” to “Cisco Workload Automation (CWA)”. Added the new Installing the Hadoop Client Libraries section. Updates to the Configuring the Hive Adapter section. Updates to the Defining a Connection section. Added the service.props configuration chapter. Updated and corrected the documentation for the 6.3 release.

Contents

1

Introducing the Hive Adapter

This chapter provides an overview of the Hive Adapter and its requirements:

Overview

The Cisco Workload Automation Hive Adapter provides the automation of HiveQL commands as part of the cross-platform process organization between Cisco Workload Automation (CWA) and the CWA Hadoop Cluster. The Adapter is designed using the same user interface approach as other Cisco Workload Automation adapter jobs, seamlessly integrating Hadoop Hive data management into existing operation processes.

The Hive Adapter allows you to access and manage data stored in the Hadoop Distributed File System (HDFS™) using Hive's query language, HiveQL. HiveQL syntax is similar to SQL standard syntax.

The Have Adapter, in conjunction with Cisco Workload Automation, can be used to define, launch, control, and monitor HiveQL commands submitted to Hive via JDBC on a scheduled basis. The Adapter integrates seamlessly in an enterprise scheduling environment.

The Hive adapter includes the following features:

- Connection management to monitor system status with a live connection to the Hive Server via JDBC
- Hive job and event management includes the following:
  - Scheduling and monitoring of HiveQL commands from a centralized work console with Cisco Workload Automation
  - Dynamic runtime overrides for parameters and values passed to the HiveQL command
  - Output-formatting options to control the results, including table, XML, and CSV
  - Defined dependencies and events with Cisco Workload Automation for scheduling control
  - Runtime MapReduce parameters overrides if the HiveQL command results in a MapReduce job.

Prerequisites

- Hive version must be 0.9.0 or above.
- Hive Server
- The Hive Server must be fully operational and accessible to the Hive Adapter.

Cisco Workload Automation Adapters require Java 8. (Refer to *Cisco Workload Automation Compatibility Guide* for further details).

Software Requirements

The 6.3 Hive Adapter is installed with the CWA 6.3 master and client and cannot be used with an earlier CWA version.

Refer to your *Cisco Workload Automation Compatibility Guide* for a complete list of hardware and software requirements.

2

Configuring the Hive Adapter

Overview

The Hive Adapter software is installed as part of a standard installation of Cisco Workload Automation. However, before the Hive Adapter can be used, the following configuration procedures must be completed:

- Installing the Hadoop Client Libraries, page 7 – Install the necessary Hadoop client libraries for Hive.
- Configuring the Hive Adapter, page 9 – Add optional configuration properties to the service.props file.
- Licensing an Adapter, page 9 – Apply the license to the Hive Adapter. You cannot define a Hive connection until you have applied the Hive license.
- Securing the Adapter, page 10 – Define Hive users that the Adapter can use to establish authenticated sessions with the Hive server and permit requests to be made on behalf of the authenticated account.
- Defining a Connection, page 13 – Define a Hive connection so the master can communicate with the Hive server.

See Configuring service.props for details about configuring service.props to control such things as polling, output, and log gathering.

Installing the Hadoop Client Libraries

Hadoop client libraries are required for processing the Hadoop-related DataMover, Hive, MapReduce, and Sqoop jobs. As of CWA 6.3, Hadoop libraries are not included with CWA. Instead, we provide a Maven script (POM.xml) to install the required libraries.

If you do not already have Maven, you must download and install it. Obtain the POM.xml file from the folder/directory named " Hadoop" in the CD and run the file script to download the required Hadoop client libraries. Instructions for obtaining Maven and downloading the Hadoop libraries are included in these sections:

- Installing Maven, page 7
- Downloading the Hadoop Client Library, page 8

Note: The instructions here are for Windows.

Installing Maven

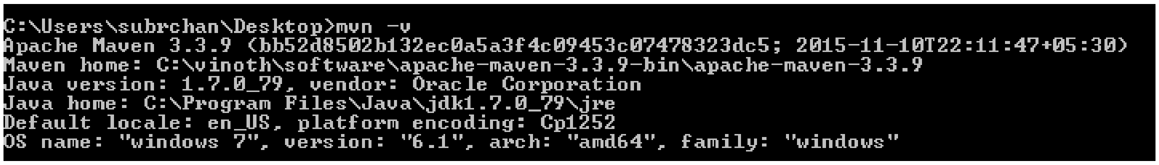
If you do not have Maven installed, follow the instructions below.

Maven Prerequisites

- JDK must be installed.
- The JAVA_HOME environment variable must be set and point to your JDK.

To download and install Maven:

1. Download maven 3 or above from https://maven.apache.org/download.cgi.
2. Unzip apache-maven-<3 or above>-bin.zip.
3. Add the bin directory of the created directory (for example, apache-maven-3.3.9) to the PATH environment variable
4. Confirm a successful Maven installation by running the **mvn -v** command in a new shell. The result should look similar to this:



```
C:\Users\subrchan\Desktop>mvn -v
Apache Maven 3.3.9 (bb52d8502b132ec0a5a3f4c09453c07478323dc5; 2015-11-10T22:11:47+05:30)
Maven home: C:\vinoth\software\apache-maven-3.3.9-bin\apache-maven-3.3.9
Java version: 1.7.0_79, vendor: Oracle Corporation
Java home: C:\Program Files\Java\jdk1.7.0_79\jre
Default locale: en_US, platform encoding: Cp1252
OS name: "windows 7", version: "6.1", arch: "amd64", family: "windows"
```

Downloading the Hadoop Client Library

With Maven installed, you can now download the Hadoop client library. Maven scripts (POM.xml) are provided for the following distributions of Hadoop:

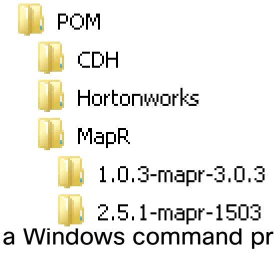
Hadoop Distribution Type	Versions
Cloudera	CDH5
Hortonworks	HDP 2.4.x
MapR	5.1.0

Note: The *Cisco Workload Automation Compatibility Guide* contains the most current version information.

To download and install the Hadoop client library

1. Download the POM.zip file. This file is provided in the /Hadoop directory in the CWA 6.3 distribution package.
2. Unzip the POM.zip.

   The POM xml files needed by Maven are saved in the directory structure shown here:



```
POM
  CDH
  Hortonworks
  MapR
    1.0.3-mapr-3.0.3
    2.5.1-mapr-1503
```

3. Open a Windows command prompt and navigate to the directory for the Hadoop distribution in which you are interested. For example, navigate to the CDH directory if you want to download Hadoop client libraries for Cloudera.
4. Edit the POM.xml file to mention exact versions of MapR, Hadoop, Hive, and Sqoop that you are using. For example, for Cloudera the required properties could be edited as shown below:

```
<properties>
<Hadoop.version>2.6.0-cdh5.6.0</Hadoop.version>
<Hive.version>1.1.0-cdh5.7.0</Hive.version>
<Sqoop.version>1.4.6-cdh5.6.0</Sqoop.version>
</properties>
```

For MapR it is also necessary to mention the version of MapR used, as shown in the following example:

```
<properties>
<Hadoop.version>2.7.0-mapr-1602</Hadoop.version>
<Hive.version>1.2.0-mapr-1605</Hive.version>
<Sqoop.version>1.4.6-mapr-1601</Sqoop.version>
<Mapr.version>5.1.0-mapr</Mapr.version>
</properties>
```

5. From the directory containing the Hadoop distribution you want, execute this command:

   ```
   mvn dependency:copy-dependencies -DoutputDirectory=<directory to which you want to download the jars>
   ```

   For example, running the following command from the CDH directory:

   ```
   mvn dependency:copy-dependencies -DoutputDirectory=C:\CDHlib
   ```

   would insert the Cloudera Hadoop client libraries to the “C:\CDHlib” directory.

Configuring the Hive Adapter

The service.props file contains optional parameters that can also be set to control things like logging and connections.

To configure the Hive adapter:

1. Stop the Master.
2. In the {207463B0-179B-41A7-AD82-725A0497BF42} directory, create a Config subdirectory.
3. If necessary, create the service.props file in the Config directory (see Configuring service.props).
4. (Optional) Modify the properties in service.props as desired to control polling, output, and log gathering. See Configuring service.props.
5. Restart the Master.

Licensing an Adapter

Each CWA Adapter must be separately licensed. You cannot use an Adapter until you apply the license file. If you purchase the Adapter after the original installation of CWA, you will receive a new license file authorizing the use of the Adapter.

You might have a Demo license which is good for 30 days, or you might have a Permanent license. The procedures to install these license files are described below.

To license an Adapter:

1. Stop the master:

   Windows:

   a. Click on **Start** and select **All Programs>Cisco Workload Automation>Scheduler>Service Control Manager**.

   b. Verify that the master is displayed in the **Service** list and click on the **Stop** button to stop the master.

   UNIX:

   Enter **tesm stop**

2. Create the license file:

   - For a Permanent license, rename your Permanent license file to *master.lic*.
   - For a Demo license, create a file called *demo.lic*, then type the demo code into the *demo.lic* file.

3. Place the file in the **C:\Program Files\TIDAL\Scheduler\Master\config** directory.

4. Restart the master:

   Windows:

   Click **Start** in the Service Control Manager.

   UNIX:

   Enter **tesm start**

   The master will read and apply the license when it starts.

5. To validate that the license was applied, select **Registered License** from **Activities** main menu.

Securing the Adapter

There are two types of users associated with the Hive Adapter; **Runtime Users** and **Schedulers**. Although all connections to the Hive Server are anonymous, Cisco Workload Automation's job model requires at least one Hive user be defined. You maintain definitions for both types of users from the **Users** pane.

- **Runtime Users**

  Hive Server connections are anonymous, but Cisco Workload Automation's job model requires at least one Hive runtime user. Therefore when defining the Hive runtime user, the password can be of any value as it is being used at runtime.

- **Schedulers**

  Schedulers are those users who will define and/or manage Hive jobs. There are three aspects of a user profile that grant and/or limit access to scheduling jobs that affect Hive:

  - Security policy that grants or denies add, edit, delete and view capabilities for Hive jobs and events.
  - Authorized runtime user list that grants or denies access to specific authentication accounts for use with Hive jobs.
  - Authorized agent list that grants or denies access to specific Hive Adapter connections for use when defining Hive jobs.

Defining Runtime Users

To define a runtime user:

1. From the **Navigator** pane, expand the **Administration** node and select **Runtime Users** to display the defined users.

2. Right-click **Runtime Users** and select **Add Runtime User** from the context menu (*Insert* mode).

-or-

Click the **Add** button on the menu bar.

The **User Definition** dialog displays.

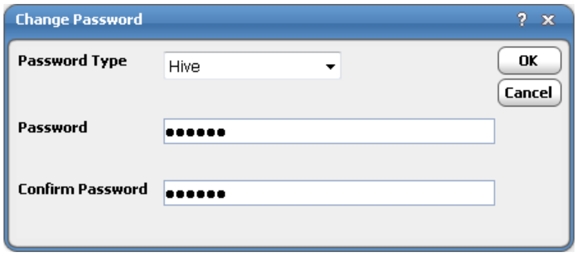
3. Enter the new user name in the **User Name** field.

4. For documentation, enter the **Full Name** or description associated with this user.

5. In the **Domain** field, select a Windows domain associated with the user account required for authentication, if necessary.

6. To define this user as a runtime user for Hive jobs, click **Add** on the **Passwords** tab.

The **Change Password** dialog displays.



Password Type: Hive

Password: ••••••

Confirm Password: ••••••

7. Select **Hive** from the **Password Type** list.

8. Enter a password (along with confirmation) in the **Password/Confirm Password** fields.

Since the password entered here is only used to satisfied the Cisco Workload Automation's job model but is not used to authenticate user to Hive Server. The password can be of any value in this case.

9. Click **OK** to return to the **User Definition** dialog.

The new password record displays on the **Passwords** tab.

10. Click **OK** to add or save the user record in the Cisco Workload Automation database.

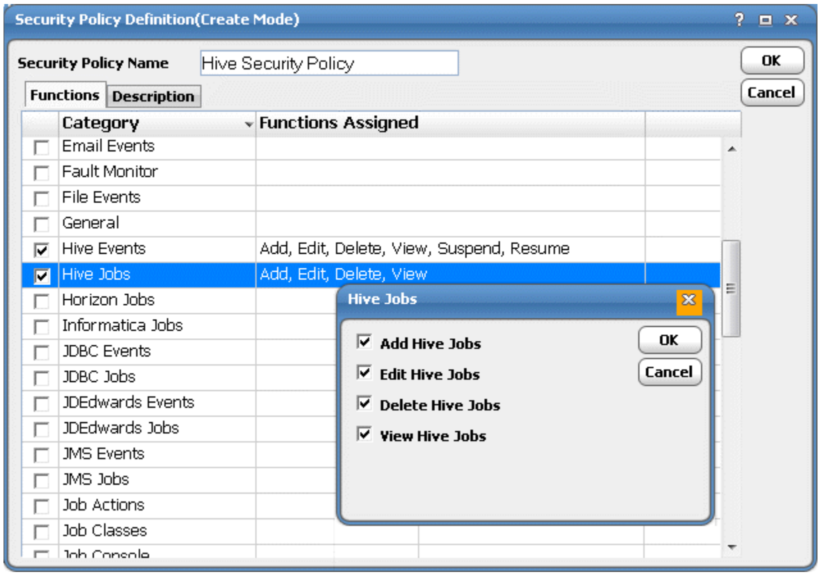
For further information about the **User Definition** dialog, see your *Cisco Workload Automation User Guide*.

Authorizing Schedulers to Work With Hive Jobs

To authorize Schedulers:

1. From the **Navigator** pane, select **Administration>Security Policies** to display the **Security Policies** pane.

2. Right-click **Security Policies** and select **Add Security Policy** from the context menu. You can also right-click to select an existing security policy in the **Security Policies** pane and select **Edit Security Policy**.



Security Policy Name: Hive Security Policy

	Category	Functions Assigned
□	Email Events	
□	Fault Monitor	
□	File Events	
□	General	
☑	Hive Events	Add, Edit, Delete, View, Suspend, Resume
☑	Hive Jobs	Add, Edit, Delete, View
□	Horizon Jobs	
□	Informatica Jobs	
□	JDBC Events	
□	JDBC Jobs	
□	JDEdwards Events	
□	JDEdwards Jobs	
□	JMS Events	
□	JMS Jobs	
□	Job Actions	
□	Job Classes	
□	Job Console	

☑ Add Hive Jobs
☑ Edit Hive Jobs
☑ Delete Hive Jobs
☑ View Hive Jobs

3. In the **Security Policy Name** field, enter a name for the policy.

4. On the **Functions** page, scroll to the **Hive Jobs** category, click the ellipses on the right-hand side of the dialog and select the check boxes next to the functions that are to be authorized under this policy (**Add**, **Edit**, **Delete** and **View Hives Jobs**).

5. Click **Close** on the **Function** dropdown list.

6. Click **OK** to save the policy.

For further information about setting up security policies, see your *Cisco Workload Automation User Guide*.

Defining Scheduler Users of Hive Jobs

To define a Scheduler user to work with Hive jobs:

1. From the **Navigator** pane, expand the **Administrative** node and select **Interactive Users** to display the defined users.

2. Right-click **Interactive Users** and select **Add Interactive User** from the context menu (*Insert* mode). You can also right-click a user in the **Interactive Users** pane and select **Edit Interactive User** from the shortcut menu (*Edit* mode).

The **User Definition** dialog displays.

3. If this is a new user definition, enter the new user name in the **User/Group Name** field.

4. For documentation, enter the **Full Name** or description associated with this user.

5. In the **Domain** field, select a Windows domain associated with the user account required for authentication, if necessary.

6. On the **Security** page, select the **Other** option and then select the security policy that includes authorization for Hive jobs.

7. Click the **Runtime Users** tab.

8. Select the Hive users that this scheduling user can use for Hive authentication from Hive jobs.
9. Click the **Agents** tab.
10. Select the check boxes for the Hive connections that this scheduling user can access when scheduling jobs.
11. Click the **Kerberos** page. If your Hadoop cluster is Kerberos secured, the Kerberos Principal and Kerberos Key page file is required. The Key page file is relative to the Master's file system and contains one or more Kerberos principals with their defined access to Hadoop.
12. Click **OK** to save the user definition.

Defining a Connection

You must create one or more Hive connections before Cisco Workload Automation can run your Hive jobs. These connections also must be licensed before Cisco Workload Automation can use them. A connection is created using the **Connection Definition** dialog.

Adding a Connection

To add a connection:

1. From the **Navigator** pane, navigate to **Administration>Connections** to display the **Connections** pane.
2. Right-click **Connections** and select **Add Connection>Hive Adapter** from the context menu.

The **Hive Adapter Connection Definition** dialog displays.

3. On the **General** page, enter a name for the new connection in the **Name** field.
4. In the **Job Limit** field, select the maximum number of concurrent active processes that Cisco Workload Automation should submit to the Hive server at one time.
5. (Optional) From the **Default Runtime User** dropdown list, you can select the name of a default user for Hive jobs.

   Only authorized users that have been defined with Hive passwords display in this list. The selected user is automatically supplied as the default runtime user in a new Cisco Workload Automation Hive job definition.
6. (Optional) Select the **Use as default for Hive Jobs** option if you wish for this connection to be used as the default whenever creating Hive jobs.

7. Click the **Hive Connection** tab.

8. In the **Distribution Type** field, choose the type of distribution you are using.

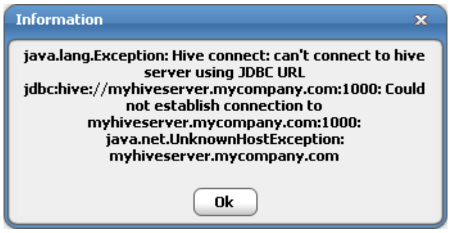
9. In the **Client Jar Directory** field, enter the path to the directory that contains the Hadoop client libraries. See Installing the Hadoop Client Libraries, page 7 to obtain these client libraries.

   Note: Hiveserver 2 supports authentication. For any authentication mechanism that is based on a username-password combination, append the following to your Hive Server JCBC URL: " ;user=<username>;password=<password>"

10. In the **Hive Server JDBC URL** field, enter the URL of the Hive server in the format specified above.

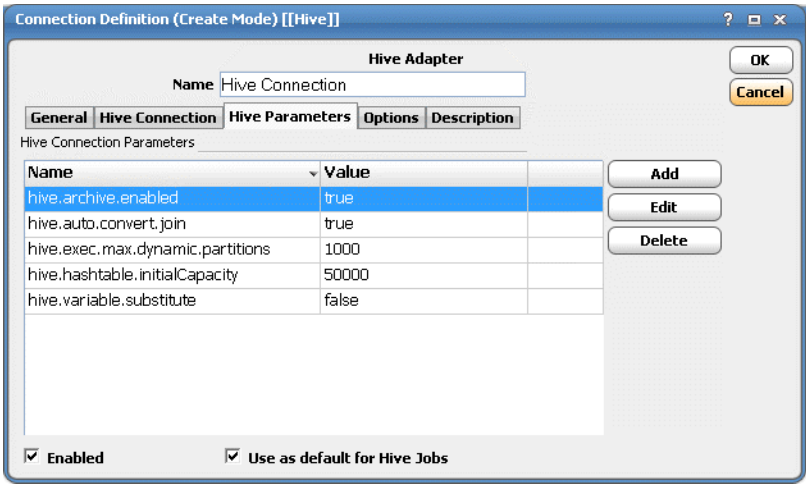
11. Click the **Test** button to test connectivity using the specified Hive Server JDBC URL.

If the test is unsuccessful, the following information dialog box displays:



java.lang.Exception: Hive connect: can't connect to hive server using JDBC URL jdbc:hive://myhiveserver.mycompany.com:1000: Could not establish connection to myhiveserver.mycompany.com:1000: java.net.UnknownHostException: myhiveserver.mycompany.com

12. Click the **Hive Parameters** tab to specify MapReduce job runtime parameters. These parameter values are applied if the HiveQL command results in a MapReduce job. Setting these parameter values at the connection level will globally set the job parameters for every job that is run on this connection. This configuration may be overridden at the job level by defining the corresponding configuration for the job options during the job definition.



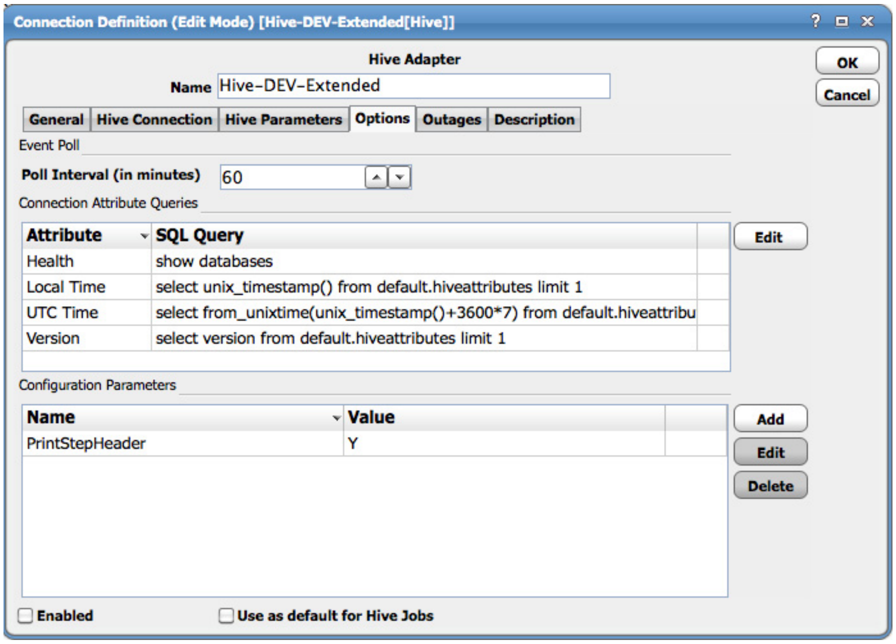
Name: Hive Connection

Hive Connection Parameters

Name	Value
hive.archive.enabled	true
hive.auto.convert.join	true
hive.exec.max.dynamic.partitions	1000
hive.hashtable.initialCapacity	50000
hive.variable.substitute	false

☑ Enabled ☑ Use as default for Hive Jobs

13. Click the **Options** tab, then in the **Event Poll** section, enter the number of minutes between successive poll checks of an event monitor that the Adapter uses to check whether it should raise an event. The default value is 60 minutes. Since MapReduce jobs are created and run each time a poll check occurs, it is recommended to keep the poll interval as high as possible.



Name: Hive-DEV-Extended

Event Poll

Poll Interval (in minutes): 60

Connection Attribute Queries

Attribute	SQL Query
Health	show databases
Local Time	select unix_timestamp() from default.hiveattributes limit 1
UTC Time	select from_unixtime(unix_timestamp()+3600*7) from default.hiveattribu
Version	select version from default.hiveattributes limit 1

Configuration Parameters

Name	Value
PrintStepHeader	Y

☐ Enabled ☐ Use as default for Hive Jobs

14. In the **Connection Attribute Queries** section, define custom queries to provide additional information about the connection. Since these custom queries result in frequent running of MapReduce jobs at the Hadoop cluster in order to provide output, it is recommended these settings be skipped unless the information is vital to the running of jobs.

- **Version** – Queries the version number of the database.
- **Local Time** – Queries the time of the database's server.

- **UTC Time** – Queries the UTC time for the database server. The local time and the UTC are used to get the time zone and calculate the time difference between the database server and the scheduling master.
- **Health** – Queries the health of the database based on any custom SQL select statement.

If the query returns a result set, the connection is considered healthy (green). If an empty result set is returned, the connection is considered unavailable (red). If none is specified, the connection uses query “show databases” by default to check the health of the connection.

The following is an example Connection Attribute query:

a. Create a custom table in Hive.

```
CREATE TABLE HIVEATTRIBUTES (VERSION STRING) ROW FORMAT DELIMITED;
```

b. Prepare a datafile in **/tmp/version.txt** that includes the version.

```
0.9.0
```

c. Load the data.

```
LOAD DATA LOCAL INPATH '/tmp/version.txt' OVERWRITE INTO TABLE HIVEATTRIBUTES;
```

d. Local Time

```
select unix_timestamp() from HIVEATTRIBUTES limit 1;
```

e. Version

```
select version from hiveattributes limit 1;
```

f. UTC Time

```
select from_unixtime(unix_timestamp()) from hiveattributes limit 1;
```

15. Click **OK** to save the new Hive connection.

The status light next to the connection indicates whether the Cisco Workload Automation Master is connected to the Hive server. If the light is green, the Hive server is connected.

A red light indicates that the master cannot connect to the Hive server. Hive jobs will not be submitted without a connection to the Hive server. You can only define jobs from the Client if the connection light is green.

If the light is red, you can test the connection to determine the problem. Right-click the connection and select **Test** from the shortcut menu. A message displays on the **Test Hive Connection** dialog describing the problem. Or, go to **Operator>Logs** to look for error messages associated with this connection.

3

Using the Hive Adapter

Overview

This chapter covers these topics:

Defining Hive Jobs

This section provides instructions for defining a Hive job in Cisco Workload Automation and descriptions of the various types of tasks and options that can be included in the jobs.

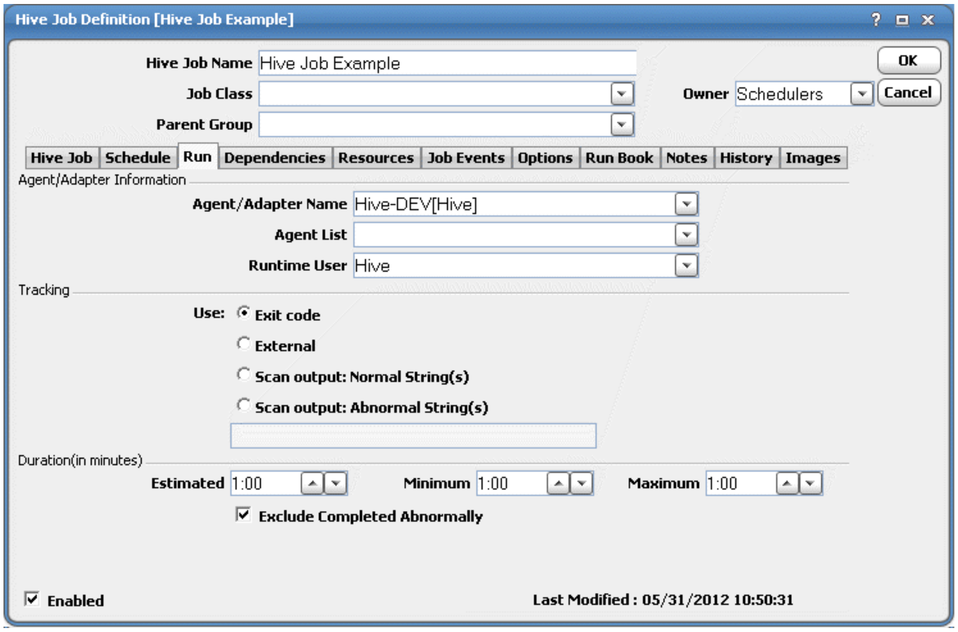
Hive Job Definition

This section describes the basic steps for defining a Hive job.

To define a Hive job:

1. In the **Navigator** pane, select **Definitions>Jobs** to display the **Jobs** pane.
2. Right-click **Jobs** and select **Add>Hive Job** from the context menu.

The **Hive Job Definition** dialog displays.



Hive Job Definition [Hive Job Example]

Hive Job Name: Hive Job Example

Job Class:

Parent Group:

Owner: Schedulers

Agent/Adapter Information

Agent/Adapter Name: Hive-DEV[Hive]

Agent List:

Runtime User: Hive

Tracking

Use: ● Exit code

○ External

○ Scan output: Normal String(s)

○ Scan output: Abnormal String(s)

Duration(in minutes)

Estimated 1:00 Minimum 1:00 Maximum 1:00

☑ Exclude Completed Abnormally

☑ Enabled

Last Modified : 05/31/2012 10:50:31

The **Run** tab is selected by default. You must first specify a name for the job, the Hive Adapter connection that will be used for the job and a valid runtime user who has the appropriate authority for the batch job being scheduled.

3. In the upper portion ofsend the dialog, specify the following information to describe the job:

   - **Job Name** – Enter a name that describes the job.
   - **Job Class** – If you want to assign a defined job class to this job, select it from the dropdown list. This field is optional.
   - **Owner** – Select the Hive job owner.
   - **Parent Group** – If this job exists under a parent group, select the name of the parent group from the dropdown list. All properties in the Agent Information section are inherited from its parent job group.

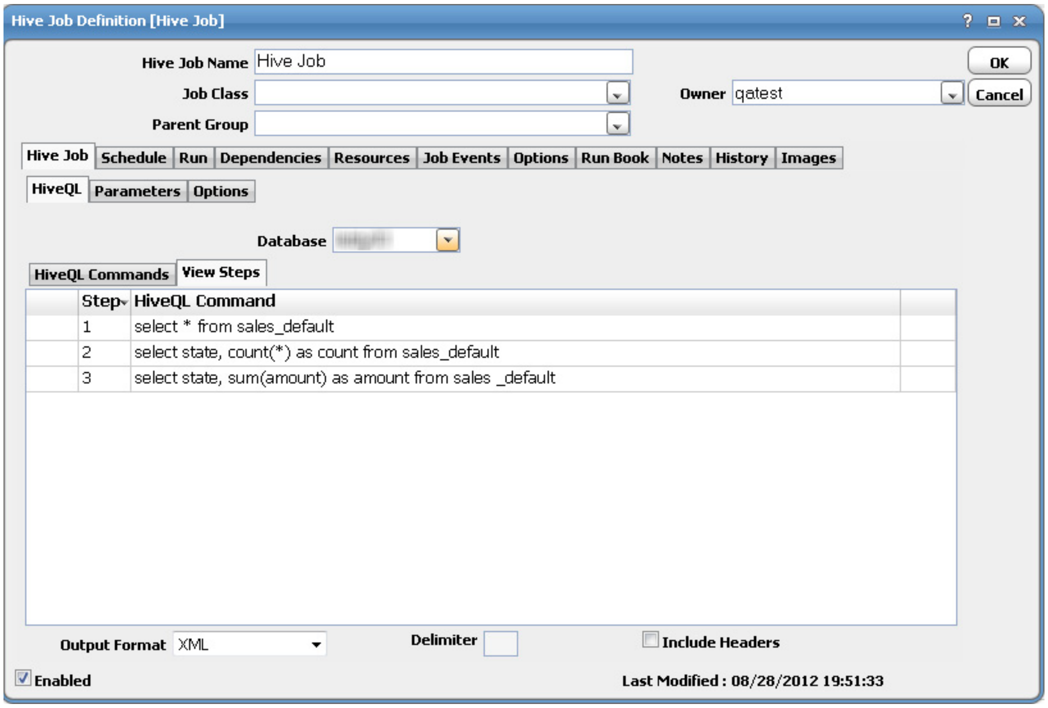
4. Specify the following connection information in the **Agent/Adapter Information** section:

   - **Agent/Adapter Name** – Select the Hive Adapter connection to be used for this job from the dropdown list.
   - **Agent List Name** – Select a list for broadcasting the job to multiple Hive servers rather than a specific agent/adapter.
   - **Runtime User** – Select a valid runtime user with the appropriate authority for the job from the dropdown list.

5. Specify the appropriate **Tracking** and **Duration** information for the job.

   Refer to the *Cisco Workload Automation User Guide* for information on these options.

6. Click the **Hive Job** tab.



Hive Job Name: Hive Job

Owner: qatest

Step	HiveQL Command
1	select * from sales_default
2	select state, count(*) as count from sales_default
3	select state, sum(amount) as amount from sales _default

Output Format: XML

□ Include Headers

☑ Enabled

Last Modified : 08/28/2012 19:51:33

7. On the **HiveQL** tab, select a database from the **Database** drop-down list. By default, all commands are run on this “default” database.

8. On the **HiveQL Commands** subtab, define one or more HiveQL commands.

   If more than one command is defined (as in the figure above), separate them by a semicolon “;”. If parameters are necessary, precede them with a colon “:”.

   If more than one databases specification is needed by the job definition, specify the additional databases using the <database> command to switch.

   For example:

```
select * from sales_default;
use db01;
select * from sales_db01;
```

   In the example above, the database is “Default” and the first statement is run against the " Default" database. The second command switches the current database to “db01”. All commands executed from this point forward are against “db01”.

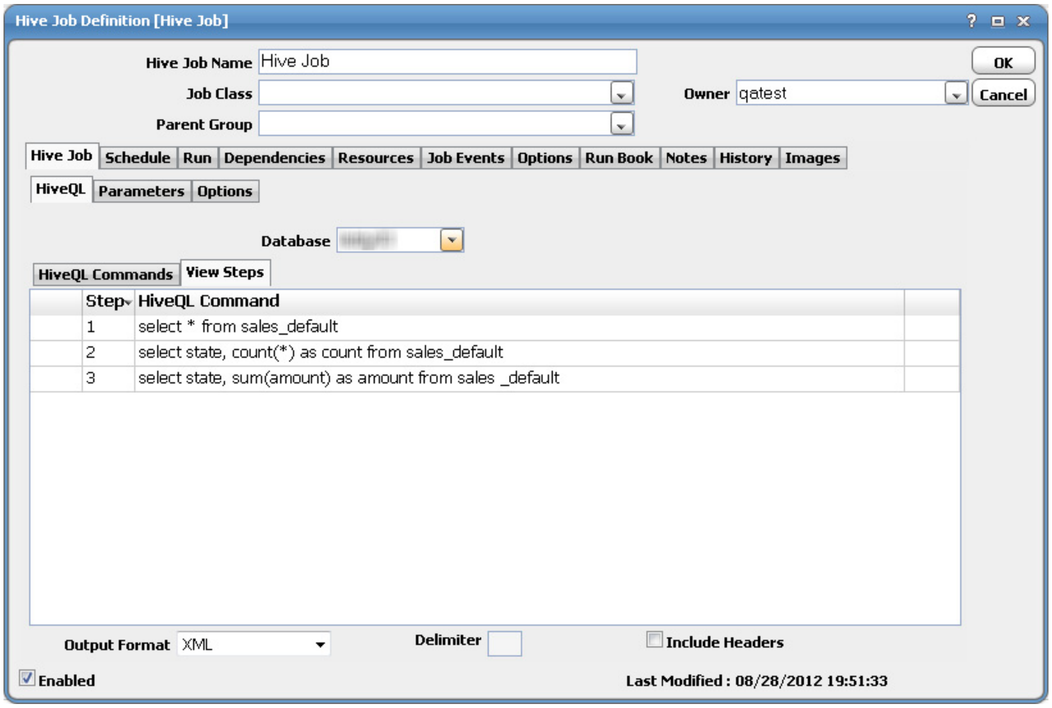
9. From the **Output Format** list, select one of the following query results formats.

   - **XML** - Writes the query results in XML format.
   - **Align Columns** - Displays the values in the most readable format.
   - **CSV Format** - Separates values with commas.
   - **Raw** - Separates values with a user-defined character.

10. In the **Delimiter** field, specify the custom character to use for delimiting the column data from the query results if **Raw** is selected for **s**.

11. (Optional) Select the **Include Headers** option to write out the column headers of the results.

    You can view the steps on the **View Steps** subtab.



Hive Job Definition [Hive Job]

Hive Job Name: Hive Job

Job Class:

Parent Group:

Owner: qatest

Database: [redacted]

Step	HiveQL Command
1	select * from sales_default
2	select state, count(*) as count from sales_default
3	select state, sum(amount) as amount from sales _default

Output Format: XML

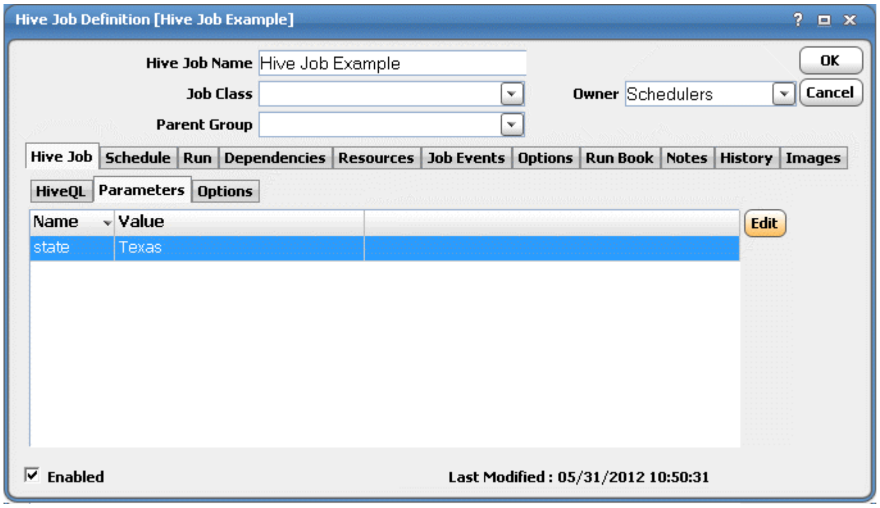
Delimiter:

□ Include Headers

☑ Enabled

Last Modified : 08/28/2012 19:51:33

12. Click the **Parameters** subtab to use parameters to be replaced at runtime.



Hive Job Definition [Hive Job Example]

Hive Job Name: Hive Job Example

Job Class:

Parent Group:

Owner: Schedulers

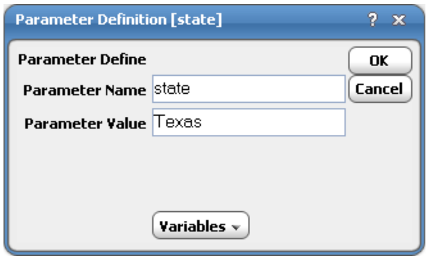
Name	Value
state	Texas

☑ Enabled

Last Modified : 05/31/2012 10:50:31

This subtab provides a list of parameters that have been preceded by a colon in the HiveQL statements. If there are no parameters defined in the SQL statement, you cannot enter parameters on this subtab.

To edit a parameter, highlight it, then click **Edit** to display the **Parameters** dialog. Edit the parameter, then click **OK**.



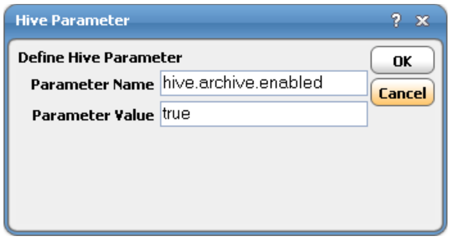
Parameter Define

Parameter Name: state

Parameter Value: Texas

You cannot modify the **Parameter Name**. The **Parameter Value** can be a literal value or a valid Cisco Workload Automation variable.

13. (Optional) On the **Options** subtab, specify additional MapReduce job runtime parameters. These parameter/values are applied if the HiveQL command results in a MapReduce job.



Define Hive Parameter

Parameter Name: hive.archive.enabled

Parameter Value: true

The parameter name/value specified at the job level overrides the value from the connection level if the same parameter name/value is specified.

14. Click **OK** to save the job.

Defining Hive Events

Using the Hive Adapter, you can define events that can be used for alerting and invoking an automated response through email and/or inserting additional jobs into the schedule. The **Event Definition** dialog is displayed when you add or edit a Hive event. Cisco Workload Automation can monitor events and then take one or more actions when the event trigger occurs. You must configure a calendar for the event from the **Schedule** tab to schedule when the event is enabled (that is, when monitoring will occur). If needed, you can configure the monitor to operate only during certain time periods or leave the monitor in operation at all times.

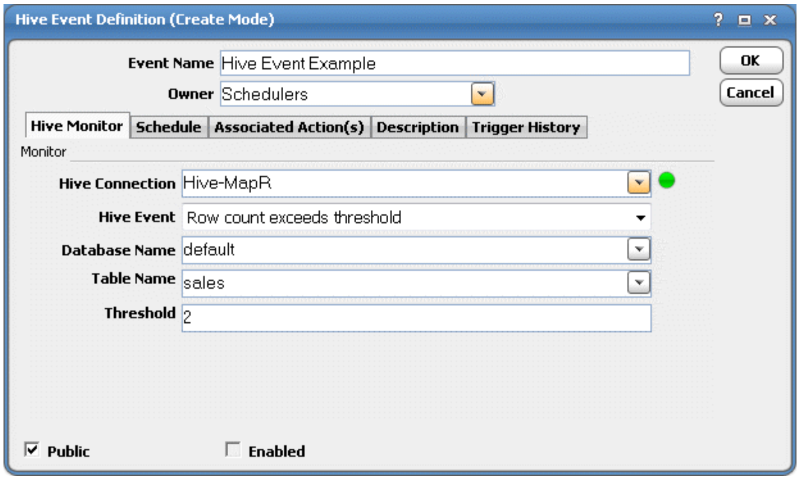
Hive Event Definition

To define a Hive event:

1. In the **Navigator** pane, select **Definitions>Events>Hive Events** to display the **Hive Events** pane.

2. Right-click **Hive Events** and select **Add Hive Event** from the context menus.

The **Hive Event Definition** dialog displays.



Event Name: Hive Event Example

Owner: Schedulers

Hive Connection: Hive-MapR ●

Hive Event: Row count exceeds threshold

Database Name: default

Table Name: sales

Threshold: 2

☑ Public □ Enabled

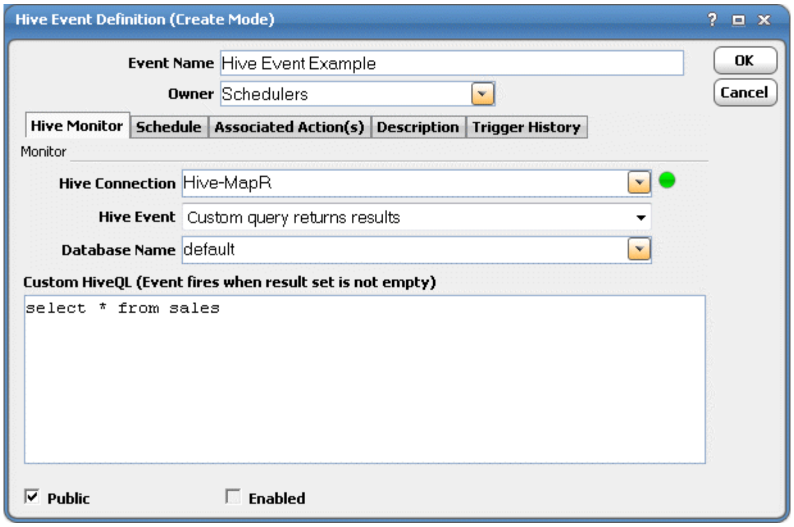
3. Enter a name for the event in the **Event Name** field and select an **Owner** from the dropdown list.

4. Click the **Hive Monitor** tab to designate the data able in Hive to be monitored and define the condition that will be considered an event. Once the designated change to the datatable is detected, the event can be linked to an action to trigger an automatic response to the change.

   Specify the following information:

- **Hive Connection** – Select the Hive connection from the list. This is the connection that will be monitored for the specified event. The status light next to the list indicates the connection status to the Hive server. If the light is green, the Hive server is connected.
- **Hive Event** – Select one of the following Hive events from the list:
  - **Row count exceed threshold** – When the threshold is exceeded and the event is triggered, the event triggers again if the row count drops below the threshold and then goes above the threshold again
  - **Column MAX value exceeds threshold** – If the largest monitored column value rises above the threshold, the event will trigger. The event will trigger again if the column maximum value drops below the threshold and rises above it again.
  - **Column MIN value below threshold** – If the smallest monitored column value falls below the threshold, the event will trigger. The event will trigger again if the column minimum value goes above the threshold and drops below it again.
  - **Custom query returns results** – A custom event triggers when a non-empty result set is returned from a custom query.
- **Database Name** – Select a database for the connection.
- **Table Name** – Select the name of the data table to be monitored.
- **Column Name** – This field displays if the event you have selected from the **Hive Event** list is either **Column MAX** or **Column MIN**. Select the name of the column for the selected table.
- **Threshold** – Select a threshold for the selected event.

- **Custom HiveQL** – This field displays if the event you have selected from the **Hive Event** list is **Custom query returns results**. Enter the HiveQL command.



Event Name Hive Event Example

Owner Schedulers

Hive Connection Hive-MapR

Hive Event Custom query returns results

Database Name default

Custom HiveQL (Event fires when result set is not empty)

```
select * from sales
```

☑ Public ☐ Enabled

Note: The other tabs on the Hive Event Definition dialog are general event configuration options and are not specific to the Hive Adapter. Any action that is available in Cisco Workload Automation, such as sending email, generating alerts, sending SNMP traps, setting variables, and adding jobs is available as a response to a Hive event.

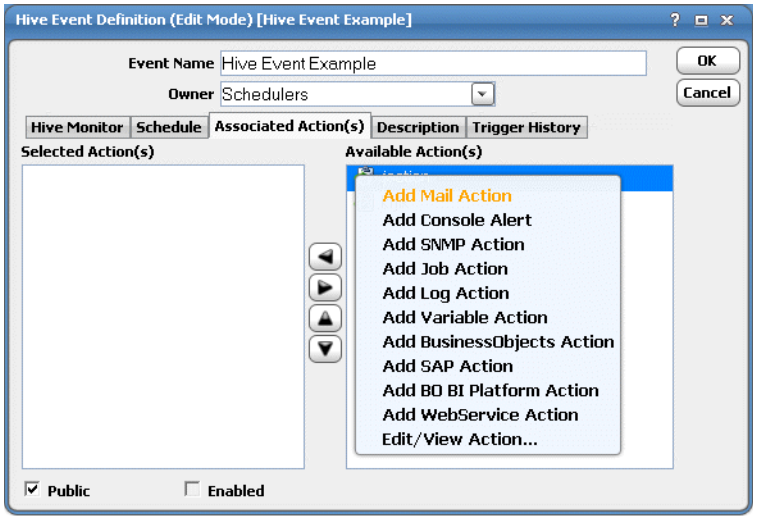
5. Click **OK** to save the event definition.

Define an Action for an Event

You can add any action for a Hive event that is available in Cisco Workload Automation.

For example, to define an email action for a Hive event:

1. From the **Hive Event Definition** dialog, click the **Associated Actions** tab.
2. In the **Available Actions** section, right-click and select **Add Mail Action**.

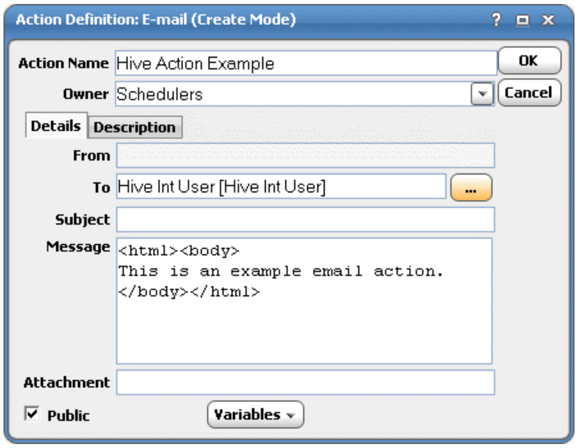


Event Name Hive Event Example

Owner Schedulers

☑ Public ☐ Enabled

The **Action Definition: E-Mail** dialog displays.



Action Name: Hive Action Example
Owner: Schedulers
From:
To: Hive Int User [Hive Int User]
Subject:
Message:

```
<html><body>
This is an example email action.
</body></html>
```

Attachment:
☑ Public

3. Complete the required fields on the **Action Definition** dialog and click **OK**.

Monitoring Hive Job Activity

As Hive tasks run as pre-scheduled or event-based jobs, you can monitor the jobs as you would any other type of job in Cisco Workload Automation using the **Job Details** dialog. You can also use Business Views to monitor job activity and view when the jobs are active (see your *Cisco Workload Automation User Guide* for instructions on using Business Views).

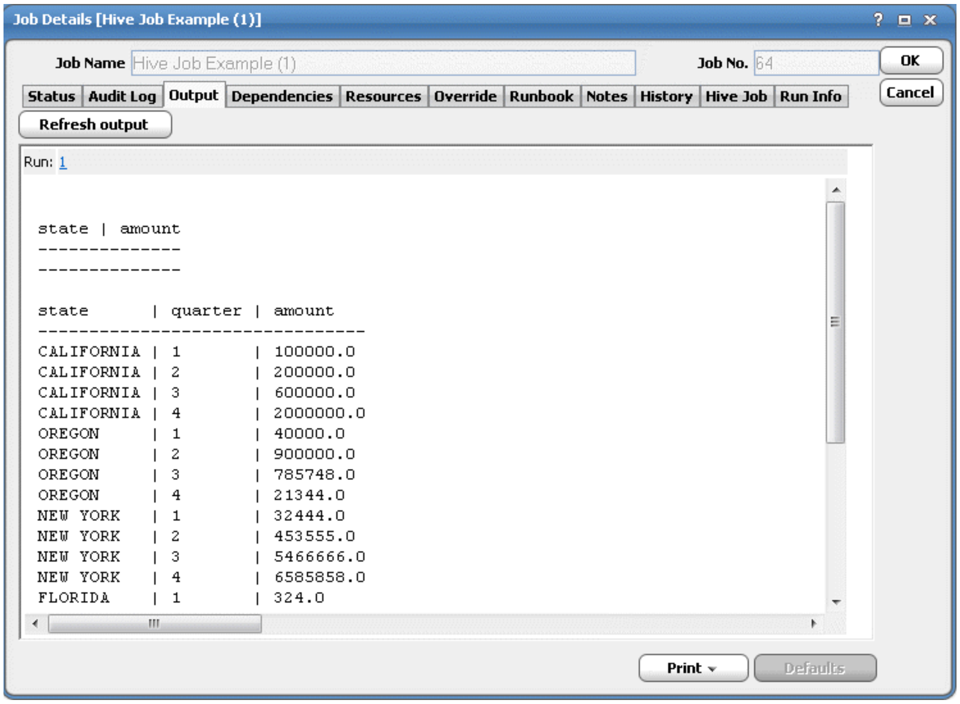
To monitor job activity:

1. In the **Navigator** pane, select **Operations>Job Activity** to display the **Job Activity** pane.

2. Right-click to select a job and choose **Details** from the context menu.

   The **Job Details** dialog displays. On the **Status** page, you can view the status of the job, the start and end time, how long it ran, and how it was scheduled. The external ID is the run ID associated with the specific execution of the batch job.

3. Click the **Output** tab to view the result sets from the HiveQL commands after the job completes.



Job Name: Hive Job Example (1)

Job No. 64

Run: 1

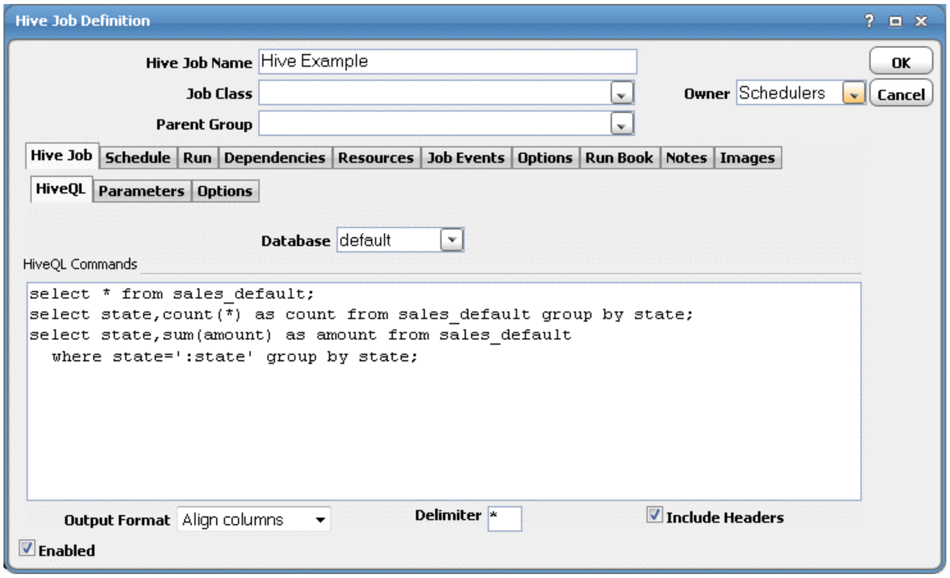
```
state | amount
--------------
--------------

state      | quarter | amount
--------------------------------
CALIFORNIA | 1       | 100000.0
CALIFORNIA | 2       | 200000.0
CALIFORNIA | 3       | 600000.0
CALIFORNIA | 4       | 2000000.0
OREGON     | 1       | 40000.0
OREGON     | 2       | 900000.0
OREGON     | 3       | 785748.0
OREGON     | 4       | 21344.0
NEW YORK   | 1       | 32444.0
NEW YORK   | 2       | 453555.0
NEW YORK   | 3       | 5466666.0
NEW YORK   | 4       | 6585858.0
FLORIDA    | 1       | 324.0
```

Cisco Workload Automation can be configured to save or discard job output by default from the **Defaults** tab of the **System Configuration** dialog. Regardless of the system default, any individual job instance can be configured from its job definition to override the system default. Each time a job is rerun, that run's output is separated by a block of number signs.

A Step Header with step number, status, and an abbreviated version of the step command is displayed before the actual result set. A Step Header can be turned off using the “PrintStepHeader” option in the connection. By default, PrintStepHeader is turned on.

If there are multiple commands in the Hive Job, each command is run as a separate step and all steps are displayed in the **Output** tab sequentially.

4. Click the **Hive Job** tab to view the original request along with variables used when this job was submitted. This tab allows you to override the HiveQL commands, output format, and parameter values prior to run or rerun. Overrides are not permitted when the job is running.



Hive Job Name: Hive Example

Job Class:

Parent Group:

Owner: Schedulers

Database: default

HiveQL Commands

```
select * from sales_default;
select state,count(*) as count from sales_default group by state;
select state,sum(amount) as amount from sales_default
  where state=':state' group by state;
```

Output Format: Align columns

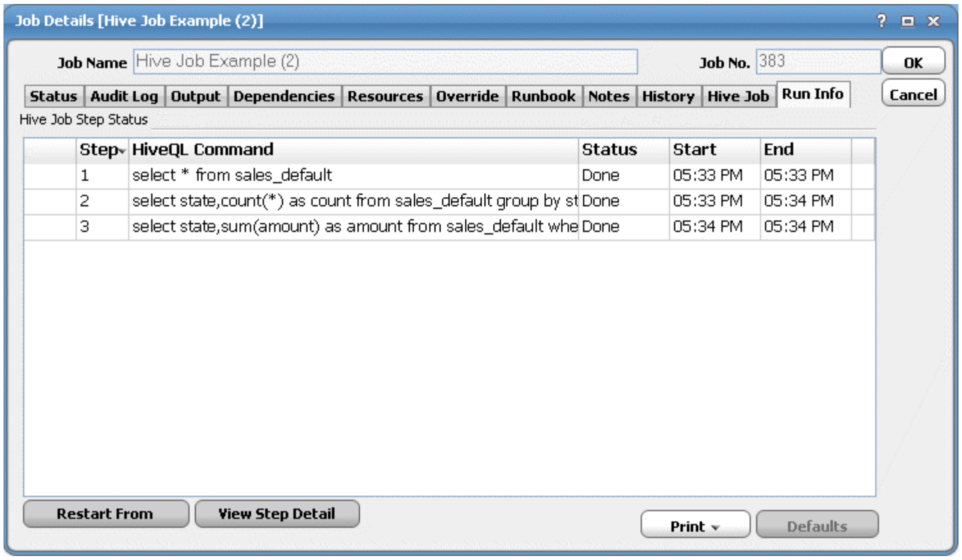
Delimiter: *

☑ Include Headers

☑ Enabled

While the job is running, the fields are disabled; however, prior to running or rerunning the job, you can override any value on this screen. Your changes here only apply to this instance of the job (the original job definition is not affected).

5. Click the **Run Info** tab to view the run status, start, and end time for each step in the Hive Job. This tab is read-only.



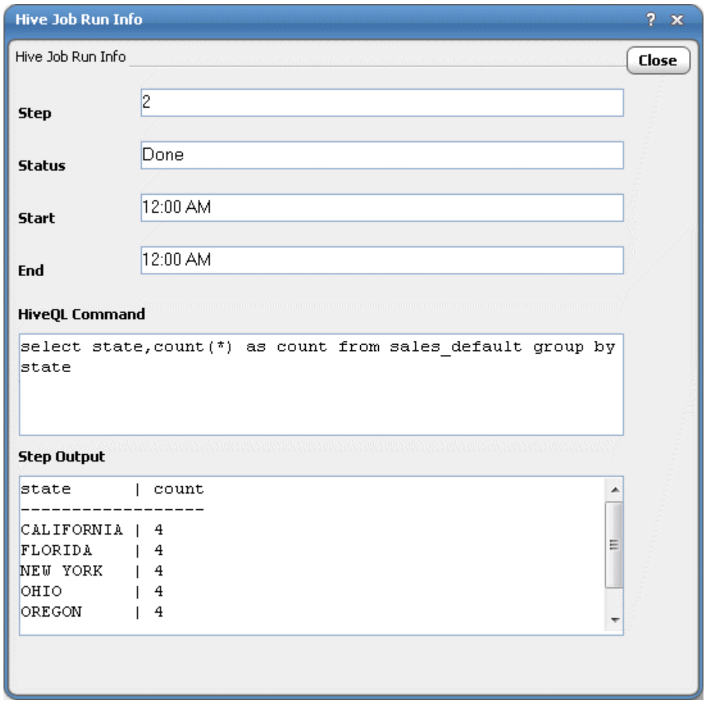
Job Name: Hive Job Example (2)

Job No.: 383

Hive Job Step Status

Step	HiveQL Command	Status	Start	End
1	select * from sales_default	Done	05:33 PM	05:33 PM
2	select state,count(*) as count from sales_default group by st	Done	05:33 PM	05:34 PM
3	select state,sum(amount) as amount from sales_default whe	Done	05:34 PM	05:34 PM

Click **View Details** to view the **Hive Job Run Info** dialog.



Step: 2

Status: Done

Start: 12:00 AM

End: 12:00 AM

HiveQL Command

```
select state,count(*) as count from sales_default group by
state
```

Step Output

```
state      | count
-------------------
CALIFORNIA | 4
FLORIDA    | 4
NEW YORK   | 4
OHIO       | 4
OREGON     | 4
```

This dialog contains the run status, start, and end time for each step in the Hive job, the HiveQL command and its output. You can override the HiveQL command and output on this dialog.

Additionally, you can select a step and click Restart From to rerun a job from a particular step.

6. When you have finished viewing the job activity details, click **OK** to close the dialog.

Controlling Adapter and Agent Jobs

Scheduler provides the following job control capabilities for either the process currently running or the job as a whole:

- Holding a Job—Hold a job waiting to run.
- Aborting a Job—Abort an active job.
- Rerunning a Job—Rerun a job that completed.
- Making One Time Changes to an Adapter or Agent Job Instance—Make last minute changes to a job.
- Deleting a Job Instance before It Has Run—Delete a job instance before it has run.

Holding a Job

Adapter/agent jobs are held in the same way as any other Scheduler jobs.

Adapter/agent jobs can only be held before they are launched. Once a job reaches the Adapter/Agent system, it cannot be held or suspended.

To hold a job:

1. From the **Job Activity** pane, right-click on the job.

2. Select **Job Control>Hold/Stop**.

Aborting a Job

Adapter/agent jobs are aborted in the same way as any other Scheduler jobs.

To abort a job:

1. From the **Job Activity** pane, right-click on the job.
2. Select **Job Control>Cancel/Abort**.

Rerunning a Job

On occasion, you may need to rerun an Adapter/Agent job. You can override parameter values first, if necessary, from the Adapter/Agent tab.

To rerun a job:

1. From the **Job Activity** pane, right-click the Adapter/Agent job you need to rerun.
2. Select **Job Control>Rerun** option from the context menu.

Making One Time Changes to an Adapter or Agent Job Instance

Prior to a run or rerun, you can edit data on the specific **Adapter/Agent** tab. To ensure that there is an opportunity to edit the job prior to its run, you can set the **Require operator release** option on the **Options** tab in the Adapter **Job Definition** dialog. Use this function to make changes to an Adapter job after it enters Waiting on Operator status as described in the following procedure.

To make last minute changes:

1. From the **Job Activity** pane, double-click the Adapter/Agent job to display the **Job Details** dialog.
2. Click the Adapter tab.
3. Make the desired changes to the job and click **OK** to close the **Job Details** dialog.
4. If this job is Waiting on Operator, perform one of the following tasks:

- To release the job, select **Job Control->Release**.
- To rerun the job with changes, select **Job Control->Rerun**.

Deleting a Job Instance before It Has Run

Adapter/Agent job instances are deleted in the same way as any other Scheduler job.

Deleting a job from the **Job Activity** pane removes the job from the Scheduler job activity only. The original definition is left in tact.

To delete a job instance:

1. From the **Job Activity** pane, right-click the Adapter/Agent job to be deleted.
2. Select **Remove Job(s) From Schedule**.

4

Setting Up SSL Connection

This chapter provides the instructions to setup Hive connection using SSL.

Procedure to setup SSL connection

To use Hive with SSL in the connection Definition, set ssl=true in the jdbc URL.

For example:

```
jdbc:hive://dummyserver:10000/;ssl=true;
```

Note: It is assumed that a JRE or JDK is available in your system PATH.

1. Launch the Command Prompt window.
2. Change to the directory in which the certificates are stored, by entering the following commands

   ```
   C:
   cd /hive-certs
   ```

3. To import a certificate using the Java keytool utility, enter the following syntax:

   ```
   keytool -genkey -alias  <alias-name> keyalg <algorithm-name> -keysize <key-size>  -keystore
   <keystore-filename>
   ```

   For example:

   ```
   C:\ hive -certs>keytool -genkey -alias hive -keyalg RSA -keysize 1024 -keystore hive.jks
   ```

4. When prompted to create a password for the keystore, enter the password at the prompt. The keystore utility displays the certificate information.
5. On the **Trust this certificate?** prompt, type **Yes** and press **Enter**. The certificate is imported into the hive.jks and a message “**Certificate was added to keystore**” is displayed.
6. Repeat this procedure for each target server.
7. Navigate to the following folder where the Enterprise Scheduler Hive Adapter is installed, and create a directory named **config**:

   ```
   <install-dir>/master/services/ {207463B0-179B-41A7-AD82-725A0497BF42} /config
   ```

8. Create a text file named service.props.
9. Open the service.props text file, and add the following line:

   ```
   TrustStore=C:\\ hive -certs\\hive.jks
   ```

Note: The double backslash is used for Windows directories.

10. Restart the Master, to establish the Hive adapter SSL connection.

5

Configuring service.props

About Configuring service.props

The **service.props** file is used to configure adapter behavior. **service.props** is located in the \config directory located under the Adapter's GUID directory, You can create both the directory and file if it does not yet exist. Properties that can be specified in service.props control things like logging and connection configuration. Many of the properties are specific to certain adapters; others are common across all adapters.

service.props Properties

The table below lists many of the parameters that can be specified in service.props. Some properties apply to all adapters (shaded in the table) and some properties are adapter-specific as indicated by the **Applicable Adapter(s)** column. The properties are listed in alphabetical order.

Property	Applicable Adapter(s)	Default	What It Controls
BYPASS_SEC_VALIDATION	Oracle Apps	N	If set to Y, the secondary user validation is bypassed. If not, secondary user validation is performed.
CLASSPATH	All	<none>	(Optional) - The path to the JDBC driver. If the default CLASSPATH used when the Adapter process is started does not include an appropriate JDBC driver jar required to connect to the PowerCenter Repository Database, you will need to specify this *service.props* configuration
CONN_SYNC	Informatica, Oracle Apps, SAP	N	Setting this flag to Y allows synchronous connections without overloading the RDOnly Thread. If set to N, the adapter might stop trying to reconnect after an outage or downtime.
DISCONN_ON_LOSTCONN	Informatica	N	Setting this flag to Y avoids an unnecessary logout call to the Informatica server when the connection is lost. This logout call usually hangs.
EnableDynamicPollingInterval	All	N	Use to avoid frequent polling on long-running jobs. When set to Y in service.props of a particular adapter, these properties are enabled: MinDynamicPollInterval—Minimum value should be 5 seconds. MaxDynamicPollIntervalInMin—Maximum value should be 5 minutes. PercentOfEstDuration—Default value is 5.

service.props Properties

Property	Applicable Adapter(s)	Default	What It Controls
HADOOP_JAVA_HOME	Sqoop	<none>	If the Java version used in the Hadoop environment is lower than Java 8, then install the same lower JDK version in the in the Master and include the path of the JDK in this property.
IGNORE_CODES	Informatica	<none>	This parameter can be set in service.props, job configuration and connection configuration parameters. The order of precedence is service.props (applicable for all jobs running in all connections), job level (only for that particular job), and connection (applicable for all jobs in the connection). This parameter is used to specify Informatica-specific error codes, separated by commas (,), that you want to ignore while running a job.
IGNORESUBREQ	Oracle Apps	N	Y or N. Setting this flag to Y stops huge job xml file transfers back and forth between the adapter and the AdapterHost during polls when a single request set has multiple sub-requests of more than 100. The default value is N or empty.
kerbkdc	MapReduce	<none>	If the Hadoop cluster is Kerberos secured, use this value to specify the KDC Server. For example, **`kerbkdc=172.25.6.112`**
kerbrealm	MapReduce	<none>	If the Hadoop cluster is Kerberos secured, use this value to specify the Kerberos Realm. For example, **`kerbrealm=TIDALSOFT.LOCAL`**
Keystore	BusinessObjects, BusinessObjects BI, BusinessObjects DS, Cognos, JD Edwards, Oracle Applications, UCS Manager, VMware, Web Service	<none>	Specify Keystore=c:\\<adapter_certificate_directory>\\<your_trusted_keystore>.keystore when importing certificates into a Java keystore.
LAUNCH_DELAY (in milliseconds)	Informatica	<none>	This parameter can be set in service.props, job configuration and connection configuration parameters. The order of precedence is service.props (applicable for all jobs running in all connections), job level (only for that particular job), and connection (applicable for all jobs in the connection). If a non-zero value is set for this parameter, then the jobs are delayed for the specified number of milliseconds before being submitted to Informatica.

Property	Applicable Adapter(s)	Default	What It Controls
LoginConfig	BusinessObjects BI Platform, BusinessObjects Data Services	<none>	Specifies the location of the login configuration if using WinAD or LDAP authentication. For example: LoginConfig=c:\\windows\\bscLogin.conf where "`c:\\windows\\bscLogin.conf`" is the location of the login configuration information. Note the use of \\ if this is a Windows location.
MaxLogFiles	Informatica, JDBC, PeopleSoft	50	(Optional) - Number of logs to retain.
OUTPUT_ASYNC_LOGOUT	Informatica	N	Setting this flag to Y avoids jobs getting stuck in Gathering Output status.
OUTPUT_SYNC	All	Y	Enables concurrent output gathering on a connection. To enable this feature, set the value to N.
POLL_SYNC	All	Y	Enables concurrent polling on connections of the same type. This is helpful when there is a heavily load on one connection of an adapter. The heavily loaded connection will not affect the other adapter connection. To enable this feature, set the value to N.
QUERY_TIMEOUT	Oracle Apps	N	Y or N. If set to Y, the timeout value defined using the parameter QUERY_TIMEOUT_VALUE is applied to the SQL queries. Default value is N or empty.
QUERY_TIMEOUT_VALUE	Oracle Apps	unset	The time period in seconds that SQL queries wait before timeout. If 0 or not set, there is no timeout.
READPCHAINLOG	SAP	Y	Used to control the log gathering in SAP Process Chain jobs. This property depends on the Summary Only check box of the job definition Options tab.
SCANFOR_SESSIONSTATS	Informatica	Y	Y or N - Set this parameter to N to turn off the default behavior of Informatica jobs collecting the session statistics during the job run.
SCANFOR_SESSIONSTATS_AFTER_WF_ENDS	Informatica	N	Y or N - Set this parameter to Y to turn off the gathering of session statistics during each poll for the status of Informatica jobs.
TDLINFA_LOCALE	Informatica	<none>	Points to the Load Manager Library locale directory. See "Configuring the Informatica Adapter" in the *Informatica Adapter Guide* for how to set this for Windows and Unix environments.
TDLINFA_REQUESTTIMEOUT	Informatica	<none>	(Optional) - The number of seconds before an API request times out. The default is 120 seconds, if not specified.
TDLJDBC_LIBPATH	JDBC	<none>	(Windows only, optional) An alternate path to the JDBC library files. The library file path should have been configured given system environment variables. This option is available in case you wish to use an alternate set of libraries and may be helpful for trouble-shooting purposes.

Property	Applicable Adapter(s)	Default	What It Controls
TDLJDBC_LOCALE	JDBC	<none>	The path to the JDBC locale files.
TRANSACTION_LOG_BATCH_SIZE	MS SQL	5000	Set this parameter if more than 5000 lines need to be read from the transaction table.
version_pre898	JD Edwards	N	If running on a JD Edwards server version that is less than 8.9.8, set version_pre898=Y.