Vertica Knowledge Base Article

Vertica QuickStart for Pentaho Data Integration (Linux)

Document Release Date: 6/28/2023

MICRO FOCUS | VERTICA

Legal Notices

Warranty

The only warranties for Open Text Corporation. All rights reserved. products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Micro Focus | shall not be liable for technical or editorial errors or omissions contained herein.

The information contained herein is subject to change without notice.

Restricted Rights Legend

Confidential computer software. Valid license from Micro Focus | required for possession, use or copying. Consistent with FAR 12.211 and 12.212, Commercial Computer Software, Computer Software Documentation, and Technical Data for Commercial Items are licensed to the U.S. Government under vendor's standard commercial license.

Copyright Notice

Trademark Notices

Adobe™ is a trademark of Adobe Systems Incorporated.

Microsoft® and Windows® are U.S. registered trademarks of Microsoft Corporation.

UNIX® is a registered trademark of The Open Group.

This product includes an interface of the 'zlib' general purpose compression library, which is Copyright © 1995-2002 Jean-loup Gailly and Mark Adler.

Contents

Vertica QuickStart for Pentaho Data Integration (Linux)

The Vertica QuickStart for Pentaho Data Integration is a sample ETL application powered by Vertica Analytic Database. The QuickStart extracts data from Vertica system tables and loads it into a data warehouse called VHist (Vertica History).

Details about the ETL processing and the source and target data sets are provided in the companion document, Vertica VHist ETL Overview.

The Vertica QuickStart for Pentaho Data Integration is posted for download at the following location:

https://www.vertica.com/quickstart/vertica-quickstart-pentaho-data-integration/

About the Vertica QuickStarts

The Vertica QuickStarts are free, sample applications created using front-end products from Vertica technology partners. For an overview, watch this short video.

The QuickStarts are posted for download on the Vertica QuickStart Examples page.

> Note The Vertica QuickStarts are freely available for demonstration and educational purposes. They are not governed by any license or support agreements and are not suitable for deployment in production environments.

VHist ETL Overview

VHist ETL occurs in two steps:

1. Data is extracted from system tables in the V_CATALOG and V_MONITOR schemas and loaded into a staging schema called VHIST_STAGE. Only minimal transformation occurs during this step.
2. Data is extracted from VHIST_STAGE, transformed, and loaded into the VHist star schema.

For details, see Vertica VHist ETL Overview.

Requirements

The Vertica QuickStart for Pentaho Data Integration requires:

- A Vertica database
- The JDBC driver from the Vertica client package for Linux that corresponds to your version of the Vertica database

- JDK 1.7 or above
- Pentaho Data Integration (either the Enterprise or the Community Edition)

The QuickStart was created using Pentaho Data Integration Community Edition on Linux Centos 5, Vertica Analytic Database 7.1, and Vertica JDBC driver 7.1.

Install the Software

To install the software that is required for running the QuickStart, follow these steps:

- Install Pentaho Data Integration
- Install the Vertica Database Server
- Install the JDBC Client Driver
- Install the QuickStart

Install Pentaho Data Integration

Pentaho Data Integration (PDI) is a Java-based ETL product. If you do not already have PDI, you can install the Community Edition or a free trial of the commercial version.

To install the Community Edition:

1. Download PDI Community Edition from http://sourceforge.net/projects/pentaho/files/Data%20Integration.
2. Extract the contents of the compressed file to <*PDI_Location*> on your local machine.
3. Verify that the data-integration folder is present in <*PDI_Location*>.
4. Add EXECUTE permission to all shell scripts:

```
chmod +x <PDI_location>/data-integration/*.sh
```

To install a free trial of the Enterprise Edition:

1. Go to the Big Data Integration and Analytics page on the Hitachi Vantara website.
2. Click **START YOUR 30-DAY TRIAL**.
3. Extract the contents of the compressed file to <*PDI_Location*> on your local machine.
4. Start the installer and follow the installation instructions.
5. Verify that the data-integration folder is present in <*PDI_Location*>.
6. Add EXECUTE permission to all shell scripts:

```
chmod +x <PDI_location>/data-integration/*.sh
```

Install the Vertica Database Server

The Vertica database server runs on Linux platforms. If you do not have Vertica, you can download the Community Edition free of charge:

1. Navigate to Vertica Community Edition.
2. Log in or click **Register Now** to create an account
3. Follow the on-screen instructions to download and install the Vertica Community Edition.

Install the JDBC Client Driver

Before you can connect to Vertica using PDI, you must download and install a Vertica client package. This package includes the Vertica JDBC driver that PDI uses to connect to Vertica. Follow these steps:

1. Go to the Vertica Downloads page.
2. Under **Client Drivers**, download the Linux driver for your version of Vertica.

   > Note Vertica drivers are forward compatible, so you can connect to the Vertica server using previous versions of the client. For more information, see Client Driver and Server Version Compatibility in the Vertica documentation.

3. Place the Vertica JDBC driver jar file in the Pentaho directory for external libraries. The default directory is:

   ```
   <PDI_Location>/data-integration/lib
   ```

Install the QuickStart Application

1. Navigate to vertica.com/quickstart.
2. Select **Vertica QuickStart for Pentaho Data Integration.**
3. Log in or create an account.
4. Click **Download**.
5. Save the compressed file on your machine.
6. Extract the contents of the file to <*VHIST_Jobs_Location*>. You will see these subdirectories:
   - `config`—contains information for configuring the ETL source and target
   - `JOBS`—contains the Pentaho ETL jobs
   - `sql`—contains the SQL scripts that create and populate the VHist schema
   - `setup`—contains batch scripts for creating VHist and performing ETL
   - `logs`—contains log files in which the batch scripts record events
7. Add EXECUTE permission to all shell scripts:

   ```
   chmod +x < VHIST_Jobs_Location >/VHIST_ETL/setup/*.sh
   ```

Configure the Source and Target Databases

To configure the ETL process with your source- and target-specific information, follow these steps:

1. Close PDI if it is open.
2. Open the configuration file, properties:

   ```
   <VHIST_Jobs_Location>/VHIST_ETL/config/Config.properties
   ```

3. Edit the file, supplying the following information:
   - Your source and target server
   - Your database name
   - Your database credentials
4. Copy the contents of the file.
5. Open the file, properties:

   ```
   <user_home>/.kettle/kettle.properties
   ```

6. Paste the contents of properties into kettle.properties.

   > Note If kettle.properties is not available, then run this command:
   >
   > ```
   > sh <PDI_Location>/data-integration/kitchen.sh
   > ```

Create the Data Warehouse

Creation of the stage and star schemas is a one-time task. You can perform this task from the Linux command line, or you can use the UI.

Command Line

1. Run this command, providing both parameters:

   ```
   sh <VHIST_Jobs_Location>/VHIST_ETL/setup/setup_schema.sh
        <PDI_Location> <VHIST_Jobs_location>
   ```

2. Check the execution log to determine if the script executed successfully:

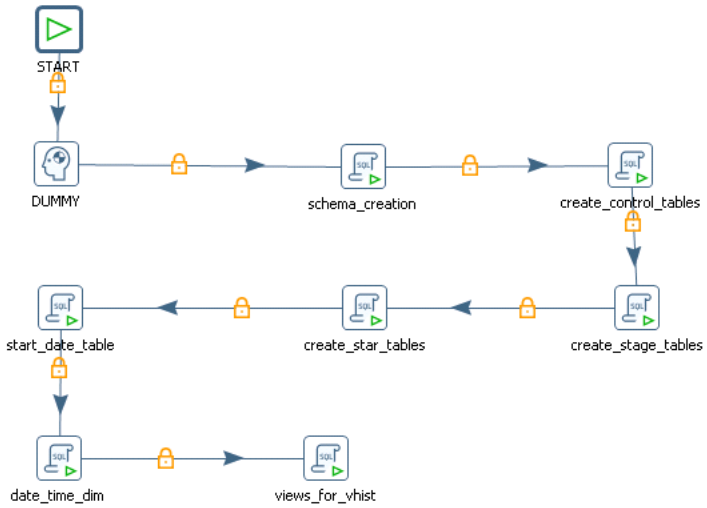
   ```
   <VHIST_Jobs_location>/VHIST_ETL/logs/create_schema.txt
   ```

User Interface

1. Start PDI.
2. From the **File** menu, choose **Open** and browse to the `JOBS` sub-folder under `VHIST_ETL`.

3. Select kjb and click **Open**.

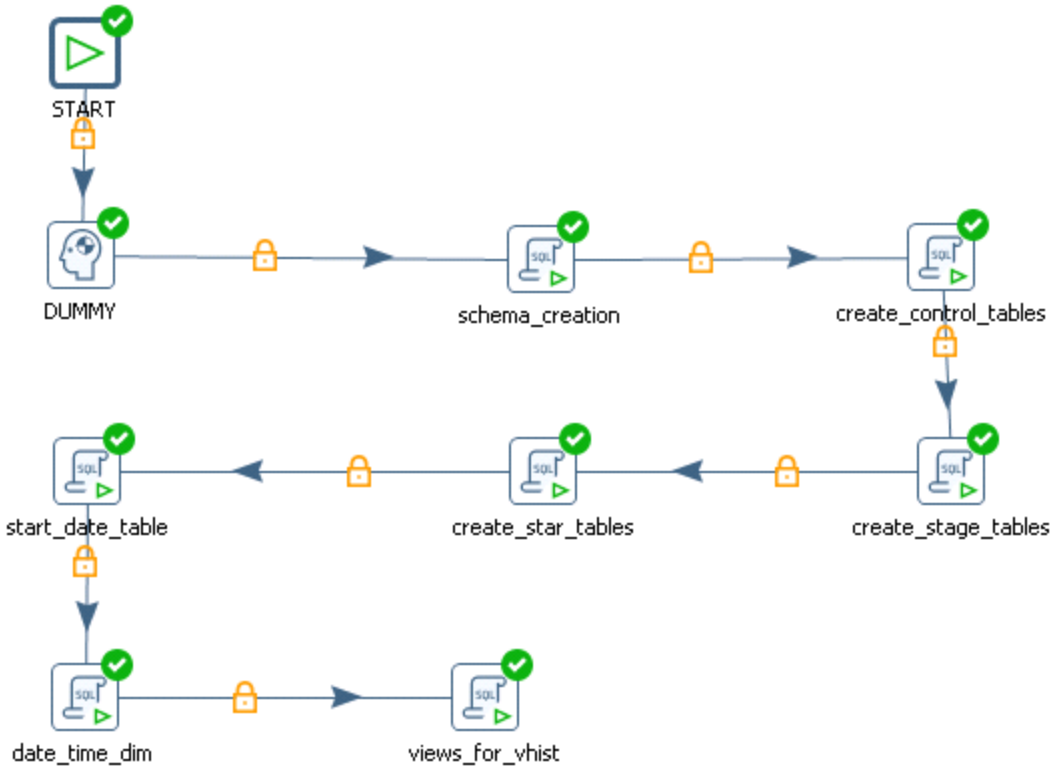
   The following Pentaho job appears:



4. Click the Run icon to run the `create_schema` job.

5. Verify the source and target information in the **Execute Job** dialog box and click **Launch**.

   A check mark appears when processing is complete.



6. Check the **Execution Results** at the bottom of the page to determine if the job executed successfully.

Populate the Data Warehouse

The initial data load populates the VHist data warehouse with data from Vertica system tables. You can perform the initial load from the Linux command line, or you can use the UI.

Command Line

1. Run this command, specifying both parameters:

```
sh <VHISt_Jobs_location>/VHIST_ETL/setup/load_VHISTDW_run.sh
      <PDI_Location> <VHIST_Jobs_location>
```

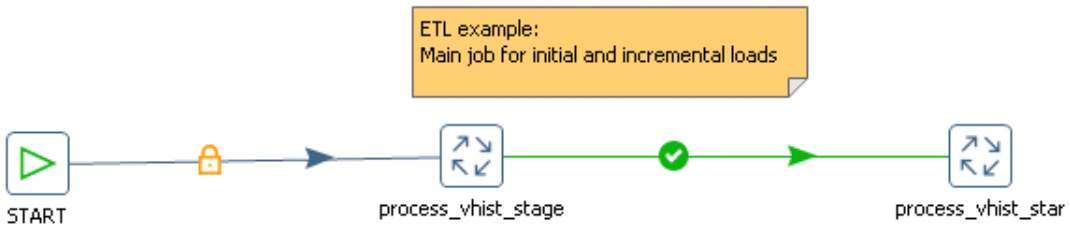
2. Check the execution log to determine if the job executed successfully.

```
/<VHIST_Jobs_location>/VHIST_ETL/logs/job.txt
```

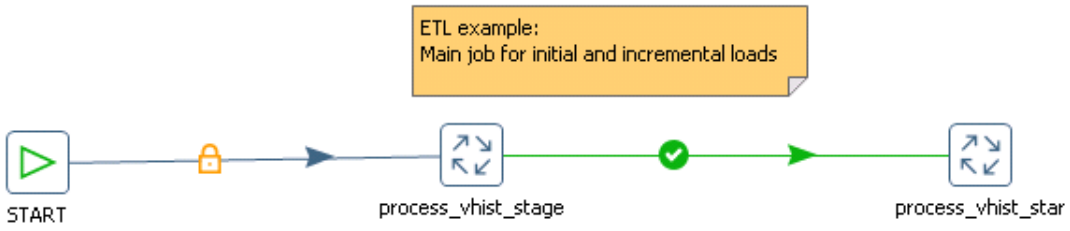
User Interface

1. Start PDI.
2. From the **File** menu, choose **Open** and browse to the `JOBS` sub-folder under `VHIST_ETL`.

3. Select kjb and click **Open**.
   A visual representation of the job displays:



4. Click the Run icon to run the job.



5. In the Execute a job dialog box, click **Launch**.
6. To verify that the job succeeded, check **Execution Results** at the bottom of the page.

Validate the ETL

The VHist ETL process records events in log tables that you can query to determine the success or failure of the data load.

To query the ETL log tables:

1. Connect to the target database using vsql or a client tool like DBVisualizer.
2. Run this query to validate the vhist_stage schema:

```
SELECT *
     FROM vhist_stage.vhist_stage_load_log
     WHERE batch_id =(SELECT max(batch_id)
          FROM vhist_stage.vhist_stage_load_log);
```

3. Run this query to validate the vhist schema:

```
SELECT *
     FROM vhist.vhist_load_log
     WHERE batch_id =(SELECT MAX(batch_id)
          FROM vhist.vhist_load_log);
```

Schedule Incremental Loads

Once the data warehouse has been created and populated, you can perform incremental loads to keep the warehouse up to date. To continually refresh the data warehouse, schedule incremental loads to run at intervals.

PDI Enterprise Edition supports a scheduling feature. You can schedule jobs using Spoon and monitor them from the PDI Server console. If you are using the Community Edition, you must use a Linux scheduling tool.

Command Line

1. At the Linux command prompt, type this command:

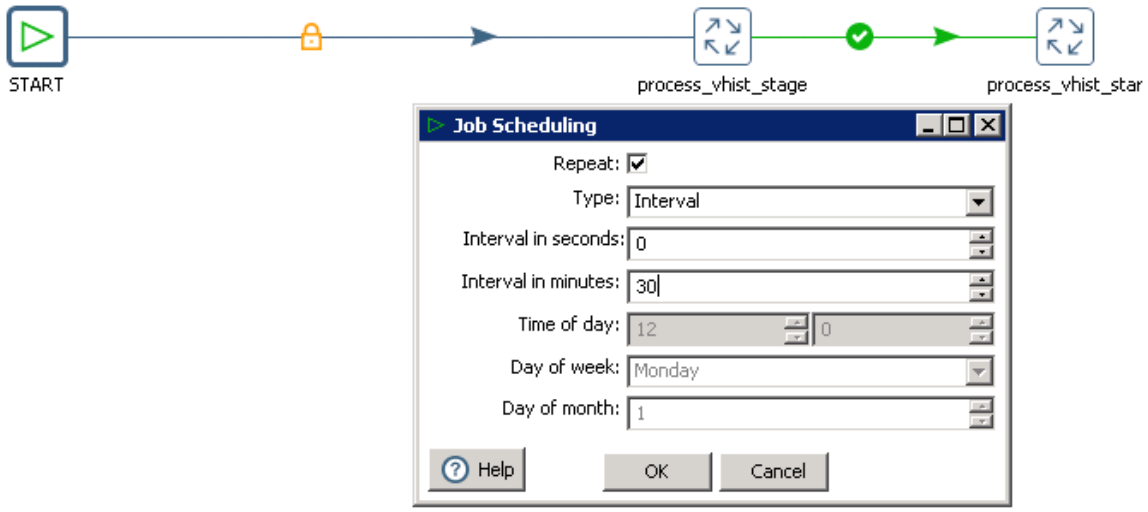
```
crontab –e
```

2. Type this line. Substitute the appropriate values for your system. In this example, incremental loads are scheduled to run every 30 minutes.

```
0,30****<VHIST_Jobs_location>/VHIST_ETL/setup/load_VHISTDW_run.sh
     <PDI_Location> <VHIST_Jobs_location>
```

User Interface (PDI Enterprise Edition)

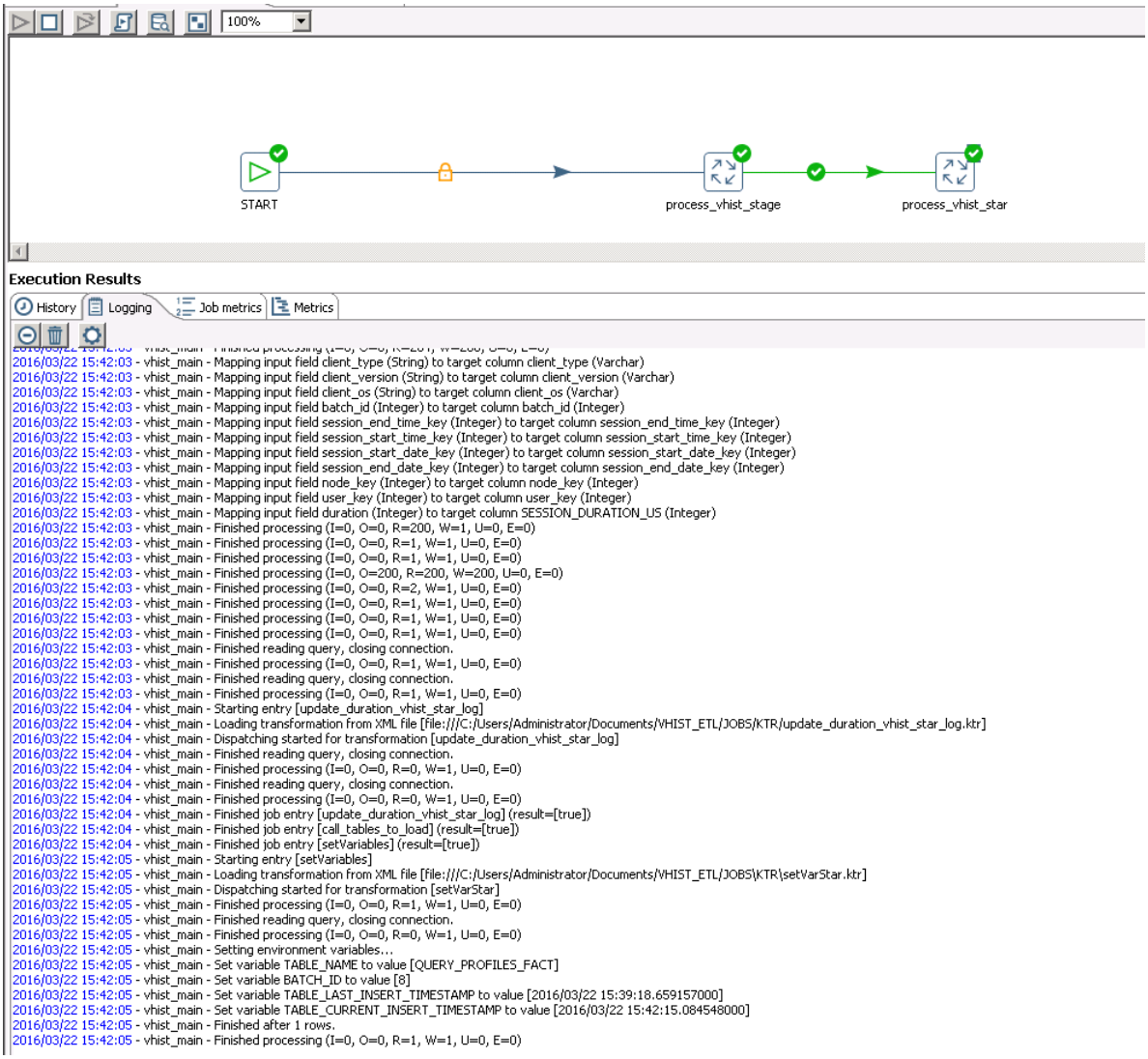
1. Start PDI.
2. From the File menu, click **Open** and select kjb.
3. Double click the **START** node of the job.
4. In the Job Scheduling window, specify the scheduling interval for the incremental loads. In this example, the incremental loads will run every 30 minutes:



5. Click **OK** to close the window.

6. Click the icon **Run this job** to start the load proces.

   From this moment on, the system will execute the load on a regular basis.



A Caution for Incremental Loads

You should take care when scheduling incremental loads to avoid placing undue demands on system resources or causing the data warehouse to grow too large. The amount of data stored in Vertica system tables is dependent on many factors, and the individual tables are not flushed at the same rate. Keep in mind the following:

- To avoid running incremental loads more often than is needed, try starting with daily loads then review the results in the log tables. If there are gaps in the results, decrease the interval between loads until you find an optimal balance.
- Repeated incremental loads increase the size of the data warehouse over time. The growth amount varies depending on system activity and frequency of collection.

> Note The data that you load into VHist counts towards the limit specified in your Vertica license.

- You may need to increase the size of the heap available to PDI. See Troubleshooting.

> Tip If you are using Vertica Community Edition, your license allows up to 1 terabyte of free storage. If you already have a licensed installation of Vertica, you can build and maintain the VHist warehouse using the Community Edition on a separate cluster.

Troubleshooting

Depending on the memory available in your environment and the amount of data that you are processing, you may need to increase the size of the heap that is available to PDI.

If you encounter this error, you need to increase the heap size:

```
exception:  java.lang.OutOfMemoryError: Java heap space
```

To increase the heap size for PDI:

1. Edit the file sh:

   ```
   <PDI_Location>/data-integration/spoon.sh
   ```

2. In the following statement, increase the values for -Xmx and -XX:MaxPermSize to values that are reasonable for your environment.

   ```
   set PENTAHO_DI_JAVA_OPTIONS="-Xmx512m" "-XX:MaxPermSize=256m"
   ```

   See the Pentaho documentation for more information:

   - How to assign max available memory to PDI: https://help.pentaho.com/Documentation/5.2/0H0/070/010/010
   - How to increase the spoon memory limit: https://help.pentaho.com/Documentation/5.2/0H0/070/020/010

Find More Information

- Pentaho
- Vertica Integration with Pentaho Data Integration: Tips and Techniques
- Vertica VHist ETL Overview
- Vertica System Tables
- Vertica Community Edition

- Vertica User Community
- Vertica Documentation