Hewlett Packard
Enterprise

HPE Vertica QuickStart for Pentaho Data Integration (Linux)

HPE Vertica Analytic Database

May, 2016

Legal Notices

Warranty

The only warranties for Hewlett Packard Enterprise 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. Hewlett Packard Enterprise 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 Hewlett Packard Enterprise 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.

Table of Contents

What is a QuickStart Application?

The HPE Vertica QuickStarts are sample applications that show how complementary technologies can work together to deliver outstanding benefits to end users. Each QuickStart uses the HPE Vertica Analytic Database with a different BI or ETL tool from a Vertica technology partner.

The QuickStarts are available for download on the HPE Big Data Marketplacehttp in the **QuickStart Examples** category.

The HPE 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.

About this Document

This document provides instructions for installing, configuring, and deploying the Vertica QuickStart for Pentaho Data Integration (PDI) on Linux. It includes an overview of the QuickStart ETL functionality.

Details about the ETL processing and the source and target data sets are provided in the companion document, *HPE Vertica VHist ETL Overview*. Both documents are included in the QuickStart download package. They are also posted on the HPE Developer Community in the Vertica Knowledge Base.

About the Vertica QuickStart for Pentaho Data Integration

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). VHist ETL occurs in two steps:

1. Extracts the data from system tables in the V_CATALOG and V_MONITOR schemas and loads it into a staging schema called VHIST_STAGE. Virtually no transformation occurs during this step. Apart from a batch ID column that is appended to each target table in VHIST_STAGE, the tables are unchanged.
2. Extracts the data from VHIST_STAGE, transforms it, and loads it into the VHist star schema.

For more information, see:

- *HPE Vertica VHist ETL Overview*
- Vertica System Tables

Requirements

The Vertica QuickStart for PDI 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 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 Download page on the Pentaho website:

   http://www.pentaho.com/download

2. Under **Pentaho Data Integration**, click the download symbol for **Linux.**
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

HPE Vertica database server runs on Linux platforms. If you do not have the Vertica database server, you can download the Community Edition free of charge from my.vertica.com. Follow these steps:

1. Go to https://my.vertica.com/.
2. Click the **Register** box.
3. Provide your information and click **Register.**
4. Follow the on-screen instructions to download and install Vertica Community Edition.

For your convenience, a link to the Vertica Community Edition is provided in the HPE Big Data Marketplace. Follow these steps:

1. Sign in to the HPE Big Data Marketplace and select **Vertica Platform.**
2. Select **HPE Vertica Analytic Database Server**.
3. Click **Get It.**
4. Provide your information and click **Register.**
5. Follow the on-screen instructions to download and install 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 HPE Vertica Downloads page on my.vertica.com.
2. Under **Client Drivers**, download the Linux driver for your version of Vertica.
3. Place the Vertica JDBC driver jar file in the Pentaho directory for external libraries. The default directory is:

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

For more information, see *Vertica Integration with Pentaho (PDI): Tips and Techniques*.

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.

Install the QuickStart Application

1. Go to the HPE Big Data Marketplace and sign in with your Marketplace credentials.
2. Select the **QuickStart Examples** category.
3. Select **HPE Vertica QuickStart for Pentaho Data Integration**.
4. Click **Download**.
5. Save the compressed file, `VHIST_ETL.zip`, 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, `Config.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, `kettle.properties`:

   <*user_home*>/.kettle/kettle.properties
6. Paste the contents of `Config.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:

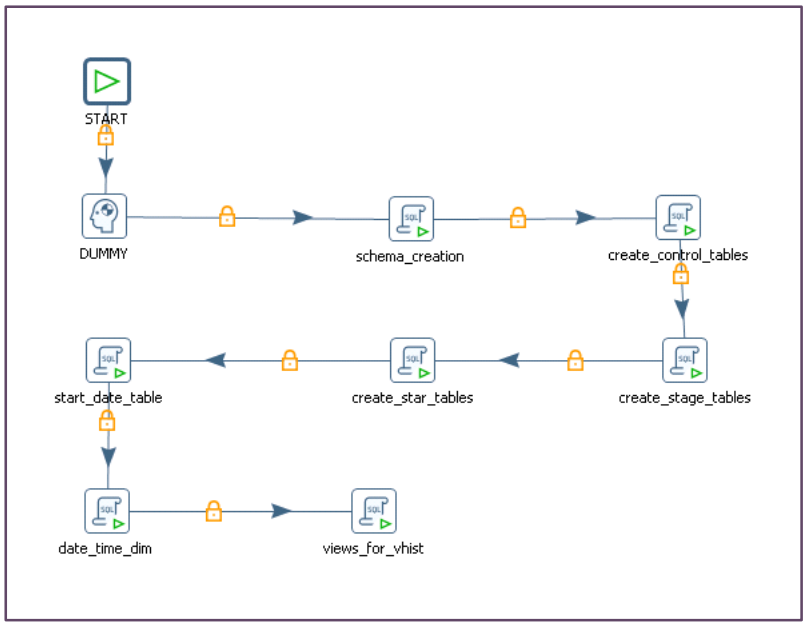
   ```
   sh <VHIST_Jobs_Location>/VHIST_ETL/setup/setup_schema.sh
        <PDI_Location> <VHIST_Jobs_location>
   ```

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

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

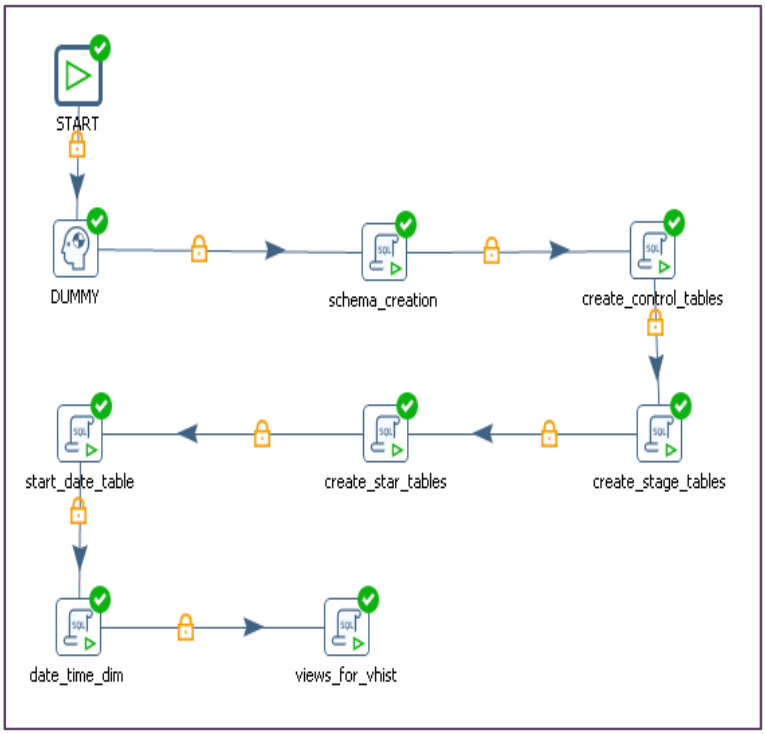
User Interface

1. Start PDI.
2. From the **File** menu, choose **Open** and browse to the `JOBS` sub-folder under `VHIST_ETL`.
3. Select `create_schema.kjb` and click **Open**.
4. The following Pentaho job appears:



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

6. Verify the source and target information in the **Execute Job** dialog box and click **Launch**.
7. A checkmark appears when processing is complete.



8. 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:

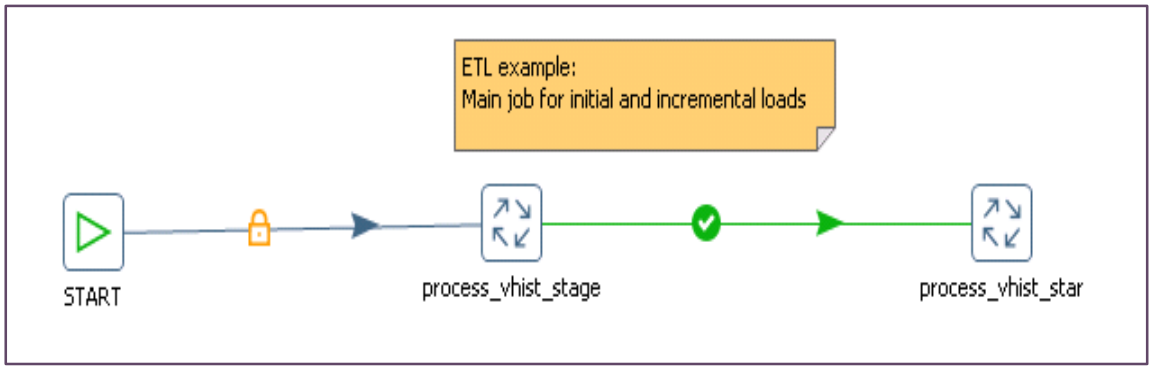
```
sh <VHISt_Jobs_location>/VHIST_ETL/setup/load_VHISTDW_run.sh
      <PDI_Location> <VHIST_Jobs_location>
```

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 `vhist_main.kjb` and click **Open**.
4. A visual representation of the job displays.



5. Click the Run icon to run the job.
6. In the **Execute a job** dialog box, click **Launch**.
7. 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
   ```

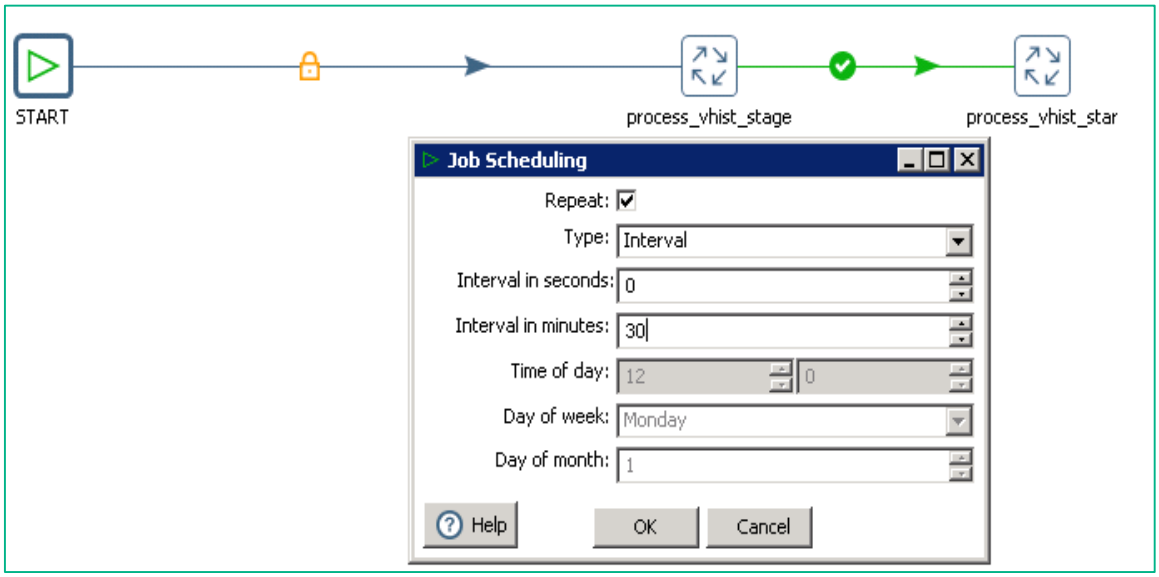
2. Type this line. Substitute the appropriate values for your system. In this example, incremental loads are scheduled to run every 30 minutes:

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

User Interface (PDI Enterprise Edition)

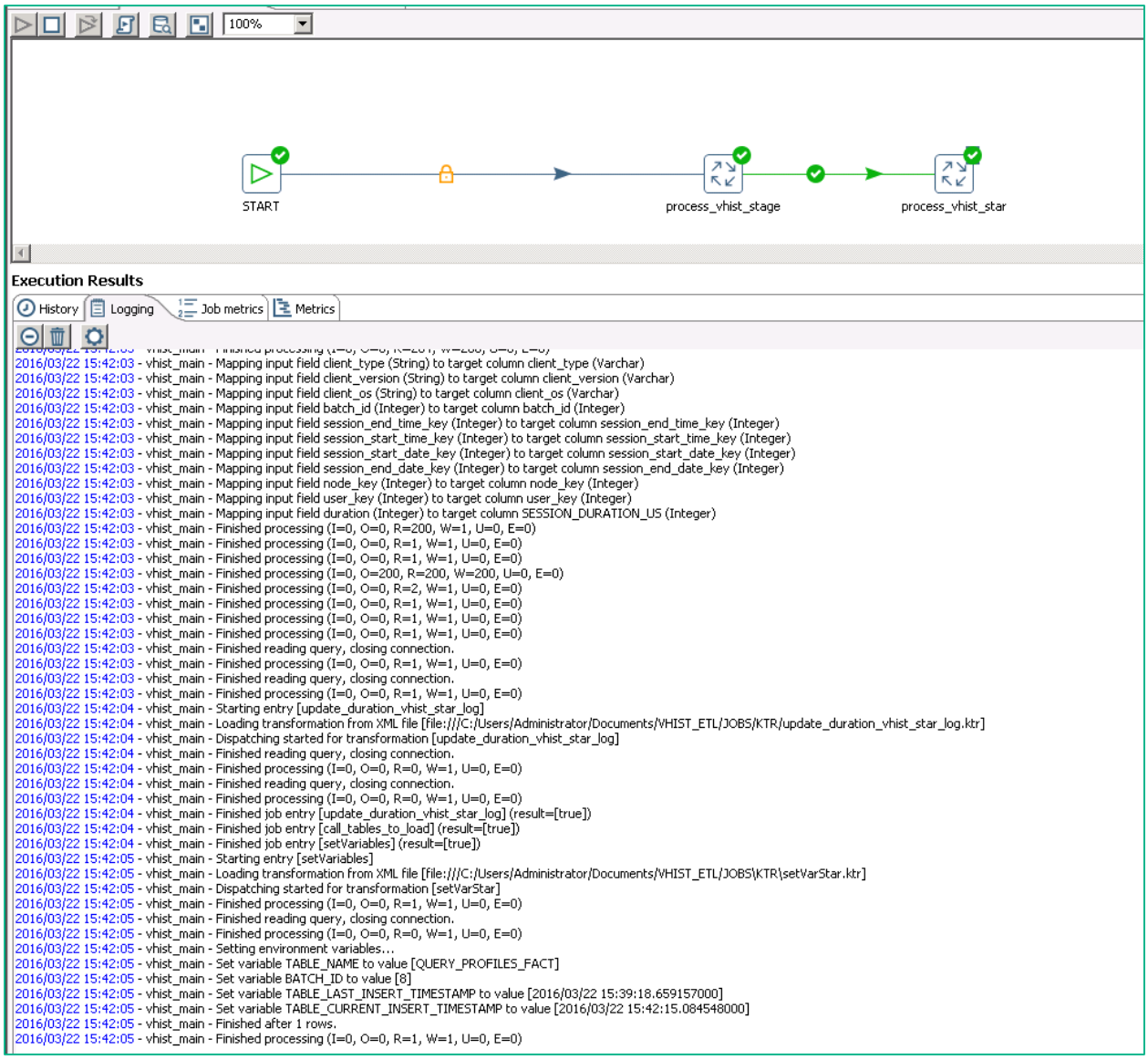
1. Start PDI.
2. From the File menu, click **Open** and select `vhist_main.kjb`.
3. Double click the **START** node of the job.

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 process.

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 the Community Edition, your license allows up to one 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 in 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 `spoon.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

Subject	Link
Pentaho	www.pentaho.com
QuickStart video	https://youtu.be/w1lyQxEx8iA
Vertica and PDI integration	Vertica Integration with Pentaho Data Integration: Tips and Techniques
VHist Overview	http://my.vertica.com/docs/Ecosystem/QS_ETL_Vhist_overview.pdf
HPE Vertica	https://my.vertica.com/
HPE Vertica documentation	https://my.vertica.com/hpe-haven-documentation/
HPE Big Data Marketplace	https://saas.hp.com/marketplace/haven
HPE Developer Community	https://community.dev.hpe.com/t5/Big-Data-and-Analytics/ct-p/bigdata_analytics

Contact Us

We welcome your feedback. If you have questions, comments, or suggestions, please contact us by clicking the **Contact developer** button on the HPE Big Data Marketplace download page for the QuickStart.