Cisco Tidal Enterprise Scheduler MapReduce Adapter Guide

Version: 6.2.1

May 4, 2016

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

Cisco Tidal Enterprise Scheduler MapReduce Adapter Guide

CONTENTS

Preface

This guide describes the installation, configuration, and usage of the MapReduce Adapter with Cisco Tidal Enterprise Scheduler (TES).

Audience

This guide is for administrators who install and configure the MapReduce Adapter for use with TES, and who troubleshoot TES installation and requirements issues.

Related Documentation

See the *Cisco Tidal Enterprise Scheduler 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 TES 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 Tidal Enterprise Scheduler MapReduce Adapter Guide*.

Version Number	Issue Date	Reason for Change
6.1.0	December 2012	• New Cisco version.
6.2.1	June 2014	• Available in online Help only.
6.2.1 SP2	June 2015	• Configuration provided in the *TES Installation Guide*; usage provided in online Help only.
6.2.1 SP3	May 2016	• Consolidated all MapReduce Adapter documentation into one document.

Introducing the MapReduce Adapter

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

Overview

Hadoop MapReduce is a software framework for writing applications that process large amounts of data (multi-terabyte data-sets) in-parallel on large clusters (up to thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

A Cisco Tidal MapReduce Adapter job divides the input data-set into independent chunks that are processed by the map tasks in parallel. The framework sorts the map's outputs, which are then input to the reduce tasks. Typically, both the input and output of the job are stored in a file-system. The framework schedules tasks, monitors them, and re-executes failed tasks.

Minimally, applications specify the input/output locations and supply map and reduce functions via implementations of appropriate interfaces and/or abstract-classes. These, and other job parameters, comprise the job configuration. The Hadoop job client then submits the job (jar/executable etc.) and configuration to the JobTracker. The client then assumes the following responsibilities:

- Distributes the software/configuration to the slaves
- Schedules and monitors tasks
- Provides status and diagnostic information to the job -client

The MapReduce Adapter serves as the job client to automate the execution of MapReduce jobs as part of a Tidal Enterprise Scheduler (TES) managed process. The Adapter uses the Apache Hadoop API to submit and monitor MapReduce jobs with full scheduling capabilities and parameter support. Alternatively, the Adapter may be configured to connect to a Cloudera Hadoop or MapR distribution. As a platform independent solution, the Adapter can run on any platform where the TES master runs.

Prerequisites

- Linux is the only supported production platform for Apache Hadoop. However, the MapReduce Adapter can run on any platform supported by the TES Master. The MapReduce Adapter is supported on Apache Hadoop version 1.0.1+, Cloudera version CDH3-Update4 and MapR (versions 2.0.xand 3.0.x).

  See the *Cisco TES Compatibility Guide* for specific version support.

Note Only one type of distribution is supported by the TES master at any time. Apache Hadoop is the default selection, which can be changed in the service.props.

MapR configuration is supported only on Windows 2008 and Linux servers.

- All files needed by the MapReduce job are already in HDFS (mapper/reducer classes, dependent libraries, resource files, etc.) The Adapter does not move dependent files in or out of HDFS.
- All hosts that use the Kerberos authentication system must have their internal clocks synchronized within a specified maximum amount of time (known as clock skew). This requirement provides another Kerberos security check. If the clock skew is exceeded between any of the participating hosts, client requests are rejected. The maximum clock skew is configurable, but typically defaulted to five minutes. Refer to Kerberos documentation for further details. Because maintaining synchronized clocks between the KDCs and Kerberos clients (Master host machine) is important, you should use the Network Time Protocol (NTP) software or other similar time service tools to synchronize them.
- Cisco Tidal Enterprise Scheduler Adapters require Java 7. (Refer to *Cisco Tidal Enterprise Scheduler Compatibility Guide* for further details).

Software Requirements

The minimum requirement for the TES master and client is 6.1.

Refer to your *Cisco Tidal Enterprise Scheduler Compatibility Guide* for a complete list of hardware and software requirements.

Installing and Configuring the MapReduce Adapter

Overview

The MapReduce Adapter software is installed as part of a standard installation of TES. However, you must perform the following steps to license and configure the adapter before you can schedule and run MapReduce jobs:

- Licensing an Adapter – Apply the license to the Adapter. You cannot define a MapReduce connection until you have applied the license from Cisco.
- Configuring the Adapter – Deploy the adapter package from the TES Master.
- Securing the Adapter – Define MapReduce users that the adapter can use to establish authenticated sessions with the MapReduce server and permit requests to be made on behalf of the authenticated account.
- Defining a Connection – Define a connection so the master can communicate with the MapReduce server.

See Configuring service.props for information about general and adapter-specific properties that can be set to control things like logging and connection properties.

Licensing an Adapter

Each TES 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 TES, 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:

Step 1 Stop the master:

Windows:

a. Click **Start** and select **Programs>TIDAL Software>Scheduler>Master>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**

Step 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.

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

Step 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.

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

Configuring the Adapter

This section describes the requirements and configuration tasks:

- MapR Client Software Requirements
- Configuring the MapReduce Adapter

See also Configuring service.props for information about general and adapter-specific properties that can be set to control things like logging and connection properties.

MapR Client Software Requirements

When using MapR:

- MapR Client software must be configured on the TES master machine.
- MapR Client software must be configured appropriately using the link http://www.mapr.com/doc/display/MapR/Setting+Up+the+Client. The Adapter will not work unless there is confirmed communication between the client and cluster.
- MapR Client allows user impersonation only for Windows TES master whereas the Linux TES master must use the same username which the cluster allows.
- When using the MapR distribution, service.props must be modified for your platform. See step 10 in Configuring the Adapter. Also see Configuring service.props for more information about service.props configuration.

Configuring the MapReduce Adapter

The MapReduce Adapter adapter must be configured before you can schedule and run MapReduce jobs.

To configure the MapReduce adapter:

Step 1 Stop the Master.

Step 2 In the {D9AC03D5-41ED-4B1E-8A45-B2EC8BDE3EA0} directory, create a subdirectory named *Config* if does not already exist.

Step 3 Create the *service.props* file in the Config directory.

Step 4 (For Apache 1.1.2 distribution only) Add the following lines in the service.props file:

```
jarlib=apache1.1.2
CLASSPATH=C:\\Program
Files\\TIDAL\\Scheduler\\Master\\services\\{D9AC03D5-41ED-4B1E-8A45-B2EC8BDE3EA0}\\lib
\\*;%CLASSPATH%
```

Step 5 (For Cloudera 3 distribution only) Add the following line in the *service.props* file:

```
jarlib=cloudera
```

Step 6 (For Cloudera 4 distribution only) Add the following lines in service.props file:

```
jarlib=cdh4
CLASSPATH=C:\\Program
Files\\TIDAL\\Scheduler\\Master\\services\\{D9AC03D5-41ED-4B1E-8A45-B2EC8BDE3EA0}\\lib
\\*;%CLASSPATH%
```

Step 7 (For MapR Distribution only) Install MapR client in the TES Master machine, and add the following lines in the *service.props* file for your platform:

Windows:

```
jarlib=mapr
JVMARGS=-Djava.library.path=C:\\opt\\maprv1\\hadoop\\hadoop-0.20.2\\lib\\native\\Windo
ws_7-amd64-64
CLASSPATH=C:\\opt\\maprv1\\hadoop\\hadoop-0.20.2\\lib\\*;%CLASSPATH%
```

Linux:

```
jarlib=mapr
JVMARGS=-Djava.library.path=/opt/mapr/hadoop/hadoop-0.20.2/lib/native/Linux-amd64-64
CLASSPATH=/opt/mapr/hadoop/hadoop-0.20.2/lib/*:${CLASSPATH}
```

All paths above are derived from the MapR Client installation. If a filename does not exactly match, use the match closest to the filename.

See MapR Client Software Requirements for more requirements for using MapR.

Step 8 (Optional) Add properties to control authentication:

kerbrealm – If the Hadoop cluster is Kerberos secured, use this value to specify the Kerberos Realm.

For example, **kerbrealm=TIDALSOFT.LOCAL**.

kerbkdc – If the Hadoop cluster is Kerberos secured, use this value to specify the KDC Server.

For example, **kerbkdc=172.25.6.112**

The Adapter supports both Simple and Kerberos authentication. For Kerberos, the following limitations exist:

- The Adapter does not support multiple Realms and KDC Servers.
- If using both Simple and Kerberos authentication when connecting to multiple Hadoop environments, you must include the kerbrealm and kerbkdc properties in service.props.

- When connecting using Simple authentication, the Kerberos options will be ignored, but are required when connecting to a Kerberos secured environment.

Step 9 (Optional) Add properties to service.prop to control the polling, output, and logging for the MapReduce Adapter. See Configuring service.props.

Step 10 Restart the Master.

Securing the Adapter

There are two types of users associated with the MapReduce Adapter, **Runtime Users** and **Schedulers**. You maintain definitions for both types of users from the **Users** pane.

- **Runtime Users**

  Runtime users in the context of MapReduce jobs represent those users and passwords required for authentication when submitting jobs. MapReduce operations require authentication against a valid user as defined by a Hadoop administrator.

- **Schedulers**

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

  – Security policy that grants or denies add, edit, delete and view capabilities for MapReduce jobs.

  – Authorized runtime user list that grants or denies access to specific authentication accounts for use with MapReduce jobs.

  – Authorized agent list that grants or denies access to specific MapReduce Adapter connections for use when defining MapReduce jobs.

Defining Runtime Users

To define a Hadoop MapReduce connection, a Hadoop user must be specified. A Hadoop MapReduce user is a user with a Hadoop password.

Note The password is not used in this initial release of the MapReduce Adapter, but is required in anticipation of future support. It is required in order to limit the user combo boxes to only MapReduce users.

To define a runtime user:

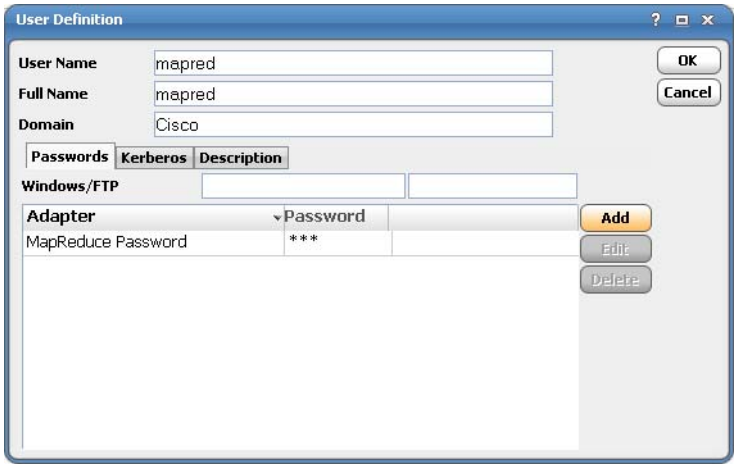
Step 1 From the **Navigator** pane, expand the **Administration** node and select **Runtime Users** to display the defined users.

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

-or-

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

The **User Definition** dialog box displays.



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

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

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

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

The **Change Password** dialog box displays.

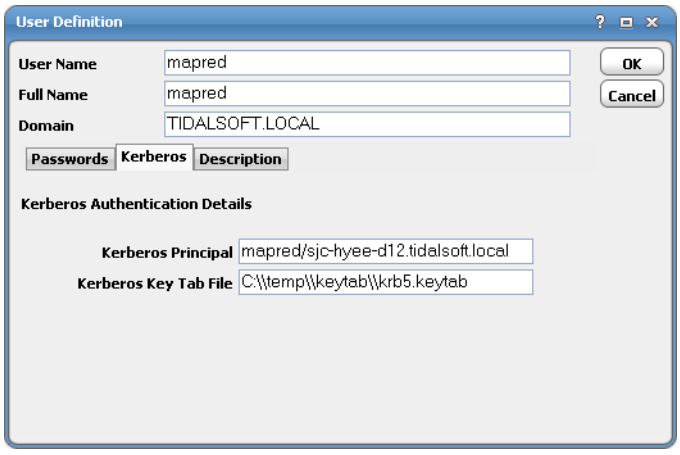
Step 7 Select **MapReduce** from the **Password Type** list.

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

Only those users with a password specified for MapReduce will be available for use with MapReduce jobs. The password might be the same as the one specified for Windows/FTP jobs.

Step 9 Click **OK** to return to the **User Definition** dialog box. The new password record displays on the **Passwords** tab.

Step 10 Click the **Kerberos** tab. If your Hadoop cluster is Kerberos secured, the Kerberos Principal and Kerberos Key Tab file is required. The Kerberos principal specifies a unique identity to which Kerberos can assign tickets. The Key Tab file is relative to the Master's file system and contains one or more Kerberos principals with their defined access to Hadoop.



Note The figure above shows the case for a Windows TES master where "\\" are uses as path separator. For Unix, the separator will be "/".

Step 11 Click **OK** to add or save the user record in the TES database.

For further information about the **User Definition** dialog box, see your *Cisco Tidal Enterprise Scheduler User Guide*.

Authorizing Schedulers to Work With MapReduce Adapter Jobs

There are two steps involved in authorizing schedulers to work with MapReduce Adapter jobs:

- Defining a Security Policy
- Defining Scheduler Users for MapReduce Adapter Jobs

Defining a Security Policy

Access to the MapReduce environment is controlled by assigning a MapReduce security policy with specified privileges to user accounts. The system administrator should create a new security policy or edit an existing policy in Scheduler as described below, that in addition to the normal user privileges, includes the capability to add and/or edit MapReduce jobs.

A user whose assigned security policy does not include MapReduce privileges cannot create and/or run MapReduce jobs.

To authorize Schedulers by defining a security policy:

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

Step 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**.

Step 3 If adding a new policy, click in the **Security Policy Name** field and enter a name for the policy.

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

Step 5 Click **Close** on the **Function** drop-down list.

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

For further information about setting up security policies, see your *Cisco Tidal Enterprise Scheduler User Guide*.

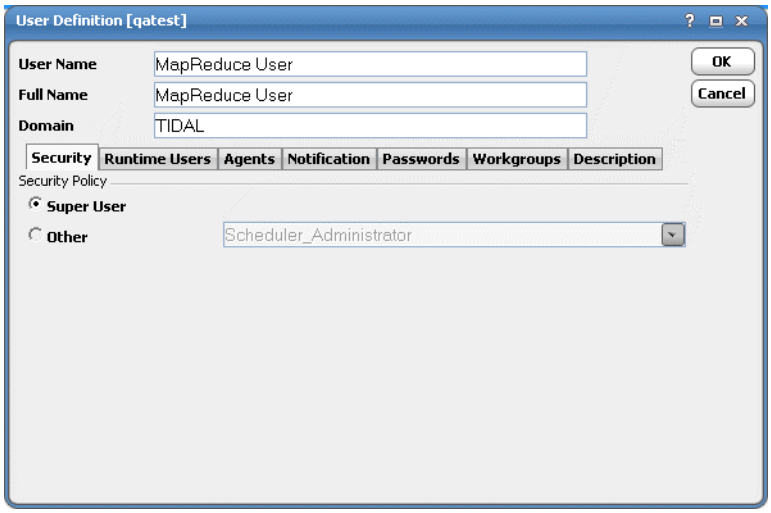
Defining Scheduler Users for MapReduce Adapter Jobs

To define a Scheduler user to work with MapReduce Adapter jobs:

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

Step 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 box displays.



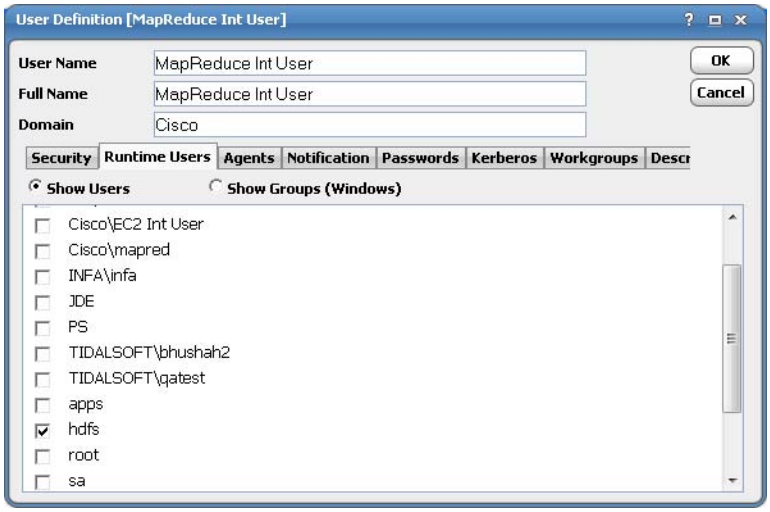
Step 3 If this is a new user definition, enter the new user name in the **User Name** field.

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

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

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

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



Step 8 Select the MapReduce Adapter users that this scheduling user can use for submitting MapReduce Adapter jobs.

Step 9 Click the **Agents** tab.

Step 10 Select the check boxes for the MapReduce Adapter connections that this scheduling user can access when scheduling jobs.

Step 11 Click **OK** to save the user definition.

Defining a Connection

You must create one or more Hadoop MapReduce connections before TES can run your MapReduce Adapter jobs. These connections also must be licensed before TES can use them. A connection is created using the **Connection Definition** dialog box.

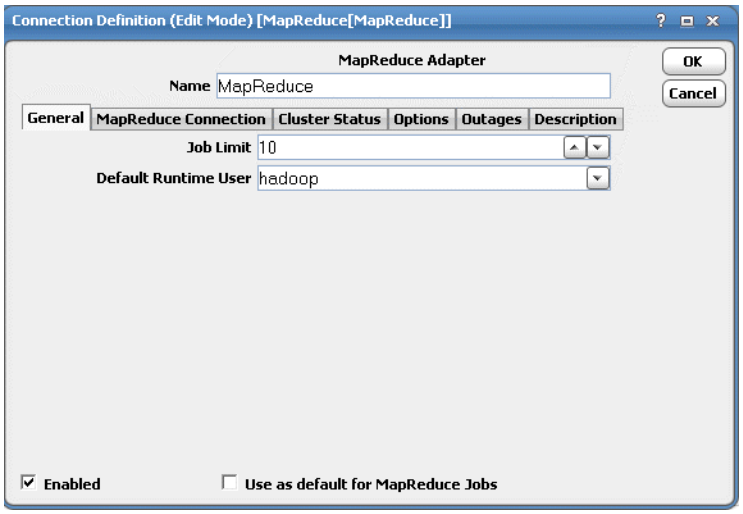
To define a Hadoop MapReduce connection, a job tracker, name node, and Hadoop user must be specified. A Hadoop MapReduce user is a user with a Hadoop password.

To define a connection:

Step 1 From the **Navigator** pane, navigate to **Administration>Connections** to display the **Connections** pane.

Step 2 Right-click **Connections** and select **Add Connection>MapReduce Adapter** from the context menu.

The **MapReduce Adapter Connection Definition** dialog box displays.



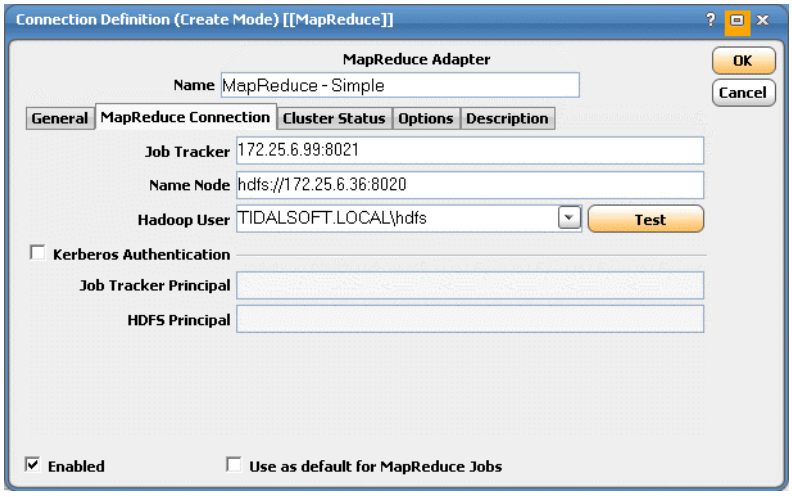
Step 3 On the **General** page, enter a name for the new connection in the **Name** field.

Step 4 In the **Job Limit** field, select the maximum number of concurrent active processes that TES submits to the Hadoop server at one time.

Step 5 From the **Default Runtime User** drop-down list, you have the option to select the name of a default user for MapReduce Adapter jobs. The runtime user is auto-selected when defining MapReduce Adapter jobs.

Only authorized users that have been defined with MapReduce passwords display in this list. The selected user is automatically supplied as the default runtime user in a new MapReduce Adapter job definition.

Step 6 Click the **MapReduce Connection** tab.



Step 7 In the **Job Tracker** field, enter the location of your Job Tracker.

Step 8 In the **Name Node** field, enter the URI of the Name node

Note For MapR, job tracker and name node must be set to “maprfs:///”.

Step 9 From the **Hadoop User** list, select the associated Runtime User for MapReduce to be used to monitor connection health and job execution.

This is a persistent user connection that is only used for administration and monitoring and for jobs with a matching runtime user. Jobs with a different runtime user specified will create additional temporary connections.

Note It is recommended that the connection's Hadoop user be a Hadoop Super User and is a requirement to display Distributed File System statistics in the Cluster Status tab.

Step 10 (Optional) Click the **Test** button to verify connectivity. If successful, the following message displays:

Step 11 (Optional) If the Hadoop cluster is secured by Kerberos, specify the Kerberos information.

Step 12 Select the **Kerberos Authentication** check box and specify the Job Tracker and HDFS Kerberos Principals.

The Kerberos User Principal and Kerberos Key Tab file associated with the Hadoop user is configured during Step 10 on page 11.

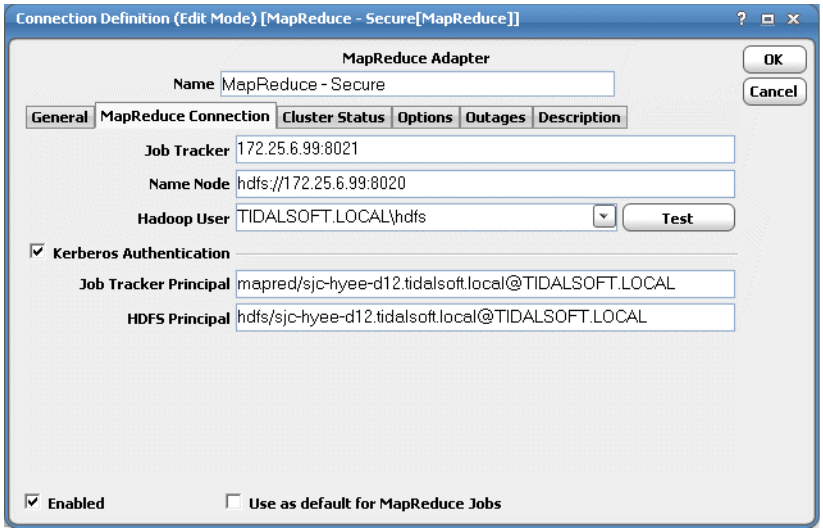
A Kerberos principal is used in a Kerberos-secured system to represent a unique identity. Kerberos assigns tickets to Kerberos principals to enable them to access Kerberos-secured Hadoop services. For Hadoop, the principals should be in the following format:

```
username/fully.qualified.domain.name@YOUR-REALM.COM
```

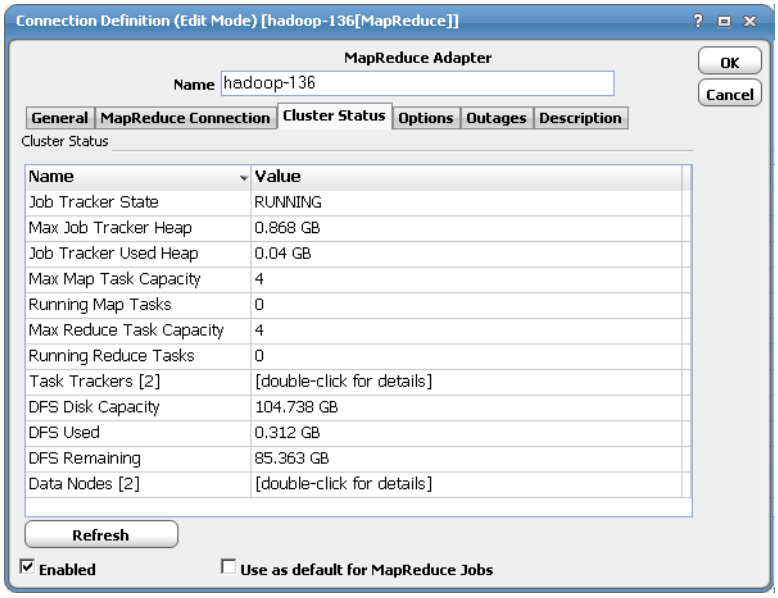
where username refers to the username of an existing Unix account, such as hdfs or mapred.

Note MapR distribution does not support Kerberos.

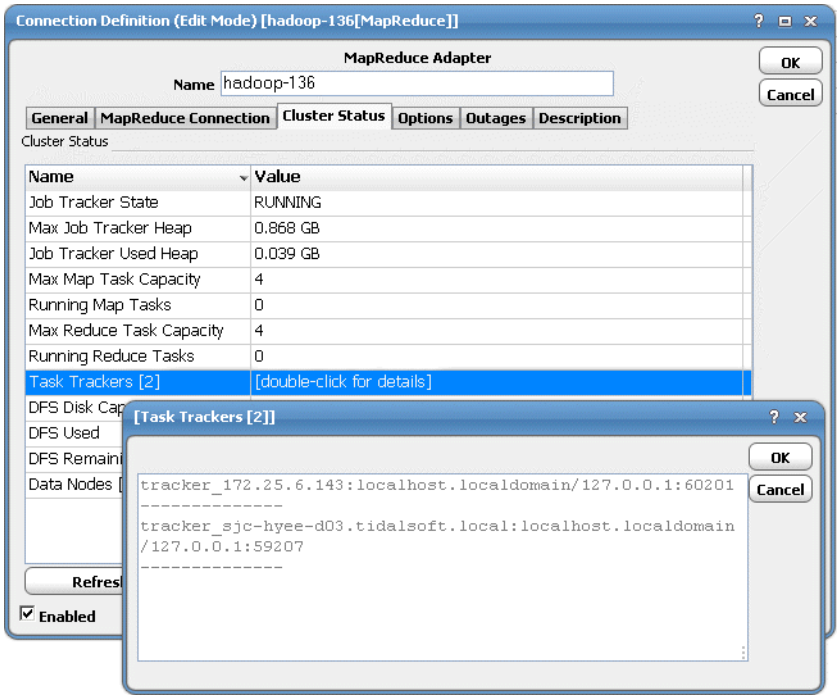
Kerberos authentication is optional, but if the Hadoop cluster is secured by Kerberos, Kerberos information must be configured.
1. Service.props
2. MapReduce User Definition
3. MapReduce Connection Definition

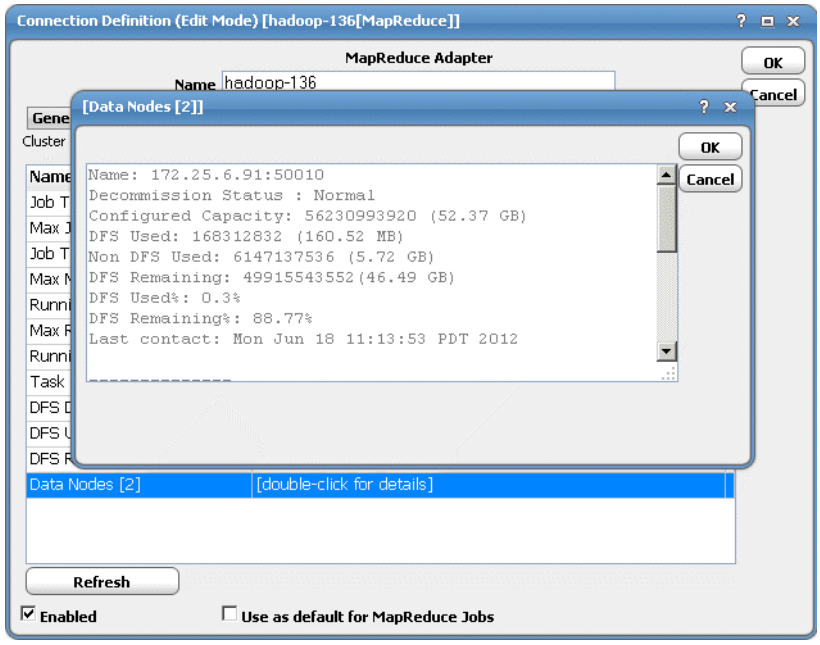


Step 13 Click the **Cluster Status** tab to display current cluster's status in real time. This is for the display of Distributed File System (DFS) info and requires a Hadoop Super User.



You can double-click **Task Trackers** and **Data Nodes** to view the following additional dialog boxes:



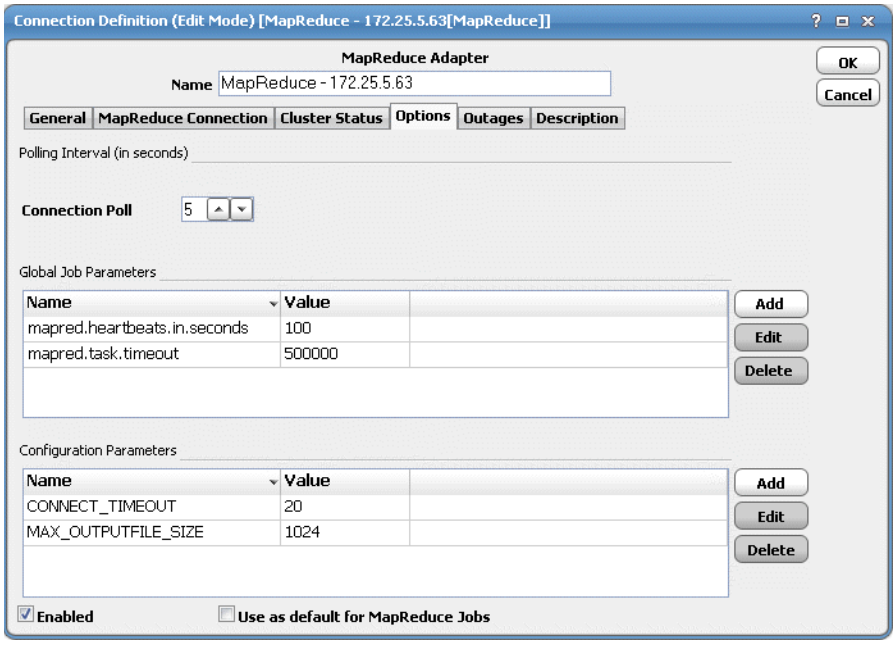


Step 14 Click the **Options** tab to specify Global Job Parameters that are applicable to all jobs using the connection. If the job definition specifies the same job parameters, the values defined in the job definition will override the corresponding connection values. The Configuration Parameters are general connection parameter options. The supported configuration parameters includes:

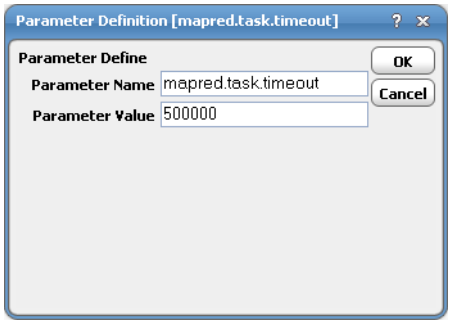
- **CONNECT_TIMEOUT** – the timeout interval in seconds (default 20) in which a failed connection will timeout, avoiding further connection retries.

- **MAX_OUTPUTFILE_SIZE** – in kbytes. (default 1024 kbytes) This option is used to determine whether or not MapReduce output should be retrieved during output collection. If the output file exceeds this configured limit, output collection for this file will be avoided.

.



Step 15 To add a parameter, click **Add** to display the **Parameter Definition** dialog box.



Step 16 Click **OK** to save the new MapReduce connection. The configured connection displays in the **Connections** pane.

Verifying Connection Status

The status light next to the connection indicates whether the TES Master is connected to the MapReduce server. If the light is green, the MapReduce server is connected.

A red light indicates that the master cannot connect to the MapReduce server. MapReduce jobs will not be submitted without a connection to the MapReduce 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 MapReduce Connection** dialog box describing the problem. Or go to **Operator>Logs** to look for error messages associated with this connection.

Using the MapReduce Adapter

Overview

This chapter describes how to use the MapReduce Adapter:

- Defining MapReduce Jobs
- Monitoring MapReduce Job Activity
- Controlling Adapter and Agent Jobs

Defining MapReduce Jobs

This section provides instructions for defining a MapReduce job in TES and descriptions of the various options that can be included in the jobs.

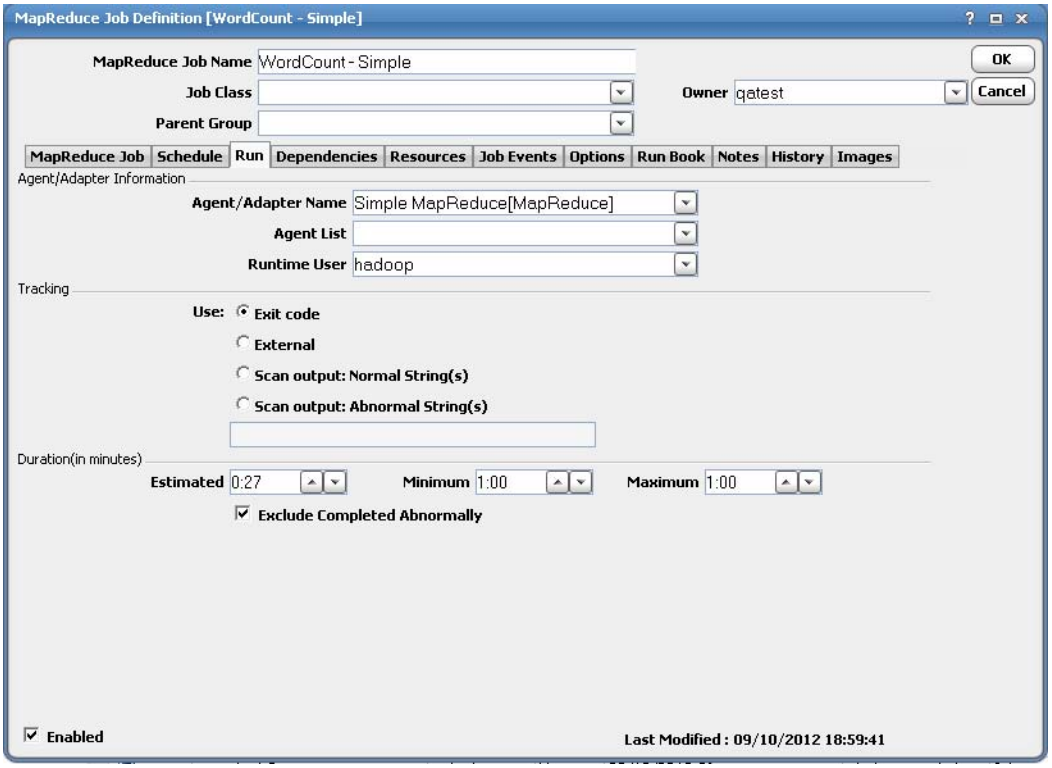
Note To execute and monitor a custom MapReduce job, all files (mapper/reducer classes, dependent libraries, resource files, etc.) needed by the MapReduce job must already exist in HDFS. The Adapter does not move dependent files in/out of HDFS.

To define a MapReduce job:

Step 1 In the **Navigator** pane, select **Definitions>Jobs** to display the **Jobs** pane.

Step 2 Right-click **Jobs** and select **Add Job>MapReduce Job** from the context menu.

The **MapReduce Job Definition** dialog box displays.



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

Step 3 In the upper portion of the dialog box, 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 drop-down list. This field is optional.
- **Owner** – Select the MapReduce owner of the selected report/ Web Intelligence. The user must have the appropriate MapReduce authority for the operation.
- **Parent Group** – If this job exists under a parent group, select the name of the parent group from the drop-down list. All properties in the **Agent Information** section are inherited from its parent job group.

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

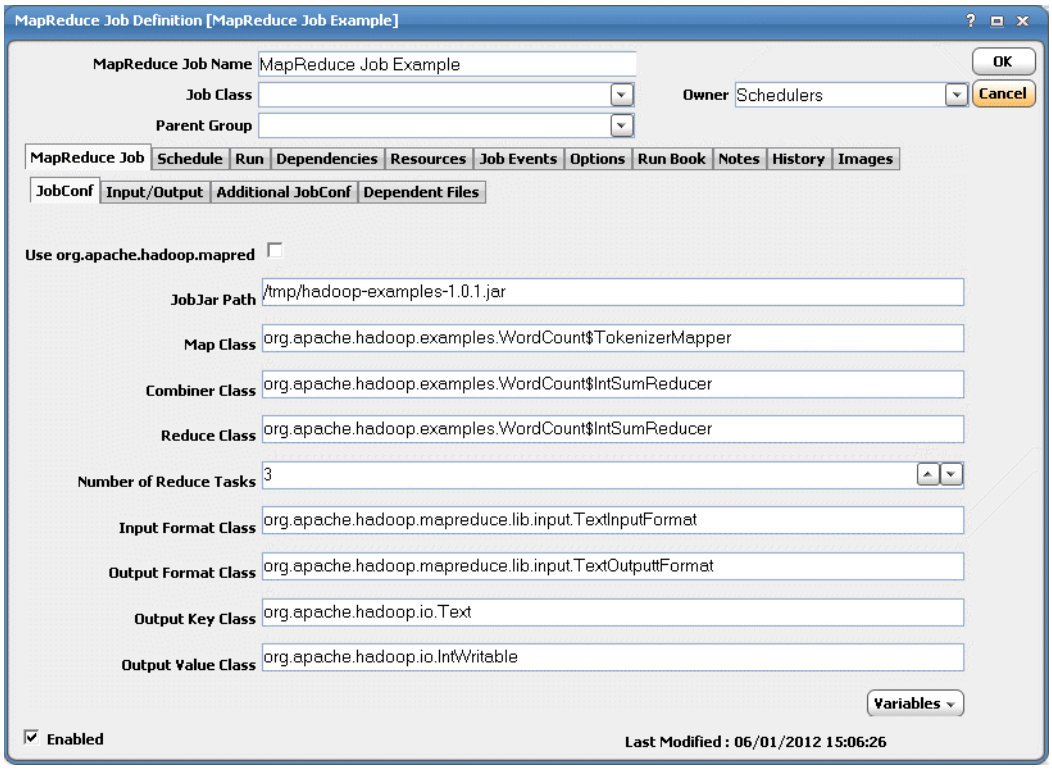
- **Agent/Adapter Name** – Select the MapReduce adapter connection to be used for this job from the drop-down list.

  -or-

  Agent List Name – Select a list for broadcasting the job to multiple servers.
- **Runtime User** – Select a valid runtime user with the appropriate MapReduce authority for the job from the drop-down list.

Step 5 Specify the appropriate Tracking and Duration information for the job. Refer to the *Cisco Tidal Enterprise Scheduler User Guide* for information on these options.

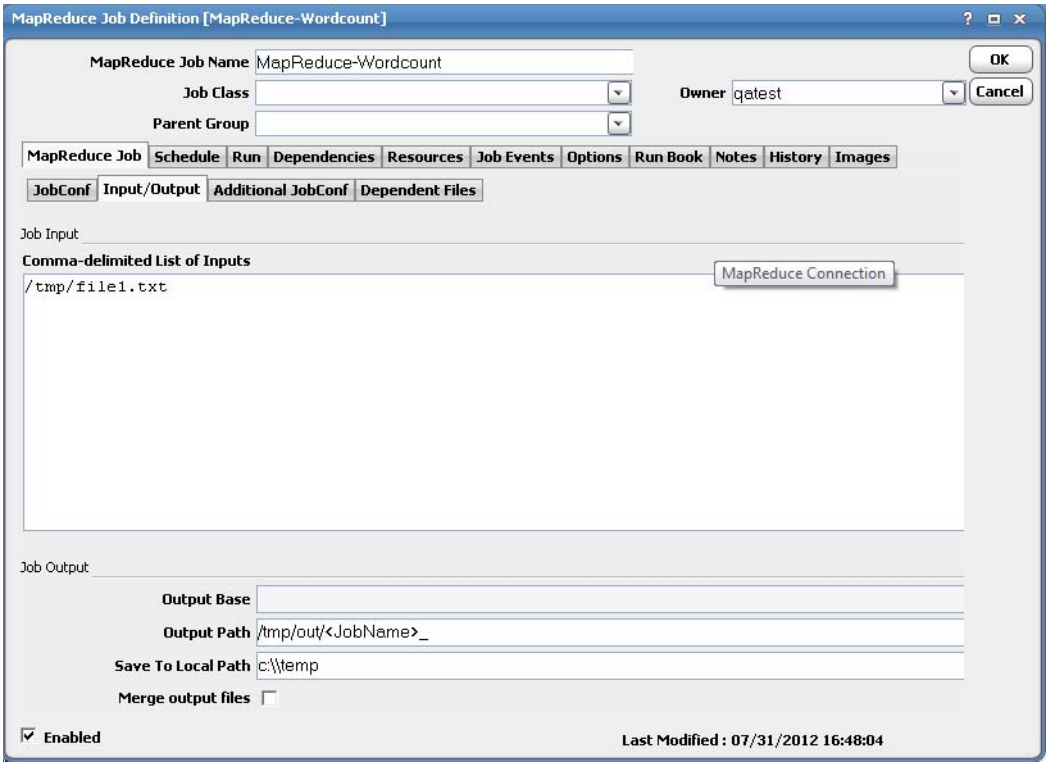
Step 6 Click the **MapReduce** tab, then click the **Job Config** subtab to specify the job configuration, which includes the classes of the MapReduce job.



This subtab contains the following:

- **Use org.apache.hadoop.mapred** – specifies the API package implementations for the map and reduce functions of the MapReduce job. Selecting this check box indicates the “mapred” package is used for the MapReduce job. By default, this is unchecked which indicates the “mapreduce” implementation. This is required by the Adapter in order to set the correct **jobConf** class properties for the map, reduce, combine, input , and output format classes.
- **JobJar Path** – the MapReduce job jar
- **Map Class** – the map class
- **Combiner Class** – the combiner class
- **Reduce Class** – the reduce Class
- **Number of Reduce tasks** (defaults to 1) – number of reduce tasks
- **Input Format Class** – the job's input format class
- **Output Format Class** – the job's output format class
- **Output Key Class** – the jobs’ output key class
- **Output Value Class** – the jobs’ output value class

Step 7 Click the **Input/Output** subtab to specify the inputs and outputs for the job.



The job input can consists of multiple file or directories, delimited by commas.

- **Job Input** – The job input can consist of multiple file or directories, delimited by commas.
- **Job Output** – In this section, specify the job output configuration.
  - **Output Base** – This option is only available if the **Use org.apache.hadoop.mapred** check box is unchecked on the **JobConf** tab. It is used to change the output base name. By default, the system generates an output name such as **part-00000** or **part-r-00000** (when using the **org.apache.hadoop.mapreduce** implementation, where **00000** is the output file number determined by the reducer number of your MapReduce job. This option allows you to override the output base (**part-**).
  - **Output Path** – Specify the output directory to be created for the MapReduce job. Upon completion, each reduce task generates output written to this directory. The trailing text box provides a default suffix to be concatenated to the output path to create a unique output path per job run.

Note The MapReduce job will fail to launch if the output directory already exists. The trailing text box specifies a suffix that can be appended to the Output Path to generate a unique output directory. It is recommended to include an output suffix that will be unique between runs. (The default when creating a job is <JobID> specifying Tidal Variable for jobrun id).

  - **Save to Local Path** – (Optional) This option is used to save the output to a local directory relative to the TES Master server. In the **Save to Local Path** field, enter the local path to a local directory relative to the TES Master server. The local directory must exist. The syntax includes an existing directory with an optional file name. The file name does not need to exist.

If no file name is included in the path, the output (which may include multiple files) will be copied from hdfs into the local system with the original file name(s).

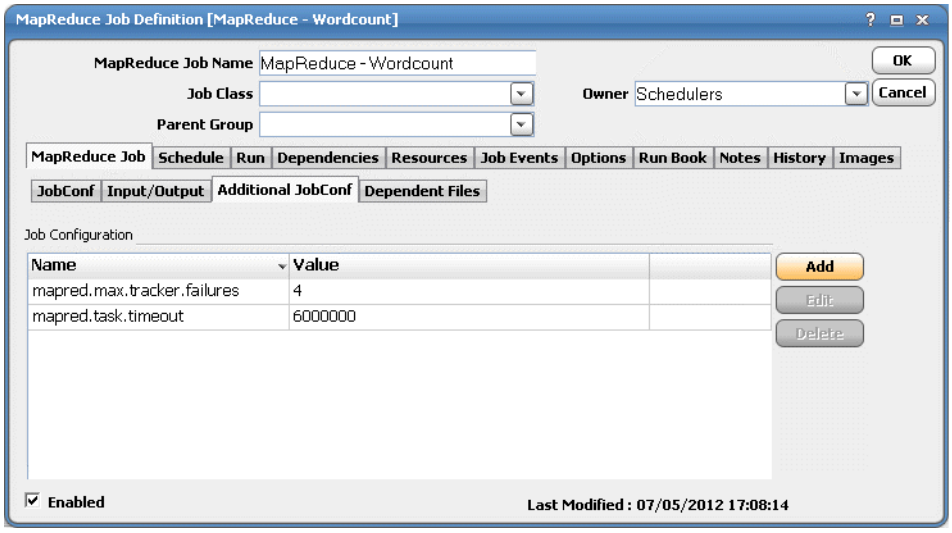
Select the **Merge Output Files** option if you are using the **Save to Local Path** option. If the MapReduce job produces multiple output files, the file will be merged into a single file on the local system.

Note You can generate a merged output without selecting this option as long as a user-specified file name is supplied.

If no filename is included, the merged file will be created with the name of the MapReduce job.

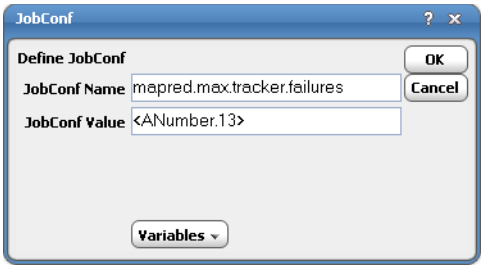
If a file name is included, the merged contents will be created with the file name specified. This behaves the same not selecting this option and including a user specified file name.

Step 8 Select the **Additional JobConf** tab to specify additional parameters for the job. Specify both the name and the value of the parameter. These options take precedence over the corresponding options specified at the connection level.

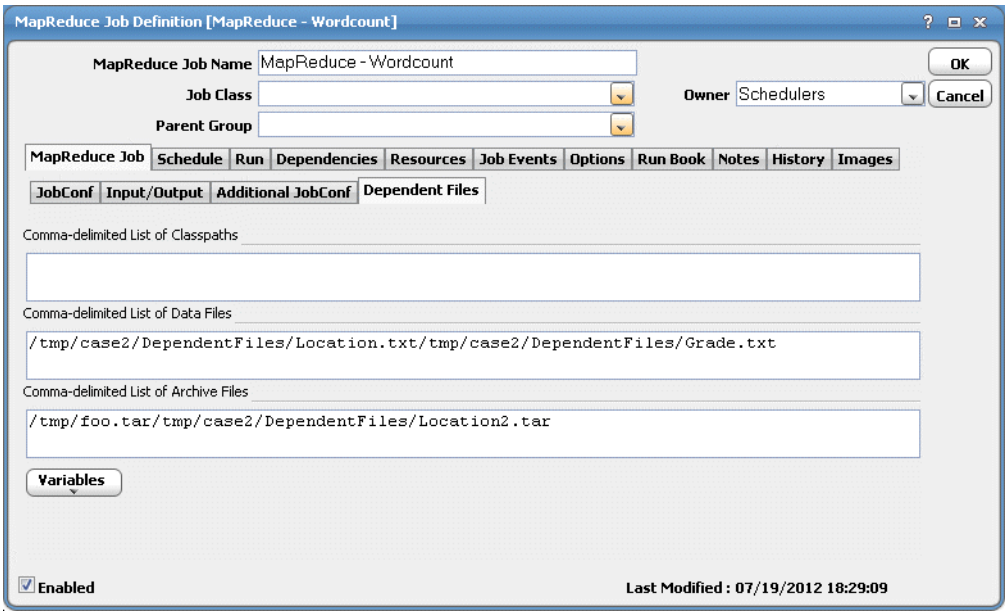


Note To assist with debugging, you can add a the tidal.debug.jobconfig parameter and set its value to Y or N. This parameter generates a local job config XML file in the Tidal MapReduce Service Temp folder. You can use this file to run a job directly on the cluster to debug job configuration related issues.

Click **Add** to add a new parameter or select an existing parameter and click **Edit** to display the **Job Config** dialog box.



Step 9 Select the **Dependent Files** tab to specify data and archive files to be copied to the Task Tracker at runtime.



Archived files are copied and unarchived. Classpath files are copied to the Task Tracker's local file system and added to the task JVM's classpath.

Step 10 Click the **Options** tab to specify Output settings such as **Summary Only**. When the **Summary Only** option is checked, Map/Reduce output will not be collected as part of output. If unchecked, Map/Reduce output will be include output for Map/Reduce tasks. For further information, see your *Cisco Tidal Enterprise Scheduler User Guide.*

Step 11 Click **OK** to save the job.

Monitoring MapReduce Job Activity

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

To monitor job activity:

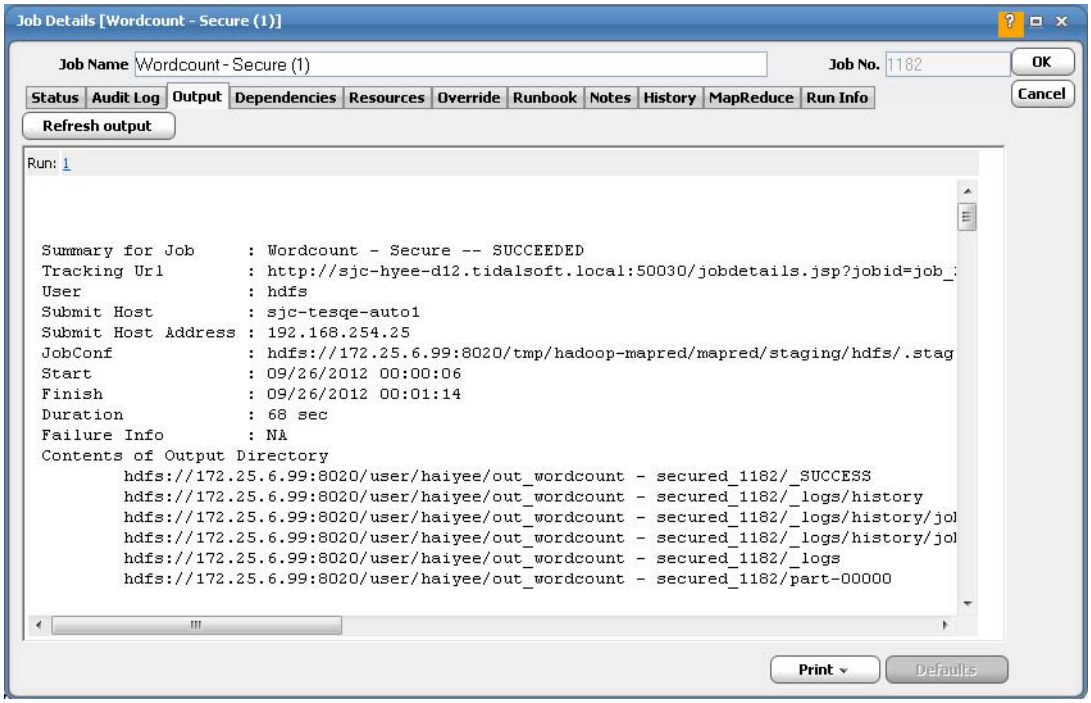
Step 1 In the **Navigator** pane, select **Operations>Job Activity** to display the **Job Activity** pane.

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

The **Job Details** dialog box displays. The **Status** page displays by default. 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 MapReduce job number.

Step 3 Click the **Audit Log** tab to view all job related messages. All other audit messages will appear in the central logs (**Operation>Logs**).

Step 4 Click the **Output** tab to view a task summary after the job completes.



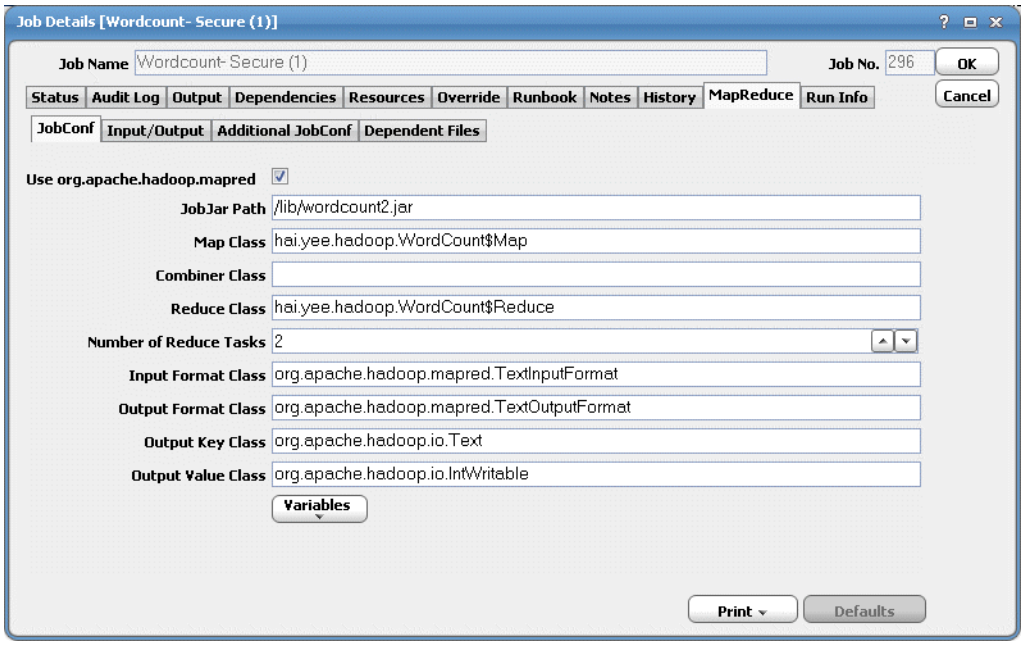
Job output consists of:

- Summary information for the MapReduce
- Summary information for each Map task
- Summary information for each Reduce task
- If **Summary-Only** is unchecked on the **Options** tab of the **Job Definition** dialog box, output contents of each Reduce task are included.

For further information on the **Options** tab of the **Job Definition** dialog box, see your *Cisco Tidal Enterprise Scheduler User Guide*.

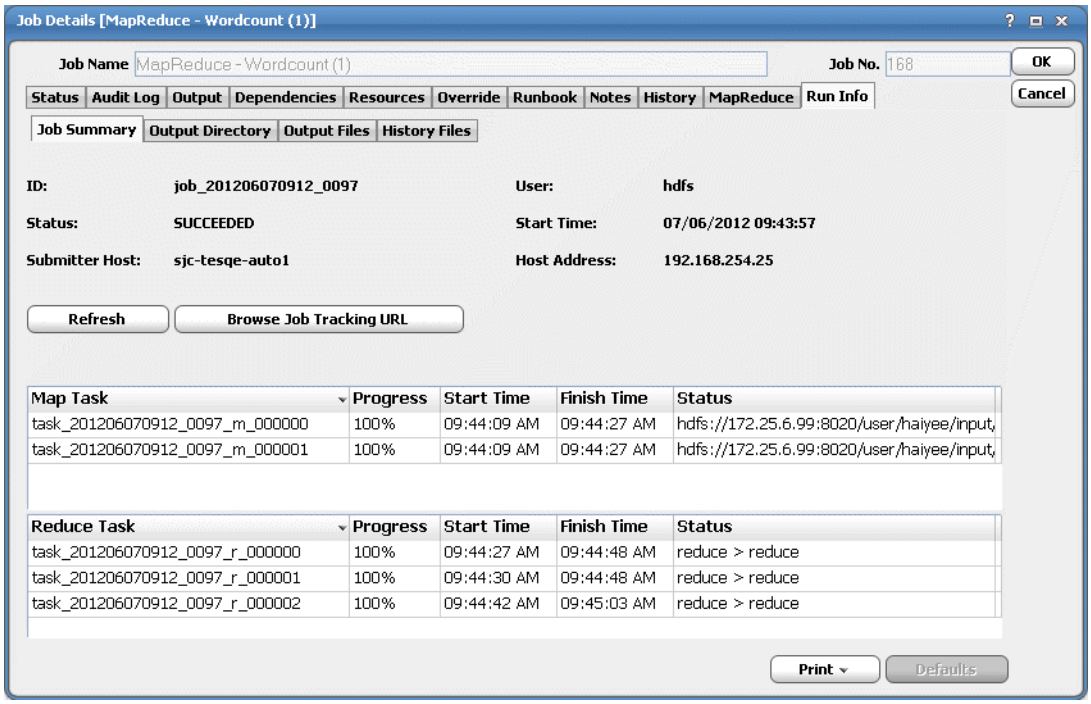
Note If the output file exceeds the limit specified by Connection configuration option MAX_OUTPUTFILE_SIZE, output contents for this file will not be included.

Step 5 Click the **MapReduce** tab to view the job definition details and the variables that were used when the job was submitted.



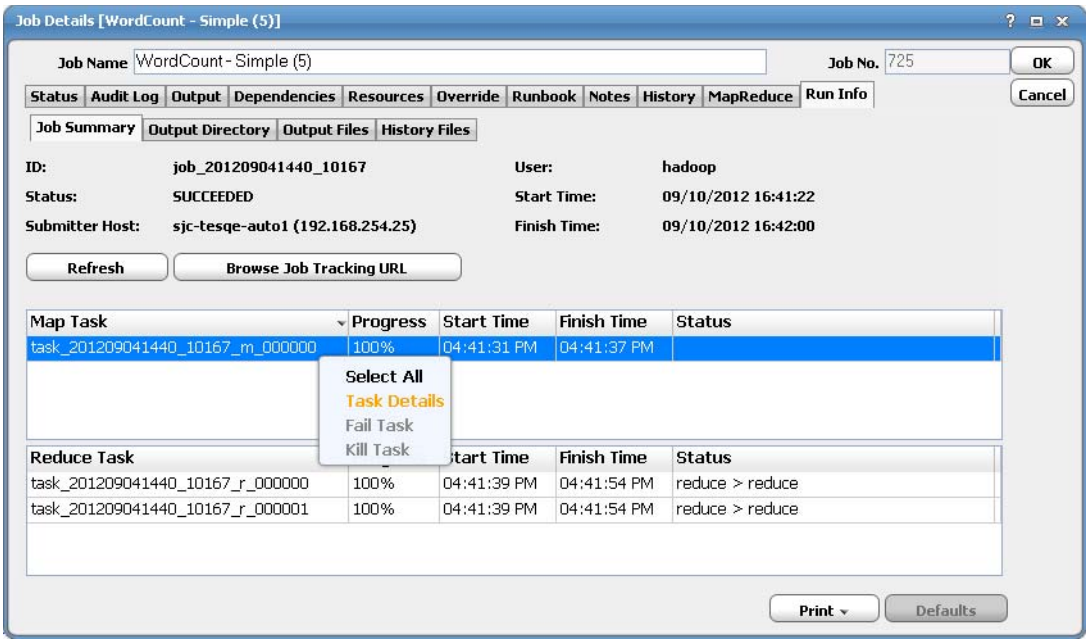
While the job is running, the fields are disabled; however, prior to running or rerunning the job, you can override any value on this tab. Your changes here only apply to this instance of the job (the original job definition is not affected).

Step 6 Click the **Run Info** tab to view real time execution data of the MapReduce job as it is running.



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).

Step 7 Click the **Job Summary** tab to view a summary of the job.



This tab includes the following summary information:

- **ID** – the Job ID
- **Submit Host** – the host name/host address which submitted the job (eg: the TES Master machine submitting the job)
- **Status** – current job status
- **User** – user who submitted the job
- **Start Time** – time the job started
- **Finish Time** – time the job finishes

Note Finish time is available as long as the job is configured to collect job history.

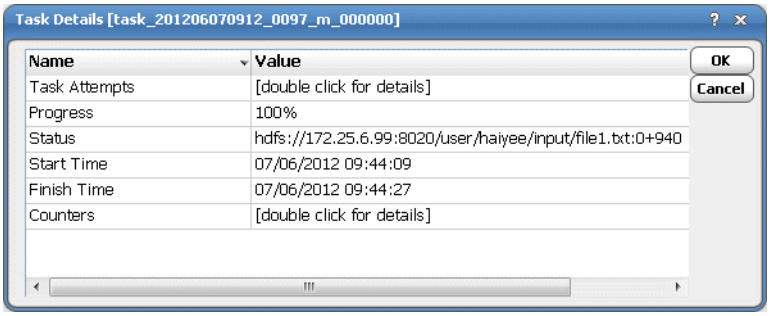
- **Refresh** – job status refresh request to update the display with current values.

Note Using refresh button after completion of a Hadoop job may result in an incomplete or empty status. This is mostly caused by a retired Hadoop job.

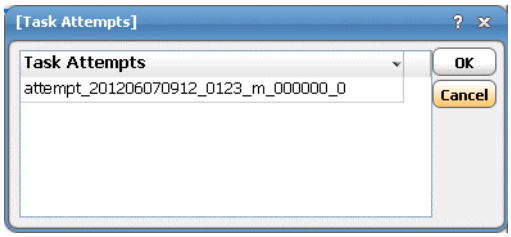
- **Browse Job Tracking URL** – opens the job tracking url in a browser.
- **Output Directory Tab** – display contents of the output directory
- **Output Files Tab** – displays a list of output files, if any (will not include 0 byte files)
- **History Files Tab** – displays the job history files, if any (will not include 0 byte files)

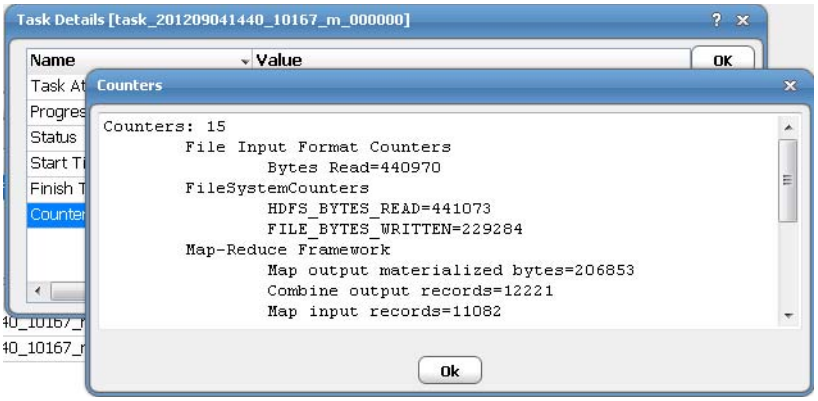
- **Tasks** – displays the current status of both map and reduce tasks. Menu options for a selected task include **Task Details**, **Fail Task** and **Kill Task**.
- **Browse Job Tracking URL** – opens the job tracking URL in a browser.

Additionally, you can view additional task details by double-clicking the task row or by right-clicking the task row, and then selecting **Task Details** from the context menu. The **Task Details** dialog box displays.



To view **Task Attempts** and **Counters** details, double-click the row to display the respective dialog boxes.

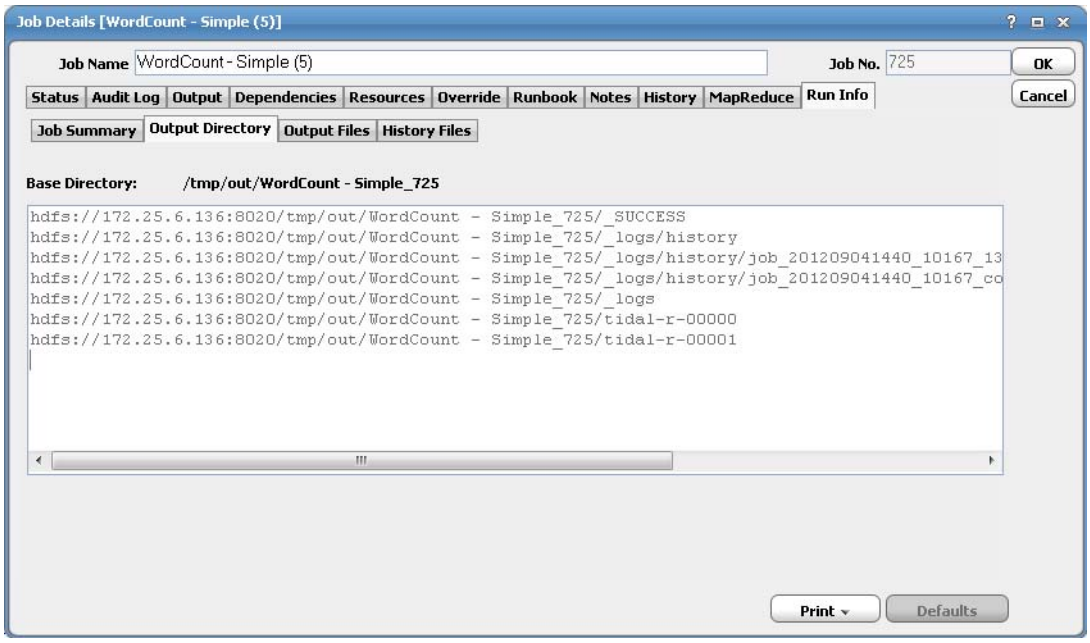




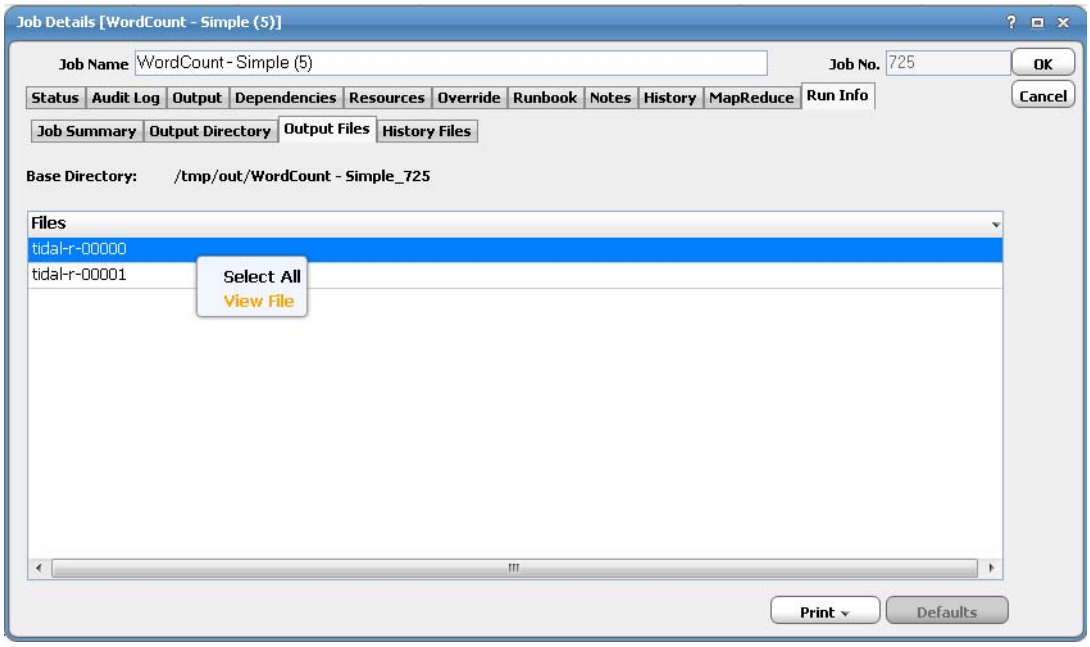
While a task is running, you can Fail or Kill the task. Refer to Hadoop MapReduce documentation for further details.

During Kill/Fail of a map/reduce task, only the latest/last task attempt will be tried. You can confirm this by clicking **Browse Job Tracking URL**.

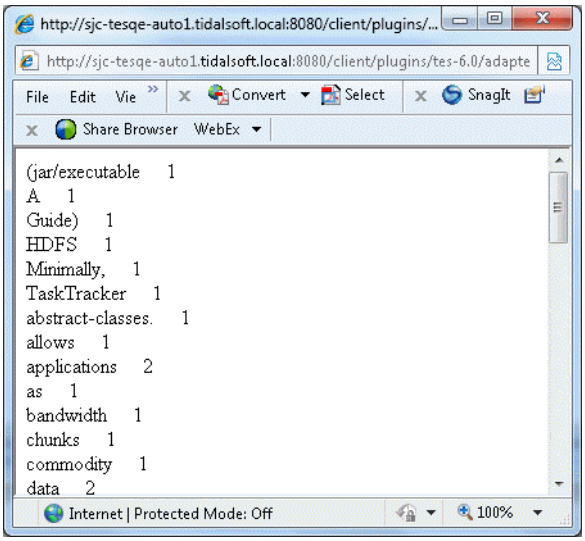
Step 8 Click the **Output Directory** tab to view the MapReduce output directory contents.



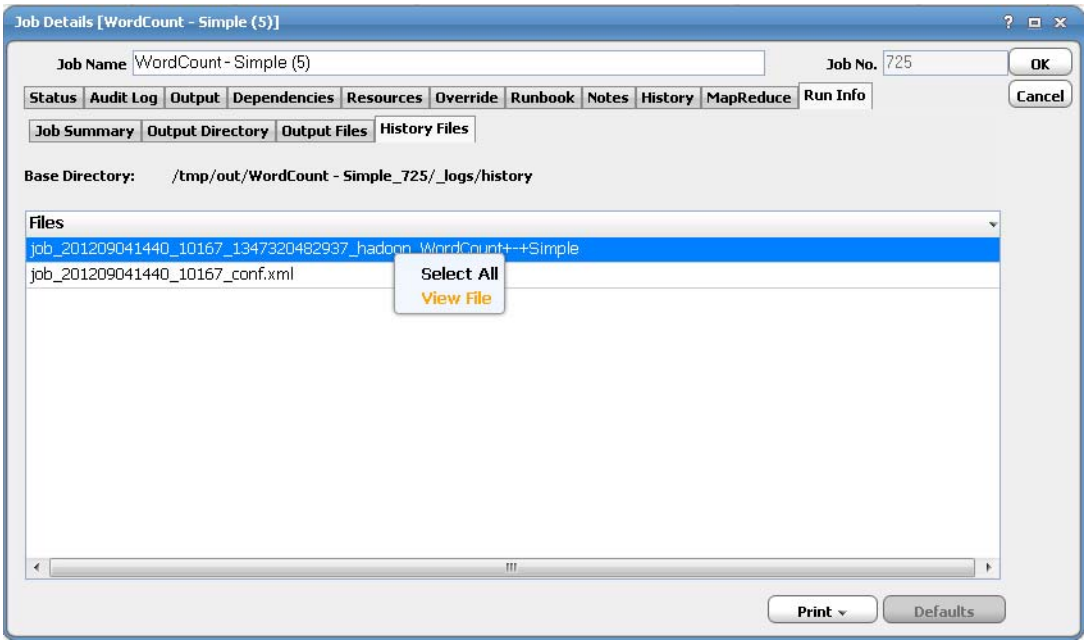
Step 9 Click the **Output Files** tab to view all reduce output files, if any, generated by the MapReduce Jobs. Output files of length 0 bytes will not be displayed.



Additionally, you can right-click a file and select **View File** from the context menu to view the file in the native Web browser.



Step 10 Click the **History Files** tab to view MapReduce job history file contents, if any.



Additionally, you can right-click a file and select **View File** from the context menu to view the file in the native Web browser.

Step 11 When you have finished viewing the job activity details, click **OK** to close the dialog box.

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:

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

Step 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:

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

Step 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:

Step 1 From the **Job Activity** pane, right-click the Adapter/Agent job you need to rerun.

Step 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:

Step 1 From the **Job Activity** pane, double-click the Adapter/Agent job to display the **Job Details** dialog.

Step 2 Click the Adapter tab.

Step 3 Make the desired changes to the job and click **OK** to close the **Job Details** dialog.

Step 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:

Step 1 From the **Job Activity** pane, right-click the Adapter/Agent job to be deleted.

Step 2 Select **Remove Job(s) From Schedule**.

Troubleshooting the MapReduce Adapter

Overview

This chapter describes how to troubleshoot issues for the MapReduce Adapter:

Review Service Log Files for More Information

Refer to the log files for further information regarding an issue.

Connection Failures

- Verify the hostname to IP address mappings in the hosts file
- For secured connections, verify *service.props* has been correctly set up to support Kerberos.
- Verify that the user keytab file exists and is accessible to TES master.
- Hadoop:GSSException: No valid credentials provided :Clock skew too great
  - Occurs when the clock skew between the KDC and clients exceed a maximum threshold (default 5 minutes). Maintaining synchronized clocks between the KDCs and Kerberos clients (TIDAL) is required, therefore Network Time Protocol (NTP) software or other similar time service tools must be used to synchronize them. Updating the Master's clock to the KDC server time will temporarily address the issue, but a time service software must be used to keep them synchronized.

Job Failures

- Verify your job is configured correctly.
- Verify your job can be run via the Hadoop CLI before running thru TES.
- Check the Adapter logs and verify how your job ran from the Hadoop Admin Console.
- Verify the file paths and names are case-sensitive and that they exist on the HDFS.
- Class Not Found Exception.

  Verify that the job JAR, as well as all referenced library JARs, exists in HDFS and that the path specified is correct.

Adapter Is Out-of-Memory

Adapter memory sizes are verified on a 10-node cluster and can be increased in the Adapter *service.props*.

Output Files Cannot Be Viewed

The output file either contains binary data or its size is greater than default size 1MB.

Cloudera, MapR, or Apache Connections Are RED

The Adapter cannot run with both cluster types at one time. By default, the Adapter runs with Apache Hadoop and can be switched to Cloudera or MapR in the *service.props* and after restarting the TES master.

Jobs Cannot Run on Apache Hadoop Version 1.0.2 or 1.0.3

These minor versions are unsupported, however, you can run your job by setting the following configuration in every job definition.

```
io.compression.codecs =
org.apache.hadoop.io.compress.DefaultCodec,org.apache.hadoop.io.compress.GzipCodec,org
.apache.hadoop.io.compress.BZip2Codec
```

MapReduce Job Runs Fine on the Hadoop Client but Fails via TES Adapter

Most likely, your TES MapReduce job configuration is not setup correctly. Compare the Hadoop job configuration file on the cluster with TES to resolve the problem (See your *Cisco Tidal Enterprise Scheduler User Guide* for capturing Hadoop job configuration on TES master.)

CHAPTER 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	All	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.

Property	Applicable Adapter(s)	Default	What It Controls
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.
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.
jarlib	Hive and MapReduce	<none>	Specifies the specific Java library to use for the adapter: • For Apache 1.1.2, add: **`jarlib=apache1.1.2`** • For Cloudera 3, add: **`jarlib=cloudera`** • For Cloudera 4, add: **`jarlib=cdh4`** • For MapR add: **`jarlib=apache1.1.2`**
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`**

Property	Applicable Adapter(s)	Default	What It Controls
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.
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	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.

Property	Applicable Adapter(s)	Default	What It Controls
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.
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.