TIBCO ActiveMatrix BusinessWorks™ Plug-in for Big Data User's Guide

Software Release 6.0
May 2014

Two-Second Advantage®

Important Information

Contents

TIBCO Documentation and Support Services

All TIBCO documentation is available in the TIBCO Documentation Library, which can be found here:

https://docs.tibco.com

Product-Specific Documentation

The following documents for this product can be found in the TIBCO Documentation Library:

- *TIBCO ActiveMatrix BusinessWorks Plug-in for Big Data Installation*
- *TIBCO ActiveMatrix BusinessWorks Plug-in for Big Data User's Guide*
- *TIBCO ActiveMatrix BusinessWorks Plug-in for Big Data Release Notes*

How to Contact TIBCO Support

For comments or problems with this manual or the software it addresses, contact TIBCO Support as follows:

- For an overview of TIBCO Support, and information about getting started with TIBCO Support, visit this site:

  http://www.tibco.com/services/support

- If you already have a valid maintenance or support contract, visit this site:

  https://support.tibco.com

  Entry to this site requires a user name and password. If you do not have a user name, you can request one.

How to Join TIBCOmmunity

TIBCOmmunity is an online destination for TIBCO customers, partners, and resident experts. It is a place to share and access the collective experience of the TIBCO community. TIBCOmmunity offers forums, blogs, and access to a variety of resources. To register, go to:

http://www.tibcommunity.com

Plug-in Overview

This plug-in allows users to manage deal with big data in a user-friendly way.

TIBCO ActiveMatrix BusinessWorks is an easy to use integration product suite for enterprise, web, and mobile applications. It uses the Eclipse graphical user interface (GUI) for defining business processes and the process engine to execute them.

TIBCO ActiveMatrix BusinessWorks Plug-in for Big Data plugs into TIBCO ActiveMatrix BusinessWorks, which bridges TIBCO ActiveMatrix BusinessWorks with Hadoop. It allows ActiveMatrix BusinessWorks users to established non-code approach to integrate with Hadoop family projects, such as Hadoop Distributed File System (HDFS).

The plug-in provides the following common functionalities:

- Copy files between HDFS and local file system, rename and delete files in HDFS.
- List the status of a file or directory, read and write data in a file.
- Create and queue a Hive, MapReduce, or Pig job.

Before running operations by using this plug-in, ensure that the user has appropriate permissions on HDFS server.

Getting Started

A typical workflow for using the plug-in to achieve different goals includes creating a project, designing a process, and deploying the application.

TIBCO ActiveMatrix BusinessWorks enables users to create services and integrate applications, and deploy them at runtime. It uses the Eclipse graphical user interface (GUI) for defining business processes and the process engine to execute them.

To design a process and deploy it at runtime, you need to complete the following tasks:

1. Creating a Project
2. Creating an HCatalog Connection
3. Designing a Process
4. Debugging and Running a Process
5. Deploying an Application

Creating a Project

BusinessWorks application modules are Eclipse projects that are created in TIBCO Business Studio, which contain various resources.

Procedure

1. Start TIBCO Business Studio.
2. Select **File > New > BusinessWorks Resources**.
3. Click the **BusinessWorks Application Module** resource in the **BusinessWorks Resource** wizard. Click **Next**.
4. Type a name for the project that you are creating in the **Project name** field.
5. Select the **Create empty process** check box and type a process name in the **Name** field.
6. Select the **Create Application** check box and type an application name in the **Name** field. Click **Finish**.
   A project with the default settings is opened in the Project Explorer view.

Creating a Connection

An HDFS Connection shared resource, which contains all the parameters for connecting to HDFS, is required when using the activities in the HDFS palette. An HCatalog Connection shared resource, which contains all the parameters for connecting to HCatalog, is required when using the activities in the Hadoop palette.

Prerequisites

The HDFS Connection and the HCatalog Connection shared resources are available at the Resources level. Before creating a Connection, you need to Creating a Project.

Procedure

1. Expand the created project in the Project Explorer view.
2. Right-click the **Resources** folder and click **New > HDFS Connection** or **HCatalog Connection**.
3. Type a name for the shared resource in the **Resource Name** field in the window. Click **Finish**.

4. Configure the shared resource in the editor.
   See HDFS Connection or HCatalog Connection regarding the configuration fields.
5. Click **Test Connection** to validate the connection.

What to do next

After the HCatalog connection is established successfully, you can click the **Database and Table** tab to introspect databases and tables existing in the HCatalog. See Database and Table for more information.

Designing a Process

Processes capture and manage the flow of business information in an enterprise between different data sources and destinations. You can design a process by using activities and adding conditions.

By default, an empty process is created when Creating a Project with the **Create empty process** check box selected.

See *TIBCO ActiveMatrix BusinessWorks Application Deployment* for more details about creating processes.

Procedure

1. In the Process editor, select and drop activities from the Palette view.
2. Click **Link**  to create links between the activities.
3. Configure the added activities.
4. Click **File > Save** to save the process.

Debugging and Running a Process

Debug the application you have configured to ensure that the application configuration is correct.

Procedure

1. Open the process you have configured in TIBCO Business Studio.
2. On the toolbar, click **Run > Debug Configurations**.
3. Click **BusinessWorks Application > BWApplication** in the left panel.
4. Ensure only the application you want to debug and run is selected in the **Applications** tab in the right panel.
5. Click the **Advanced** tab and click **Browse** to locate the `logback` file.
   By default, the log file resides in the *`TIBCO_HOME`*`/bw/6.1/config/design/logback` directory and error logs are captured. See Managing Logs for more details.
6. Click **Debug**.
   TIBCO Business Studio changes to the Debug perspective. Logs are displayed in the Console view.

Deploying an Application

You can manage BusinessWorks applications by using TIBCO Enterprise Administrator after deploying the applications.

Prerequisites

The following tasks are required before deploying applications:

- Creating a Project.
- Generating an EAR File.

A typical workflow of deployment includes:

1. Upload an EAR file.
2. Deploy an application.
3. Configure an application.
4. Start an application.

You can deploy an application EAR file by using the command-line mode with the **bwadmin** utility. See *TIBCO ActiveMatrix BusinessWorks Administration* for more details about how to deploy an application.

Generating an EAR File

Application archives are the enterprise archive (EAR) files that are created in TIBCO Business Studio. An EAR file is required when deploying an application.

Prerequisites

An application project has already been created, as explained in Creating a Project.

There are many ways to generate an EAR file, the following is one method. See *TIBCO ActiveMatrix BusinessWorks Administration* for more information.

Procedure

1. Go to File Explorer and click **Open Directory to Browse**  .
2. Select the folder where you want to generate the EAR file and click **OK**.
   The new folder is displayed in the File Explorer view
3. Drag the application from the Project Explorer to the new folder in the File Explorer.

   The EAR file is generated with the name `<application>_<version>.ear`.

Shared Resources

The HDFS palette contains one shared resource, HDFS Connection. The Hadoop palette contains one shared resource, HCatalog Connection (Database and Table).

HDFS Connection

The HDFS Connection shared resource contains all necessary parameters to connect to HDFS. It can be used by the HDFS Operation, ListFileStatus, Read, Write activities and the HCatalog Connection shared resource.

General

The **General** panel has the following fields.

Field	Module Property?	Description
Package	No	The name of the package where the shared resource is located.
Name	No	The name for the shared resource.
Description	No	A short description for the shared resource.

HDFSConfiguration

The **HDFSConfiguration** pannel has the following fields.

Field	Module Property?	Description
HDFS Url	Yes	The web HDFS URL to connect to HDFS.
User Name	Yes	A unique user name to identify each user connecting to HDFS.

Test Connection

Clicking **Test Connection** to test whether the specified configuration fields result in a valid connection to HDFS for design time configuration and runtime execution of the activities.

HCatalog Connection

The HCatalogConnection shared resource contains all the necessary parameters to connect to HCatalog. It can be used by the Hive, Mapreduce, Pig, and WaitForJobCompletion activities.

General

The **General** panel has the following fields.

Field	Module Property?	Description
Package	No	The name of the package where the shared resource is located.

Field	Module Property?	Description
Name	No	The name for the shared resource.
Description	No	A short description for the shared resource.

HCatalogConnection Configuration

The **HCatalogConnection Configuration** pannel has the following fields.

Field	Module Property?	Description
HCatalog Url	Yes	The web HCatalog URL to connect to HCatalog.
User Name	Yes	A unique user name to identify each user connecting to HCatalog.
HDFS Connection	No	An HDFS Connection shared resource. Click ... to select an HDFS Connection.

Test Connection Button

Clicking **Test Connection** to test whether the specified configuration fields result in a valid connection to HCatalog for design time configuration and runtime execution of the activities.

Database and Table

The Database and Table Editor allows you to introspect the databases and tables existing in HCatalog.

Creating or modifying databases and tables are not supported by this plug-in in this release.

Introspect Database

When you click **Introspect Database**, the plug-in introspects all the databases that currently exist in the specified HCatalog. When you select one of them from the database list, the plug-in generates a database with definitions.

Remove

Click **Remove** to remove the selected database or tables from the **Database** list.

Introspect Table

When you click **Introspect Table**, the plug-in introspects all the tables that currently exist in the specified database. When you select one of them from the table list, the plug-in generates a table with definitions.

Generate Schema

This button is enabled for table. Clicking **Generate Schema** allows you to generate the schema of a table. After clicking **Generate Schema**, input the name of the schema file, which is stored in the Schemas folder in your project.

Database

Database is an administrative container for a set of tables.

Configuration

The **Configuration** pannel has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.
Description	No	A short description of the activity.
Database Location	No	Specifies the directory where the database is located.
Database Comment	No	Specifies comments for the database.
Database Properties	No	Specifies properties for the database. A property is associated with a name and a value.

Table

Table provides shared virtual storage for data. It is a shared entity that can be accessed by multiple applications concurrently, each one of these has the same coherent view of the data contained in the table.

Configuration

The **Configuration** pannel has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.
Description	No	A short description of the activity.
Comment	No	Specifies comments for the table.
Location	No	Specifies the directory where the table is located.
Table Columns	No	Specifies columns of the table. A column is associated with a name, a type, an advanceType, and a Comment.  The advanceType column indicates the type of complex data type.

Advanced

The **Advanced** pannel has the following fields.

Field	Module Property?	Description
Partitioned	No	Specifies whether the table is partitioned.
Output Format	No	Specifies the output format of the table.
Owner	No	Specifies the owner of the table.
Input Format	No	Specifies the input format of the table.
Permission	No	Specifies the permission of the table.
Group	No	Specifies the group associated with the table.
Partition Columns	No	Specifies the partition columns of the table. A partition column is associated with a name, a type, and a comment.
Table Properties	No	Specifies the properties of the table. A table property is associated with a name and a value.

Palettes

This plug-in has two palettes, HDFS palette and Hadoop palette, which contain activities that can be added to your business processes.

HDFS Palette

The HDFS palette allows you to do some operations on the files in HDFS.

- HDFSOperation
- ListFileStatus
- Read
- Write

HDFSOperation

The HDFSOperation activity is used to do basic operations on files in HDFS, including copying local files to HDFS, copying files from HDFS to local system, and renaming or deleting files in HDFS.

- Folders cannot be copied between HDFS and local system.
- File names containing space characters in HDFS cannot be operated by all activities.

Genaral

The **Genaral** tab has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.
HDFS Connection	Yes	Click  to select an HDFS Connection shared resource. If no matching HDFS Connection shared resources are found, click **Create Shared Resource** to create one.
HDFSOperation	No	Specifies the HDFS operation. • **PUT_LOCAL_TO_HDFS**: Copy local files to HDFS. • **GET_HDFS_TO_LOCAL**: Copy files from HDFS to the local file system. • **RENAME_HDFS**: Rename files in HDFS. • **DELETE_FROM_HDFS**: Delete files from HDFS.

Description

Provide a short description for the activity.

Input

The input of this activity varies depending on the HDFS operation you chose in the **General** tab.The following table specifies the possible input of the activity.

Input Item	Data Type	Description
sourceFilePath	string	Specifies the path of the source file.
destinationFilePath	string	Specifies the path of the destination file.
overwrite	boolean	Specifies whether to overwrite the existing file if a file that has the same name already exists in the specified destination path, `1` (true) or `0` (false).
blockSize	long	Specifies the block size of the file. The value in this field must be greater than 0.
replication	short	Specifies the number of replications of the file. The value in this field must be greater than 0.
permission	integer	Specifies the permission of the file. The value in this field must be in the range 0 to 777.
offset	long	Specifies the starting byte position. The value in this field must be 0 or greater.
length	long	Specifies the number of bytes to be processed.
bufferSize	integer	Specifies the size of the buffer used in transferring data. The value in this field must be 0 or greater.
recursive	boolean	Specifies whether to operate on the content in the subdirectories, `1` (true) or `0` (false).

Output

The output of the activity is as follows.

Output Item	Data Type	Description
HDFS		
status	integer	Returns standard HTTP status code to indicate whether the execution has succeeded or not.
msg	string	Returns the execution message.

Fault

The **Fault** tab lists the exceptions that can be thrown by this activity.

HDFSException	Description
msg	The error message description returned by the plug-in.

HDFSException	Description
msgCode	The error code returned by the plug-in.
exception	Occurs when the plug-in has internal errors.
message	The error message returned by the server.
javaClassName	The name of the Java Class where an error occurred.

ListFileStatus

The ListFileStatus activity is used to list the status of the specified file or directory.

Regarding the specified directory, the activity returns the status of all files and directories in the specified directory.

General

The **General** tab has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.
HDFS Connection	Yes	Click  to select an HDFS Connection shared resource. If no matching HDFS Connection shared resources are found, click **Create Shared Resource** to create one.

Description

Provide a short description for the activity.

Input

The input of the activity are as follows.

Input Item	Data Type	Description
path	string	Specifies the path of a file or a directory.  If you specify an empty folder in this field, the **Output** tab will empty.

Output

The output of the activity are as follows.

Output Item	Data Type	Description
fileinfo		

Output Item	Data Type	Description
accessTime	long	Returns the access time of the file in milliseconds.
blockSize	long	Returns the block size of the file.
length	long	Returns the number of bytes in a file.
modificationTime	long	Returns the modification time of the file in milliseconds.
replication	long	Returns the number of replications of the file.
owner	string	Returns the user name of the file owner.
type	string	Returns the type of the path object, `FILE` or `DIRECTORY`.
group	string	Returns the group associated with the file.
permission	string	Returns the permission of the file.
pathSuffix	string	Returns the path suffix of the file.

Fault

The **Fault** tab lists the exceptions that can be thrown by this activity.

HDFSException	Description
msg	The error message description returned by the plug-in.
msgCode	The error code returned by the plug-in.
exception	Occurs when the plug-in has internal errors.
message	The error message returned by the server.
javaClassName	The name of the Java Class where an error occurred.

Read

The Read activity is used to read data from a file in HDFS and place its content into the **Output** tab of the activity.

Due to the limitations of TIBCO ActiveMatrix BusinessWorks, this activity cannot read more than 2 GB data at one time. You can use the group to iteratively read the data in a file of more than 2 GB.

General

The **General** tab has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.

Field	Module Property?	Description
HDFS Connection	Yes	Click  to select an HDFS Connection shared resource. If no matching HDFS Connection shared resources are found, click **Create Shared Resource** to create one.
Read As	No	Specifies in what format the file to be read as, `text` or `binary`.

Description

Provide a short description for the activity.

Input

The input of the activity are as follows.

Input Item	Data Type	Description
fileName	string	Specifies the path of the file to be read.
offset	long	Specifies the starting byte position to be read. The value in this field must be greater than 0.
length	long	Specifies the number of bytes to be read.
bufferSize	long	Specifies the size of the buffer used in transferring data. The value in this field must be greater than 0.

Output

The output of the activity varies depending on the file format you chose in the **General** tab .

Output Item		Data Type	Description
fileContent	binary Content	base64 Binary	Returns the file content in binary format. This item is displayed when **binary** is selected in the **Read As** field in the **General** tab.
	textContent	string	Returns the file content in text format. This item is displayed when **text** is selected in the **Read As** field in the **General** tab.
end		boolean	Returns whether the file has been read to the end.

Fault

The **Fault** tab lists the exceptions that can be thrown by this activity.

HDFSException	Description
msg	The error message description returned by the plug-in.
msgCode	The error code returned by the plug-in.
exception	Occurs when the plug-in has internal errors.
message	The error message returned by the server.
javaClassName	The name of the Java Class where an error occurred.

Write

The Write activity is used to write data to a specified file in HDFS.

General

The **General** tab has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.
HDFS Connection	Yes	Click  to select an HDFS Connection shared resource. If no matching HDFS Connection shared resources are found, click **Create Shared Resource** to create one.
Append	No	Specifies whether to append the data to an existing file.  Data cannot be appended to the same file in HDFS by multiple processes simultaneously.
Write Type	No	Specifies the format of the data to be written. • **text** • **binary** • **file** • **StreamObject**
Overwrite	Yes	Specifies whether to overwrite the existing data in the specified file.  This field is displayed when the **Append** check box is cleared.

Description

Provide a short description for the activity.

Input

The input of this activity varies depending on the data format you chose in the **General** tab. The following table specifies the possible input of the activity.

Input Item	Data Type	Description
fileName	string	Specifies the path of the file to be written in.
fileContent	string	Specifies the content to be written in the file.  This field is displayed when **text** is selected in the **Write Type** field in the **General** tab.
binaryData	base64 Binary	Specifies the binary data to be written in the file.  This field is displayed when **binary** is selected in the **Write Type** field in the **General** tab.
sourceFile Path	string	Specifies the path of the file to be written.  This field is displayed when **file** is selected in the **Write Type** field in the **General** tab.
inputStream Object	Object	Specifies the streaming object to be written.  This field is displayed when **StreamObject** is selected in the **Write Type** field in the **General** tab.
overwrite	boolean	Specifies whether to overwrite the existing data in the source file.
blockSize	long	Specifies the block size of the file. The value in this field must be greater than 0.
replication	short	Specifies the number of replications of the file. The value in this field must be greater than 0.
permission	integer	Specifies the permission of the file. The value in this field must be in the range 0 to 777.
buffersize	integer	Specifies the size of the buffer used in transferring data. The value in this field must be greater than 0.

Output

The output of the activity are as follows.

Output Item	Data Type	Description
status	integer	Returns standard HTTP status code to indicate whether the execution has succeeded or not.
msg	string	Returns the execution message.

Fault

The **Fault** tab lists the exceptions that can be thrown by this activity.

HDFSException	Description
msg	The error message description returned by the plug-in.
msgCode	The error code returned by the plug-in.
exception	Occurs when the plug-in has internal errors.
message	The error message returned by the server.
javaClassName	The name of the Java Class where an error occurred.

Hadoop Palette

The Hadoop palette allows you to utilize the benefits of Hive, Mapreduce, and Pig based on Hadoop.

- Hive
- MapReduce
- Pig
- WaitForJobCompletion

Hive

The Hive activity is used to facilitate querying and managing large datasets residing in distributed storage.

- If you run this activity on the Redhat platform, you need to upgrade XML User Interface Language (XUL) Runner to version 1.8 or later. After the upgrading, you need to reinstall Mozilla Firefox.
- This activity do not support to upload Hive data from local clusters.

General

The **General** tab has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.
HCatalog Connection	Yes	Click  to select an HCatalog Connection shared resource. If no matching HCatalog Connection shared resources are found, click **Create Shared Resource** to create one.
IsFileBase	No	Select this check box if Hive scripts are from a file.

Field	Module Property?	Description
Hive Script File	Yes	Specifies the path of the file containing Hive scripts.  This field is displayed when the **IsFileBase** check box is selected.
HiveEditor	No	Specifies Hive scripts. The keywords of the scripts are highlighted automatically.  This field is displayed when the **IsFileBase** check box is cleared.
Define	No	Specifies Hive configuration variables. A variable is associated with a name and a value.
Status Directory	Yes	Specifies the directory where the status of the Hive job is located.
WaitForResult	Yes	Select this check box if you want the process to wait for the Hive operation to complete.  When this check box is selected, the Hive activity does not support to query more than 2 GB result data at one time due to the limitations of TIBCO ActiveMatrix BusinessWorks.

Description

Provide a short description for the activity.

Input

The values specified in this tab takes precedence over the ones in the corresponding fields in the **General** tab.

Input Item	Data Type	Description
HiveFile	string	Specifies path of the HDFS file that contains commands.  This item is displayed when the **IsFileBase** check box is selected.
HiveScript	string	Specifies Hive scripts directly.  This item is displayed when the **IsFileBase** check box is cleared.
Defines	string	Specifies Hive configuration variable. Each variable is associated with a name and a value.
Status Directory	string	Specifies the directory where the status of the Hive job is located.

Output

The output of the activity are as follows.

Output Item	Data Type	Description
jobId	string	Returns the job ID of the Hive operation.  • This item is displayed when the **WaitForResult** check box is cleared. • You can use the WaitForJobCompletion activity to wait for the job to complete. The exitValue in the **Output** tab of the WaitForJobCompletion activity shows the exit value of Hive SQL execution.
content	string	Returns the result of the job.  This item is displayed when the **WaitForResult** check box is selected

Fault

The **Fault** tab lists the exceptions that can be thrown by this activity.

HDFSException	Description
msg	The error message description returned by the plug-in.
msgCode	The error code returned by the plug-in.

MapReduce

The Mapreduce activity is used to create and queue a standard Mapreduce job or a streaming Mapreduce job.

General

The **General** tab has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.
HCatalog Connection	Yes	Click  to select an HCatalog Connection shared resource. If no matching HCatalog Connection shared resources are found, click **Create Shared Resource** to create one.
Streaming	No	Select this check box to create and run Mapreduce streaming jobs.

Field	Module Property?	Description
The following four fields are displayed when the **Streaming** check box is selected.		
Input	Yes	Specifies the path of the input data in Hadoop.
Output	Yes	Specifies the path of the output data.
Mapper	Yes	Specifies the path of the mapper program in Hadoop.
Reducer	Yes	Specifies the path of the reducer program in Hadoop.
The following four fields are displayed when the **Streaming** check box is cleared.		
Jar Name	Yes	Specifies the name of the .jar file for Mapreduce to use.
Main Class	Yes	Specifies the name of the class for Mapreduce to use.
Lib Jars	Yes	Specifies the comma separated .jar file to be included in the classpath.
Files	Yes	Specifies the comma separated .jar files to be copied to the Mapreduce cluster.
Status Directory	Yes	Specifies the directory where the status of Mapreduce jobs are stored.

Arguments	No	Specifies the program arguments.
		• If the **Streaming** check box is cleared, specify Java main class arguments.
		• If the **Streaming** check box is selected, specify a list of arguments that contain space-separated strings to pass to the Hadoop streaming utility.
		For example,
		`- files /user/hdfs/file`
		`- D mapred.reduce.task=0`
		`- input format`
		`org.apache.hadoop.mapred.lib.NLineInputFormat`
		`- cmdenv info=wc-reducer`

The following field is displayed when the **Streaming** check box is cleared.		
Define	No	Specifies the Hadoop configuration variables. A variable is associated with a name and a value.

Description

Provide a short description for the activity.

Input

The values specified in this tab takes precedence over the ones in the corresponding fields in the **General** tab. The following table specifies the possible input of the activity.

Input Item	Data Type	Description
The following four fields are displayed when the **Streaming** check box is selected.		
Input	string	Specifies the path of the input data in Hadoop.
Output	string	Specifies the path of the output data.
Mapper	string	Specifies the path of the mapper program in Hadoop.
Reducer	string	Specifies the path of the reducer program in Hadoop.
The following four fields are displayed when the **Streaming** check box is cleared.		
JarName	string	Specifies the name of the .jar file for Mapreduce to use.
ClassName	string	Specifies the name of the class for Mapreduce to use.
Libjars	string	Specifies the comma separated .jar file to be included in the classpath.
Files	string	Specifies the comma separated .jar files to be copied to the Mapreduce cluster.
StatusDirectory	string	Specifies the directory where the status of Mapreduce jobs are stored.
Arguments	string	Specifies the program arguments.
The following field is displayed when the **Streaming** check box is cleared.		
Defines	string	Specifies the Hadoop configuration variables. A variable is associated with a name and a value.

Output

The output of the activity are as follows.

Output Item	Data Type	Description
jobId	string	Returns the job ID of the Mapreduce operation.  You can use the WaitForJobCompletion activity to wait for the job to complete. The `exitValue` in the **Output** tab of the WaitForJobCompletion activity shows the exit value of Mapreduce execution.

Fault

The **Fault** tab lists the exceptions that can be thrown by this activity.

HadoopException	Description
msg	The error message description returned by the plug-in.
msgCode	The error code returned by the plug-in.

Pig

The Pig activity is used to create and queue a Pig job.

If you run this activity on the Redhat platform, you need to upgrade XML User Interface Language (XUL) Runner to version 1.8 or later. After the upgrading, you need to reinstall Mozilla Firefox.

General

The **General** tab has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.
HCatalog Connection	Yes	Click  to select an HCatalog Connection shared resource. If no matching HCatalog Connection shared resources are found, click **Create Shared Resource** to create one.
IsFileBase	No	Select this check box if Pig scripts are from a file.
Pig File	Yes	Specifies the path of the file contains Pig scripts.  This field is displayed when the **IsFileBase** check box is selected.
PigEditor	No	Specifies Pig scripts. The keywords of the scripts are highlighted automatically.  This field is displayed when the **IsFileBase** check box is cleared.
Arguments	No	Specifies Pig arguments containing space-separated string.
Status Directory	Yes	Specifies the directory where the status of the Pig job is located.
Files	Yes	Specifies the comma separated files to be copied to the Mapreduce cluster.

Description

Provide a short description for the activity.

UDF

The **UDF** tab has the following fields.

Input Item	Module Property?	Description
UDF Directory	Yes	Specifies the directory where user define functions (UDF) is located.
Available UDF Files	No	Specifies the UDF file to be applied. Click to Sync list available UDF files under the specified UDF directory. Select a UDF file and click Register, the UDF file is displayed in the **Pig Editor** field in the **General** tab.
Upload UDF File	Yes	Specifies the UDF file to be uploaded. Click to select the UDF file to be uploaded, and then click **Upload** to upload the file to the specified directory.

Input

The values specified in this tab takes precedence over the ones in the corresponding fields in the **General** tab.

Input Item	Data Type	Description
PigScript	string	Specifies Pig scripts. This item is displayed when the **IsFileBase** check box is cleared.
PigFile	string	Specifies the comma separated files to be copied to the Mapreduce cluster. This item is displayed when the **IsFileBase** check box is selected.
Arguments	string	Specifies Pig arguments.
StatusDirectory	string	Specifies the directory where the status of the Hive job is located.
Files	string	Specifies the comma separated files to be copied to the Mapreduce cluster.

Output

The output of the activity are as follows.

Output Item	Data Type	Description
jobId	string	Returns the job ID of the Pig operation.  You can use the WaitForJobCompletion activity to wait for the job to complete. The `exitValue` in the **Output** tab of the WaitForJobCompletion activity shows the exit value of Pig scripts execution.

Fault

The **Fault** tab lists the exceptions that can be thrown by this activity.

HadoopException	Description
msg	The error message description returned by the plug-in.
msgCode	The error code returned by the plug-in.

WaitForJobCompletion

The WaitForJobCompletion activity is used to wait for the specified jobs to complete until it reaches the specified value of timeout.

General

The **General** tab has the following fields.

Field	Module Property?	Description
Name	No	The name of the activity in the process definition.
HCatalog Connection	Yes	Click  to select an HCatalog Connection shared resource. If no matching HCatalog Connection shared resources are found, click **Create Shared Resource** to create one.
Timeout	Yes	Specifies how long to wait for a job to complete (in milliseconds). The default value is 1200000 milliseconds.
Interval	Yes	Specifies the time interval in milliseconds, to check the job execution. The default value is 5000 milliseconds.

Description

Provide a short description for the activity.

Input

The values specified in this tab takes precedence over the ones in the corresponding fields in the **General** tab.

Input Item	Data Type	Description
jobId	string	Specifies the ID of the job to be waited for.
timeout	long	Specifies how long to wait for a job to complete (in milliseconds).
interval	long	Specifies the time interval in milliseconds to check the job execution.

Output

The output of the activity are as follows.

Output Item		Data Type	Description
Job			
status	startTime	string	Returns the start time of the job.
	username	string	Returns the user name of the job owner.
	jobID	string	Returns the ID of the job.
	jobACLs	string	Returns ACLs for the job.
	schedulingInfo	string	Returns the scheduling information associated with the job.
	failureInfo	string	Returns the reason for the failure of the job.
	jobId	string	Returns the ID of the job.
	jobPriority	string	Returns the priority of the job.
	runState	string	Returns the current running state of the job.
	State	string	Returns the state of the job.
	jobComplete	string	Returns the completion status of the job.
profile	url	string	Returns the URL of the job.
	jobID	string	Returns the ID of the job.
	user	string	Returns the user name of the job owner.
	queueName	string	Returns the queue name of Mapreduce.
	jobFile	string	Returns the name of the job file.
	jobName	string	Returns the name of the job.
	jobId	string	Returns the ID of the job.

Output Item	Data Type	Description
id	string	Returns the ID of the job.
parentId	string	Returns the parent ID of the job.
percentComplete	string	Returns the completion percentage of the job.
exitValue	string	Returns the exit value of the job. If a value of 0 is returned, it indicates that the job has been successful. Otherwise, the job completed with errors.  You can get error messages of the job from the `StatusDirectory\stderr` directory. `StatusDirectory` is the value specified in the Status Directory field in the **General** tab of the activity that executed the job. The ErrorHandler process in the sample project shipped with the installer shows how to get error messages of a job. See Working with Sample Project for more information.
user	string	Returns the user name of the job owner.
callback	string	Returns the callback URL.
completed	string	Returns the completion status of the job. For example, `done`.

Fault

The **Fault** tab lists the exceptions that can be thrown by this activity.

ActivityFaultData	Description
HadoopException	
msg	The error message description returned by the plug-in.
msgCode	The error code returned by the plug-in.
ActivityTimeoutException	
msg	The error message description returned by TIBCO ActiveMatrix BusinessWorks.
msgCode	The error code returned by TIBCO ActiveMatrix BusinessWorks.

Sample Project

Working through the sample project helps you understand how TIBCO ActiveMatrix BusinessWorks Plug-in for Big Data works.

TIBCO ActiveMatrix BusinessWorks Plug-in for Big Data packages one sample project and one sample data file within the installer, which are located in the *TIBCO_HOME*\bw\palettes\bigdata\6.0\Samples directory:

- The DataCollectionAndQueryProj sample project shows how to use TIBCO ActiveMatrix BusinessWorks Plug-in for Big Data to manage and query data, for example, how to operate files in HDFS, transfer files between HDFS and local directory, create Hive Table, and query data. See Working with Sample Project.
- The SampleData folder contains the customers.csv sample file. It can be used as a sample object of operation.

Importing Sample Project

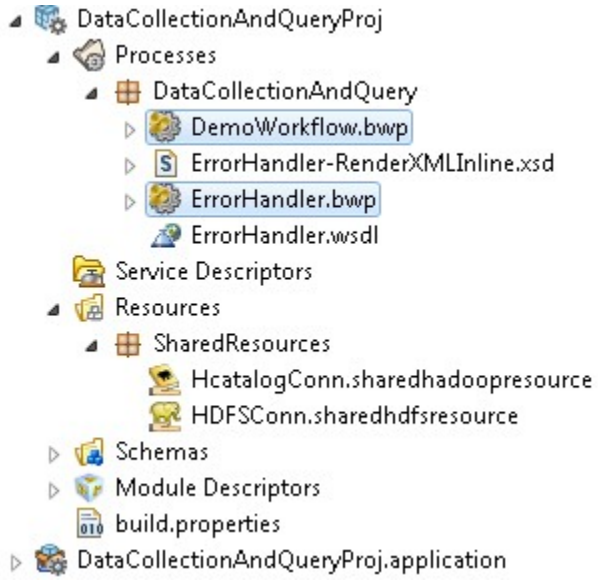
Only one sample project is packaged with the plug-in. Before running the project, you need to import the project to TIBCO Business Studio.

Procedure

1. Navigate to the *TIBCO_HOME*\bw\palettes\bigdata\6.0\Samples directory and unzip the sample.zip file to the directory.
2. Start TIBCO Business Studio.
3. Click **File > Import**.
4. In the **Import** dialog, expand the **General** folder and select **Existing Studio Projects into Workspace**. Click **Next**.
5. Click **Browse** next to the **Select root directory** field to locate the samples. Click **Finish**.
   The sample project is located in the *TIBCO_HOME*\bw\palettes\bigdata\6.0\Samples directory.

Result

The sample project is imported to TIBCO Business Studio.



Configuring Module Properties

You need to configure the module properties used in the sample project.

Prerequisites

Before configuring the shared resources, you need to import the sample project to TIBCO Business Studio. See Importing Sample Project.

Procedure

1. Expand **DataCollectionAndQueryProj > Module Descriptors** and double-click **Module Properties**.
2. Expand **Example** and configure the properties used in the shared resources.
   - `HCATALOG_URL=`*localhost_IP*
   - `HDFS_URL=`*localhost_IP*
   - `Project_Path=`*TIBCO_HOME*\bw\palettes\bigdata\6.0\Samples

Working with Sample Project

This sample process shows how to manage and query data. For example, how to operate files in HDFS, transfer files between HDFS and local file system, create Hive table, and query data.

Prerequisites

Before running the project, you need to import the sample project to TIBCO Business Studio and configure the shared resource in the sample. See Importing Sample Project and Configuring Module Properties.

Procedure

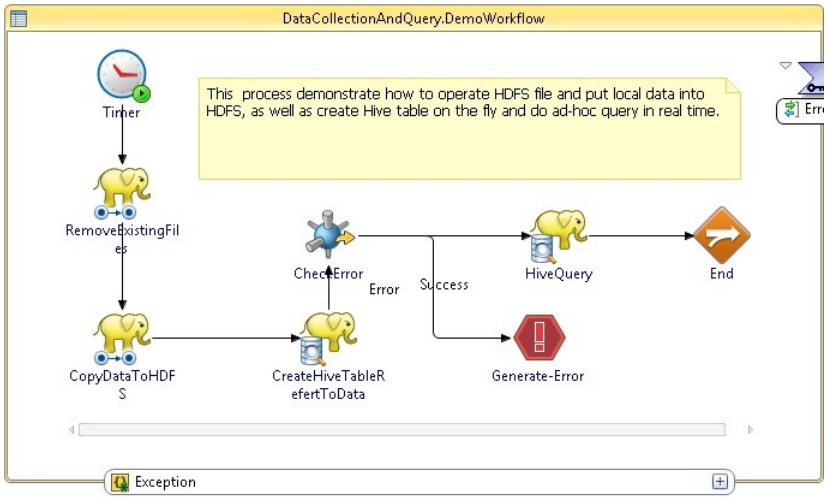
1. Expand the DataCollectionAndQueryProj project in Project Explorer.
2. Click **Processes > DataCollectionAndQuery > DemoWorkflow.bwp**.
3. Expand **Module Descriptors** and double-click **Components**.
4. Ensure only the DataCollectionAndQuery.DemoWorkflow component is selected in the Component Configurations area.
5. Click **Run > Run Configurations**.
6. Expand **BusinessWorks Application** and click **BWApplication**.
7. Ensure only the DataCollectionAndQueryProj.application is selected in the **Applications** tab.
8. Click **Run**.
9. Click **Terminate**  to stop the process.

Result

If the process performs successfully, 100 records in the `customers.csv` file created by the CreateHiveTableRefertToData activity are returned in the Output tab of the HiveQuery activity.

Configurations for DemoWorkflow Process

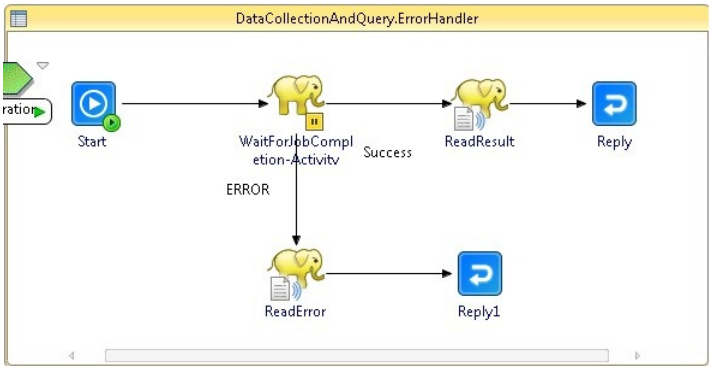
The DemoWorkflow process contains the following activities.



Activity	Description
Timer	Starts the process when the specified time interval expires.
RemoveExistingFiles	Deletes the `customers.csv` file from the `user\hdfs\bwdemo` directory in HDFS.
CopyDataToHDFS	Copies the `customers.csv` file from the *`TIBCO_HOME`*`\bw\palettes\bigdata\6.0\Samples\samples.zip\SampleData` directory to the `user\hdfs\bwdemo` directory in HDFS.
CreateHiveTableReferToData	Creates an external Hive Table named `customers` in HDFS.
CheckError	Performs the subprocess. If no error occurs during the `ErrorHandler` process works, the process goes to the HiveQuery activity. If any error occurs, the process goes to the Generate Error activity.
HiveQuery	Queries 100 records from the `customers.csv` Hive table created by the CreateHiveTableReferToData activity.
Generate Error	Generates error messages from the **Output** tab of the `ErrorHandler` subprocess.
Catch	Catches error messages and displays in the **Output** tab when any error occurs in the `DemoWorkflow` process.
End	Ends the process.

Configurations for ErrorHandler Subprocess

The ErrorHandler subprocess contains the following activities.



Activity	Description
Start	Starts the process.
WaitForJobCompletion Activity	Waits for the CreateHiveTableRefertToData activity to complete the job.
ReadResult	Reads the output of the job executed by the CreateHiveTableRefertToData activity, which is in the `user\hdfs\hive\createstatus\stdout` directory.
ReadError	Reads error messages of the job executed by the CreateHiveTableRefertToData activity, which is in the`user\hdfs\hive\createstatus\stderr` directory.
Generate Error	Generates error messages from the **Output** tab of the ReadError activity.
Catch	Catches the error messages and displays in the **Output** tab when any error occurs in the `ErrorHandler` subprocess.
Render XML	Renders the error messages in XML string.
End	Ends the process.

Managing Logs

Logs are used to trace and troubleshoot exceptions.

You can configure the `logback.xml` file which is located in the *TIBCO_HOME*`\bw\6.1\config\design\logback` directory to manage logs:

- Managing Logs in Console.
- Exporting Logs to a File.

Log Levels

The plug-in can capture logs at different levels.

Log Level	Description
Info	Indicates normal plug-in operations. No action is needed. A tracing message tagged with Info indicates that a significant processing step is reached and logged for tracking or auditing purposes. Only info messages preceding a tracking identifier are considered as significant steps.
Error	An unrecoverable error occurs. Depending on the error severity, the plug-in may continue with the next operation or may stop altogether.
Debug	A developer-defined tracing message.

Managing Logs in Console

The plug-in logs are written to Console by default.

If neither the plug-in log or the BusinessWorks log is configured in the `logback.xml` file, the error logs of the plug-in will be displayed in Console.

If the plug-in log is not configured but the BusinessWorks log is configured in the `logback.xml` file, the configuration for BusinessWorks log is implemented by the plug-in.

Procedure

1. Navigate to the *TIBCO_HOME*`\bw\6.1\config\design\logback` directory and open the `logback.xml` file.

2. Add the following nodes in the Console Appender area to specify the plug-in log.

```
<logger name="com.tibco.bw.palette.hadoop">
    <level value="DEBUG"/>
  </logger>
```

```
<logger name="com.tibco.bw.palette.webhdfs">
    <level value="DEBUG"/>
  </logger>
```

```
<logger name="com.tibco.bw.sharedresource.webhdfs">
    <level value="DEBUG"/>
  </logger>
```

```
<logger name="com.tibco.bw.sharedresource.hadoop">
    <level value="DEBUG"/>
  </logger>
```

The `level` tag that defines the log level and the value can be `Error` or `Debug`.

If you want to check the activity output in Console, you need to set `level` to `Debug`.

3. Save the file.

Exporting Logs to a File

You can export logs to a file by modifying the `logback.xml` file.

Procedure

1. Navigate to the *`TIBCO_HOME`*`\bw\6.1\config\design\logback` directory and open the `logback.xml` file.

When deploying an application in TIBCO Enterprise Administrator, you need to navigated to the *`TIBCO_HOME`*`\bw\domains\mydomain\appnodes\myspace\mynode` directory to find the `logback.xml` file.

2. Add the following node to specify the file location.

```
<appender name="FILE" class="ch.qos.logback.core.FileAppender">
   <file>c:/bw6-bigdata.log</file>
      <encoder>
        <pattern>%d{HH:mm:ss.SSS} [%thread] %-5level %logger{36}-%msg%n</pattern>
      </encoder>
</appender>
```

The `file` tag defines the location to which the log is exported and the value is the absolute path of the file that is detailed to the file name.

3. Add the following node to the `root` node at the bottom of the `logback.xml` file to enable exporting logs to a file.

```
    <appender-ref ref="FILE" />
<root level="DEBUG">
   <appender-ref ref="STDOUT" />
   <appender-ref ref="FILE" />
</root>
```

4. Save the file.

Error Messages

The error messages for each shared resource and activity are listed with corresponding description and resolution.

Error Messages for HDFS

Error Code	Role	Messages	Resolution
TIBCO-BW-PALETTE-HDFS-2000000	debug Role	{0}	No action.
TIBCO-BW-PALETTE-HDFS-2000001	debug Role	Start execute: {0}	No action.
TIBCO-BW-PALETTE-HDFS-2000002	debug Role	End execute: {0}	No action.
TIBCO-BW-PALETTE-HDFS-2000003	debug Role	Customer thread start, thread ID: {0}	No action.
TIBCO-BW-PALETTE-HDFS-2000004	debug Role	Input data: {1}	No action.
TIBCO-BW-PALETTE-HDFS-2000005	debug Role	Output data: {1}	No action.
TIBCO-BW-PALETTE-HDFS-2000006	debug Role	{0} : {1}	No action.
TIBCO-BW-PALETTE-HDFS-2000007	debug Role	Cached tasks:{0}	No action.
TIBCO-BW-PALETTE-HDFS-5000001	error Role	Please specify HDFS connection shared resource.	Specify the HDFSConnection shared resource.
TIBCO-BW-PALETTE-HDFS-5000002	error Role	Remote Error while run {0}	No action.
TIBCO-BW-PALETTE-HDFS-5000003	error Role	Error Occurred: {0}	No action.

Error Code	Role	Messages	Resolution
TIBCO-BW-PALETTE-HDFS-5000004	error Role	Invalid hdfs operation: {0}.	No action.
TIBCO-BW-PALETTE-HDFS-5000005	error Role	Invalid HDFS URL, Please specify a correct one.	Specify correct HDFS URL.
TIBCO-BW-PALETTE-HDFS-5000006	error Role	{0} is not a valid buffer size, valid buffer size should > 0	Specify a value greater than 0.
TIBCO-BW-PALETTE-HDFS-5000007	error Role	{0} is not a valid replication, valid replication should > 0	Specify a value greater than 0.
TIBCO-BW-PALETTE-HDFS-5000008	error Role	{0} is not a valid offset, valid offset should >= 0	Specify a value greater than or equal to 0.
TIBCO-BW-PALETTE-HDFS-5000009	error Role	{0} is not a length, valid length should >= 0	Specify a value greater than or equal to 0.
TIBCO-BW-PALETTE-HDFS-5000010	error Role	{0} is not a valid block size, valid block size should > 0	Specify a value greater than 0.
TIBCO-BW-PALETTE-HDFS-5000011	error Role	{0} is not a valid permission, valid values should 0 - 777	Specify a value in the range 0 - 777.
TIBCO-BW-PALETTE-HDFS-5000012	error Role	Unknow result returned: {0}	No action.
TIBCO-BW-PALETTE-HDFS-5000013	error Role	IOException occurred while retrieving XML Output for activity [{0}].	No action.

Error Messages for Hadoop

Error Code	Role	Messages	Resolution
TIBCO-BW-PALETTE-HADOOP-2000000	debug Role	{0}	No action.
TIBCO-BW-PALETTE-HADOOP-2000001	debug Role	Start execute: {0}	No action.

Error Code	Role	Messages	Resolution
TIBCO-BW-PALETTE-HADOOP-2000002	debug Role	End execute: {0}	No action.
TIBCO-BW-PALETTE-HADOOP-2000003	debug Role	Customer thread start, thread ID: {0}	No action.
TIBCO-BW-PALETTE-HADOOP-2000004	debug Role	Input data: {1}	No action.
TIBCO-BW-PALETTE-HADOOP-2000005	debug Role	Output data: {1}	No action.
TIBCO-BW-PALETTE-HADOOP-2000006	debug Role	Define name field empty so ignore it, Value is: {0}	No action.
TIBCO-BW-PALETTE-HADOOP-2000007	debug Role	{0} : {1}	
TIBCO-BW-PALETTE-HADOOP-2000008	debug Role	Cached tasks:{0}	No action.
TIBCO-BW-PALETTE-HADOOP-3000000	info Role	{0}	No action.
TIBCO-BW-PALETTE-HADOOP-3000001	info Role	{0} is done({1}ms)	No action.
TIBCO-BW-PALETTE-HADOOP-5000000	error Role	Please specify HCatalog connection shared resource.	Specify the HCatalogConnection shared resource.
TIBCO-BW-PALETTE-HADOOP-5000001	error Role	Error Occurred: {0}	No action.
TIBCO-BW-PALETTE-HADOOP-5000002	error Role	Invalid {0} URL, Please specify a correct one.	Specify a correct one.
TIBCO-BW-PALETTE-HADOOP-5000003	error Role	Unknow result returned: {0}	No action.
TIBCO-BW-PALETTE-HADOOP-5000004	error Role	IOException occurred while retrieving XML Output for activity [{0}].	No action.
TIBCO-BW-PALETTE-HADOOP-5000005	error Role	Remote Error while run {0}	No action.

Error Code	Role	Messages	Resolution
TIBCO-BW-PALETTE-HADOOP-5000006	error Role	Hive error occurred: {0}	No action.
TIBCO-BW-PALETTE-HADOOP-5000007	error Role	HDFS URL is empty, please specify one.	Specify an HDFS URL.
TIBCO-BW-PALETTE-HADOOP-5000008	error Role	Status directory is empty, please specify one.	Specify a status directory.
TIBCO-BW-PALETTE-HADOOP-5000009	error Role	Hive script cannot be null.	Specify Hive script.
TIBCO-BW-PALETTE-HADOOP-5000010	error Role	Hive file script cannot be null, please select one script file.	Specify the file contains Hive scripts.
TIBCO-BW-PALETTE-HADOOP-5000011	error Role	Pig script cannot be null.	Specify Pig script.
TIBCO-BW-PALETTE-HADOOP-5000012	error Role	Pig file script cannot be null, please select one script file.	Specify the file contains Pig scripts.
TIBCO-BW-PALETTE-HADOOP-5000013	error Role	Mapreduce jar name is required, please provide one.	Specify the Jar name in the Jar Name field.
TIBCO-BW-PALETTE-HADOOP-5000014	error Role	Mapreduce main class is required, please provide one.	Specify the main class in the Main Class field.
TIBCO-BW-PALETTE-HADOOP-5000015	error Role	Mapreduce streaming input is required, please provide one.	Specify the input in the Input field.
TIBCO-BW-PALETTE-HADOOP-5000016	error Role	Mapreduce streaming mapper is required, please provide one.	Specify the mapper in the Mapper field.
TIBCO-BW-PALETTE-HADOOP-5000017	error Role	Mapreduce streaming reducer is required, please provide one.	Specify the reducer in the Reducer field.