White Paper

PWX CONNECTORS FOR INFORMATICA FOR EMC GREENPLUM DATABASES

Abstract

This white paper explains how the EMC® Greenplum® PowerExchange (PWX) connector is used in conjunction with the Informatica Workflow Manager to create tasks that leverage the bulk load capability of the Greenplum database. It explains the uses, configuration, setup procedure, and best practices for the PWX connectors.

April 2011

Part Number h8226

Table of Contents

Executive summary

Informatica PowerCenter is a popular extract-transform-load (ETL) tool for EMC® Greenplum® database customers. By using the Informatica Workflow Manager, customers can easily create tasks that load the source data into the target Greenplum database. Greenplum also created a PowerExchange (PWX) connector for Informatica. This PWX connector facilitates the database loading process and integrates the Informatica PowerCenter Workflow Manager with the Greenplum database. This leveraging of the bulk loading capabilities of the Greenplum `gpload` utility results in better data loading performance.

Audience

This white paper is intended for EMC field-facing employees such as sales, technical consultants, support, and customers who will be using the Greenplum PWX connector for Informatica in their daily work. It documents the PWX connector's capabilities, and shows the readers how it can be used in conjunction with Informatica's Workflow Manager in creating a work task. The reader is not expected to have any prior knowledge of the Greenplum PWX connector but should be familiar with how the Informatica Workflow Manager operates.

This is not an installation guide.

Organization of this paper

This paper covers the following topics:

- The PWX connector
- Supported versions of Informatica
- Installation and registration of the PWX connector
- Use of the PWX connector
- Best practices
- Registration

The PWX connector

Informatica is one of the most popular data integration tools in the market. Its products include the PowerCenter and PowerExchange family of products.

Customers use PowerCenter to design workflow tasks that can be used to extract input data from files or databases, and then loaded into Greenplum databases. PowerCenter also has the capability to run transformations on the input data while massaging the data into the desired input format for the target databases.

Informatica PowerExchange is a family of data access tools that help customers access, load, and deliver data, as part of the ETL process. It has customized graphical user interfaces that help customers access native data types and special

features and capabilities. For example, PowerExchange supports source and target data from Adabas, DB2, Informix, POP and IMAP email formats, and many others.

The PWX connector for Greenplum is a special adapter written by EMC Greenplum. The PWX connector fills a gap in the PowerExchange support line, specifically by supporting the Greenplum database as the target database. The connector blends in seamlessly with the Workflow Manager in PowerCenter as a target writer. On the Mapping tab of the Edit Tasks frame, you can click on the target's properties, and select "Greenplum Writer" as the default writer. The other choices may be "File Writer" and "Relational Writer."

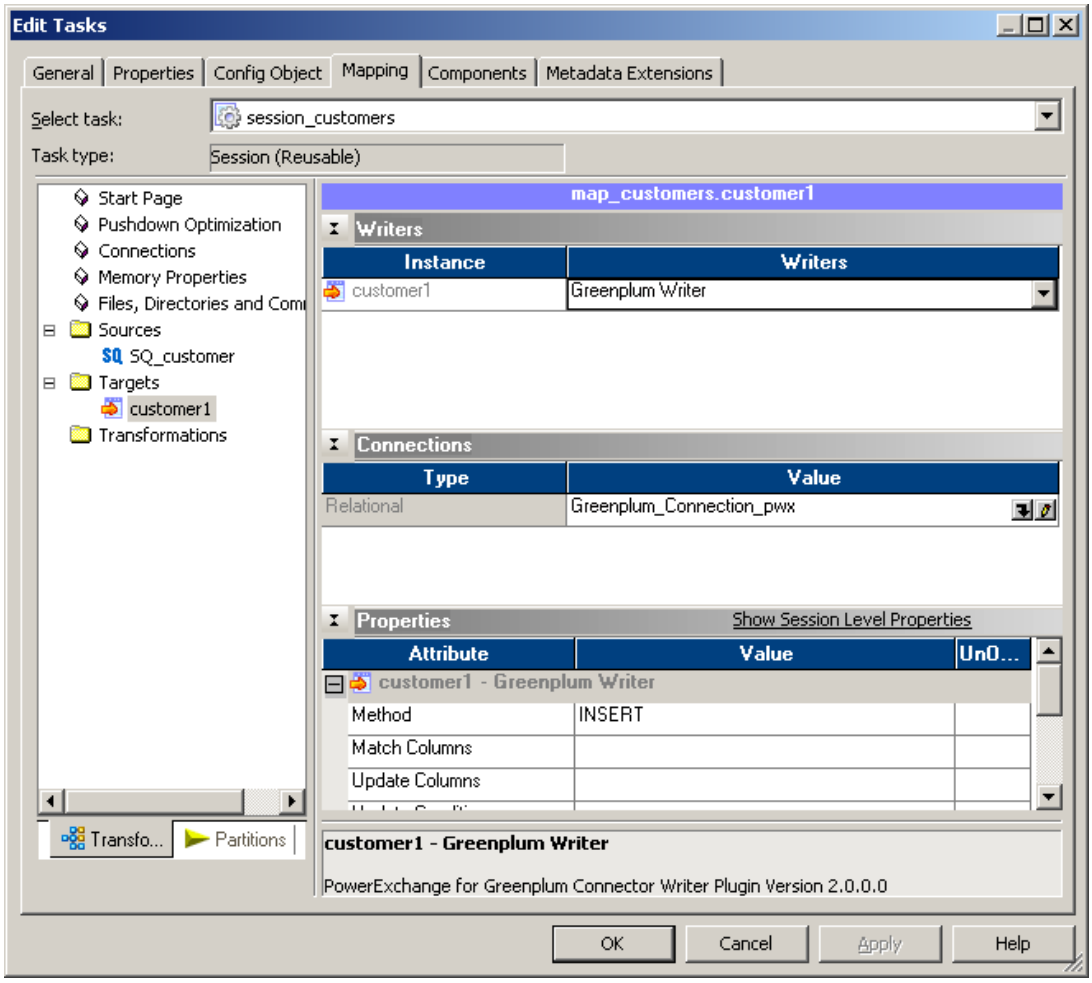

Figure 1. Workflow Manager target mapping

By using the PWX connector, customers receive another benefit. The Greenplum `gpload` utility is used internally for bulk load of data into the database directly from Informatica's PowerCenter. The `gpload` utility is an interface to the Greenplum external table parallel-loading feature.

By using a load specification control file, `gpload` executes data loading by:

- Invoking the Greenplum parallel file server program `gpfdist`
- Creating an external table definition based on the source data definition
- Executing the SQL commands to load the data

Supported versions

The PWX connector for Greenplum is supported on many versions across various operating systems:

Table 1. Informatica PWX connector versions for Greenplum

	PowerCenter 8.1.1	PowerCenter 8.6.1
GPDB 3.2	PWX 1.1.0.0	PWX 1.1.0.0
GPDB 3.3.0 – 3.3.6	PWX 1.2.0.x	PWX 1.2.0.x
GPDB 3.3.7		PWX 1.2.1.0
GPDB 4.0		PWX 2.0.0.0
GPDB 4.1		*

* GPDB 4.1 is generally available, but testing has been completed on only 64-bit Windows 2003 servers and 64-bit Red Hat Linux servers at the time of publication. Testing on GPDB 4.1 with PowerCenter 9.0.1 is in progress.

Table 2. PowerCenter versions per operating system

Operating Systems	PowerCenter Versions
Windows x86 32-bit	8.11, 8.5.1,8.6,8.6.1
Windows x86 64-bit	8.6.1
Red Hat Linux 4 x86 32-bit	8.1.1, 8.5.1, 8.6, 8.6.1
Solaris Sparc 10 32-bit	8.1.1, 8.5.1, 8.6, 8.6.1
Solaris Space 10 64-bit	8.1.1, 8.5.1, 8.6, 8.6.1

Note: You can use a 32-bit version of the PWX connector if you run a 32-bit version of PowerCenter on a 64-bit platform.

Installing the PWX connector

To use the PWX connector, you must install the following Greenplum software packages for the specific platform where the PowerCenter software is running:

Table 3. Greenplum software packages

Greenplum software	Purpose
greenplum-loader-‹version›‹platform›.bin	Installs the `gpload` and `gpfdist` utilities
greenplum-connectivity-‹version›‹platform›.bin	Installs the ODBC driver
greenplum-pwx-‹version›‹platform›.bin	Installs PWX for Greenplum

Note: The Greenplum load and Greenplum connectivity packages are freely available for download from the Greenplum Network website (http://gpn.greenplum.com), while the Greenplum PWX package is available to EMC customers only, and available for download from an internal website. EMC will be transitioning all Greenplum articles to EMC Powerlink® in the near future.

The Greenplum loader package installs the `gpload` and `gpfdist` utilities that are used to bulk load data into the Greenplum database. The connectivity package is used to import the target table schema from the Greenplum database. Importing the schema of the desired table into the Greenplum database target table is necessary in order to create the mapping between the source data and the target Greenplum database.

Using the PWX connector in Informatica Workflow Manager

Using Informatica PowerCenter as an ETL tool involves the following steps:

- Use the PowerCenter Repository Manager to add a repository and folder.
- Use the PowerCenter Designer to define the source and target data, and the mapping and transformations in between.
- Use the Workflow Manager to define tasks and workflow, and to run the tasks.
- Use the Workflow Monitor to monitor the progress of the tasks.

In the Workflow Manager, you define where the data that was extracted is to be loaded in the Greenplum database. As shown in Figure 1 on page 5, the usual method is to use the "Relational Writer." Greenplum has presented our customers with an alternative: the Greenplum Writer.

To use the Greenplum Writer, open the properties of the task by double-clicking the task icon. On the Mapping tab, click the target icon to expose the Writers drop-down list. Selecting Greenplum Writer will load the PWX connector into the task.

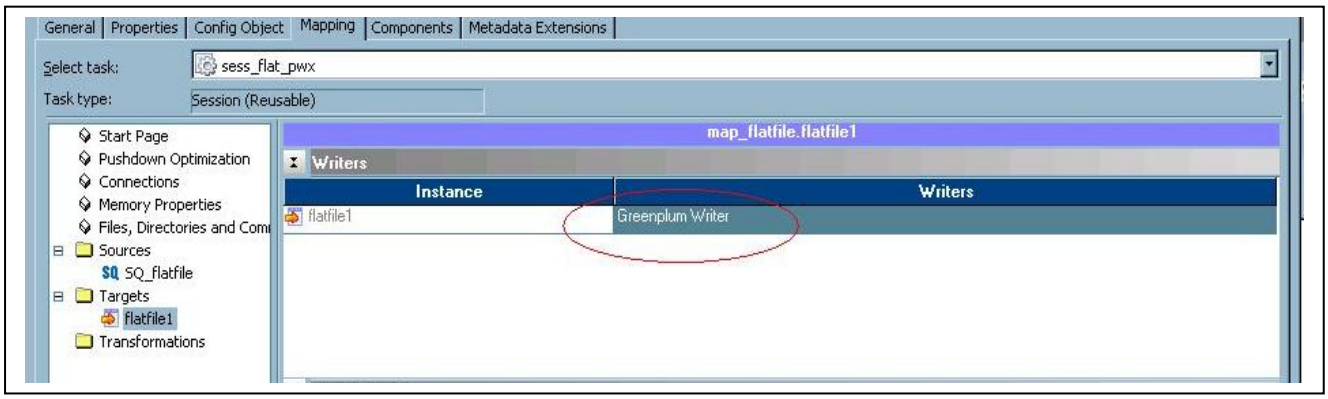

Figure 2. Using the Greenplum Writer

When Informatica Workflow Manager starts a session or task, it performs the following sequence of actions:

- The session is initialized.
- The repository is opened, integration service is contacted, and a folder is opened.
- The workflow is opened, and run ID is issued.
- The mapping is opened. At this point, PowerCenter has all the information it needs to start the session.
- The parallel pipeline engine is started.
- The reader starts reading the source data. At this point, the initialization task is completed.
- The target writer is initialized. If a relational writer is selected, it will be started and the target database will be contacted. If a PWX connecter is used, a control file (YAML file) will be created, and `gpload` will start. The `gpload` is a data-loading utility that acts as an interface to the external table parallel-loading feature.
- The `gpload` utility reads the source data and creates external tables.
- The gpload now calls `gpfdist` (Greenplum file distribution program) to load the data into the Greenplum database on the least used segment servers, then balances the data loading as evenly between the segment servers as possible.

Best practices and tips

- When you run PowerCenter at the client computer, `gpload` is invoked. The `gpload` utility calls up `gpfdist` to send rows of records to the segment servers. It is therefore important to have the network set up correctly in order to facilitate data communications between the load server, the integration service server, and the segment servers.

  For example, the segment servers must be able to reach the client server and the integration server. You can do this through DNS setup or through IP addresses set up in /etc/hosts (or its Windows equivalent). If you are using DCA, then on the DCA master you can use `gpftp` or `gpssh` to set up all segment servers all at once.

- Use character datatypes (char, vchar) whenever possible. Character datatypes are more efficient to process than other datatypes.

  For instance, if you are reading an input field of numeric type, and are not using it as a numeric datatype in the target data field, then it is more efficient to use a character datatype field as the target field. For example, if the input field is zip codes (all numerals), but you are really just storing it as is and not really using the field as a numeral, then make the target field a character datatype field.

- For debugging PWX errors, the session log is your best friend. To get the log, go to the Workflow Monitor. On the task list pane, right-click on the task, and select session log. Look for the entries that say [ERROR]. You will probably see errors such as:
  - "Short write error" – This error seems to be a catch-all. Read further down and solve the next error to remove this error.
  - "Unable to connect to the database" – Check the username and password specified in the ODBC manager, and also in the Connection in the Workflow Manager. Verify that the hostname or the host's IP address is correct, and that you can ping it from all the hosts.
  - "No privilege to create external tables" – The user specified in the Greenplum connection does not have sufficient privileges to create the external tables. Log in to the database and use "alter role" to grant the user sufficient privileges (for example, the createdb privilege.)
  - "Source file cannot be found" – This is not a PWX error. If the source file is not in the default directory, PowerCenter expects it to be in the SrcFiles directory under server › infa_shared directory of the Informatica installation directory of the integration server. Verify that the source file has been stored there.
  - "Unable to create the `gpload` process" – Python is not installed, or a wrong version of Python was installed. Additionally, it could be caused by an incorrect environment variable.

- Ask for verbose logs while you are debugging the workflow sessions. Open the task properties by right-clicking the task icon and select Edit, or just double-click the task icon:
  - On the Properties tab, select the "Write Backward Compatible Session Log File" checkbox.
  - On the Config Object tab, in the "Error handling" group, go to line 2: "Override tracing". By default, None is selected. Click on the line and select Verbose Data. Each row of data should appear.
  - On the Mapping tab, define the debug level for the source file and the destination table. On the Sources or Targets page, in the Properties group, the Tracing Level line defaults to Normal. You can select from Normal, Verbose Initialization to Verbose Data.

- If the workflow sessions do not run successfully, try eliminating `gpload` as a source of the run errors. Run `gpload` at the system (command) prompt. Create a temporary directory where you put a small representation of your source file and a control (YAML) file, and then run `gpload.py` interactively.

Use the –v (for verbose output) switch or the –V (for very verbose output), and follow the log entries for hints of what may have gone wrong.

If `gpload.py` does not run, then:

- On Windows systems, it could be that Python is not installed. Install Python (currently Python 2.5.4 is recommended). On Linux systems, Python is automatically installed.
- Check the environment variables for PATH, GPHOME_LOADERS, and PYTHONPATH. Verify that they contain the necessary paths.
- If you reinstall a different version of Python, the path to the previous version may still be in the pathname's variable PATH. You should edit the variable and remove the path to the previous version.
- If you make changes to the environment variable, restart the Informatica services so that the new values are included into the system. On Windows servers, sometimes it is necessary to reboot the PWX client computer for this to take effect.

Registration

After you have installed the PWX connector, you will need to register the connector plug-in to the repository. You have to start up the Admin Console, and go to the "plug-in" tab to register it. Failure to do so will cause PWX not to run in Informatica. Also, if the plug-in is not registered, you will not be able to see a "Greenplum writer" in the Workflow Manager "Connections".

Conclusion

Informatica customers should take advantage of the Greenplum PWX connector when loading their source data into a Greenplum database. Having `gpload` to bulk load the data enhances the performance of the data significantly.

References

The following can be found on Powerlink:

- *Greenplum Database 4.1 Administrator Guide* (P/N: 300-012-428)
- *Greenplum Database 4.1 Load Tools for Windows* (P/N: 300-012-437)