Vertica Knowledge Base Article

Vertica QuickStart for Talend Data Integration

Document Release Date: 2/26/2019

MICRO FOCUS | VERTICA

Legal Notices

Warranty

The only warranties for Micro Focus International plc 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 Talend Data Integration

The Vertica QuickStart for Talend Data Integration is a sample ETL application powered by Vertica Analytic Database. The QuickStart uses Talend Data Integration to extract data from Vertica system tables and load 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.

You can download the Vertica QuickStart for Talend Data Integration from the following location:

https://www.vertica.com/quickstart/vertica-quickstart-for-talend-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 Talend Data Integration requires the following:

- A Vertica database
- JDK 1.8 or above
- Talend Data Integration, either the Community Edition (Talend Open Studio for Data Integration) or the Enterprise Edition.

The Vertica QuickStart for Talend Data Integration was created and tested using:

- Talend Open Studio for Data Integration version 6.1 (the Community Edition)
- Linux Centos 5
- Vertica Analytic Database 7.1.0
- Vertica JDBC driver 7.0.1
- JDK 1.8

Install the Software

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

- Install Talend Data Integration
- Install the Vertica Database Server
- Install the QuickStart Application

Install Talend Data Integration

If you want to develop the Talend QuickStart ETL job yourself, you must have Talend Data Integration installed on your computer. If you do not already have Talend Data Integration, you can download the free Open Studio version, or you can obtain a free trial of the Enterprise Edition.

To install Talend Open Studio for Data Integration:

1. Go to the Download page on the Talend website: https://www.talend.com/download/
2. Scroll down to **Download Free Talend Products**.
3. Under **Data Integration**, click **DOWNLOAD FREE TOOL**.
4. Download the compressed file and extract the contents to <*Talend_Location*> on your local machine.
5. Verify that the TOS folder for your version of Talend Data Integration is present in <*Talend_Location*>. For example: TOS_DI-20151214_1327-V6.1.1 for Talend Data Integration 6.1.
6. Add EXECUTE permission to all shell scripts:

```
chmod +x <Talend_Location>/<TOS_folder>/*.sh
```

To install a free trial of Talend Enterprise for Data Integration:

1. Go to the Download page on the Talend website: http://www.talend.com/download/
2. Under **Test Drive Talend Products**, select **Data Integration** and click **Try Now**.
3. Download the compressed file and extract the contents to <*Talend_Location*> on your local machine.
4. Start the installer, and follow the installation instructions.
5. Verify that the data-integration folder is present in <*Talend_Location*>.

6. Add EXECUTE permission to all shell scripts:

```
chmod +x <Talend_Location>/<TOS_folder>/*.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 QuickStart Application

The Vertica QuickStart for Talend Data Integration download package includes:

- The QuickStart application
- The JDBC driver that connects Talend Data Integration to the Vertica database server
- Two PDFs:
  - *Vertica QuickStart for Talend Data Integration,* which provides installation and deployment instructions (this document)
  - *VHist ETL Overview*, which provides details about VHist ETL and the source and target schemas.

To install the QuickStart:

1. Navigate to vertica.com/quickstart.
2. Select **Vertica QuickStart for Talend 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_Job_Location* >. You will see these subdirectories:
   - `config`—contains information for configuring the ETL source and target
   - `dev`—contains the source code of the VHist ETL project as a zip file (Job Designs.zip)
   - `jobs`—contains the Talend ETL jobs
   - `logs`—contains log files in which the batch scripts record events
   - *`setup`*—contains batch scripts for creating VHist and performing ETL
   - `sql`—contains the SQL scripts that create and populate the VHist_stage and VHist schemas

7. Add EXECUTE permission to all shell scripts:

```
chmod +x < VHIST_Job_Location>/VHIST_ETL/setup/*.sh
chmod +x < VHIST_Job_Location>/VHIST_ETL/jobs/create_schema/*.sh
chmod +x < VHIST_Job_Location>/VHIST_ETL/jobs/Load_VHISTDW/*.sh
```

Configure the Source and Target

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

1. Open the configuration file, properties:

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

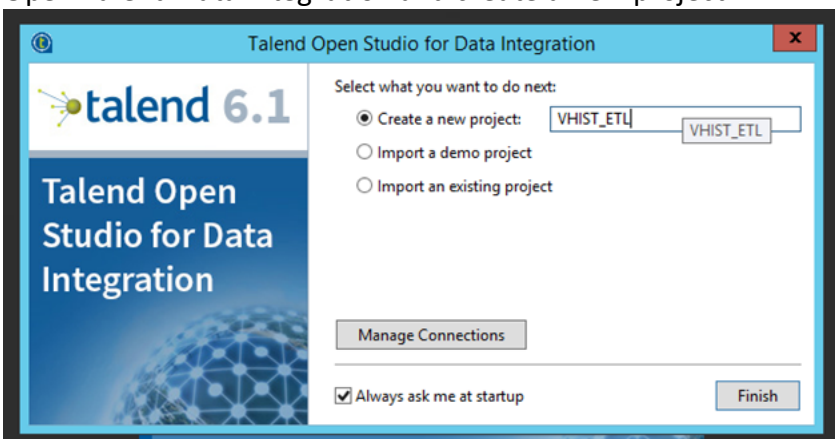
2. Edit the file, supplying the following information:
   - Your source and target server
   - Your database name
   - Your database credentials

Create the ETL Job in Developer Mode

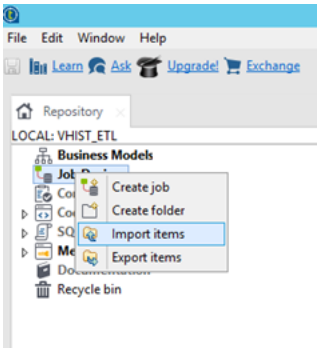
To create the ETL job in developer mode, you must create and configure a project in Talend Data Integration. Then you can create and populate the data warehouse within the project.

Create a Project

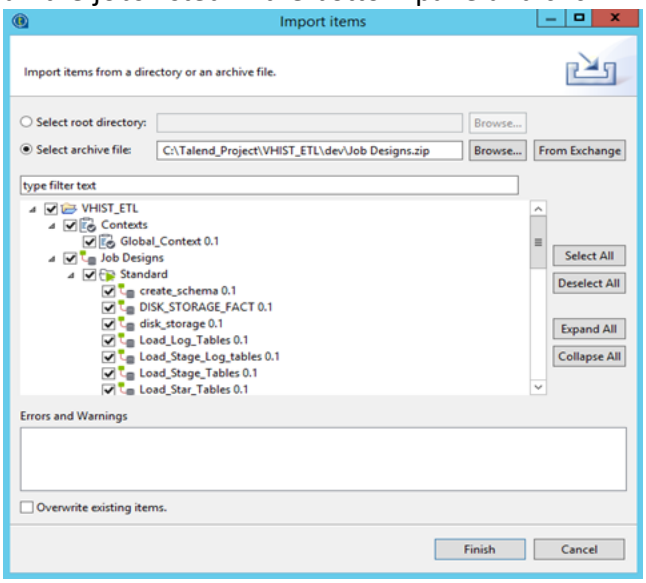
1. Open Talend Data Integration and create a new project.



2. To import the QuickStart VHist ETL source code, right click **Job Designs** and click **Import items**.

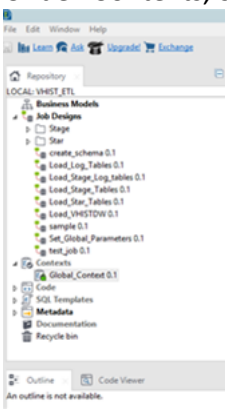


3. In the **Import items** dialog box, select the path of the source code archive file, then select all the jobs listed in the bottom pane and click **Finish**.

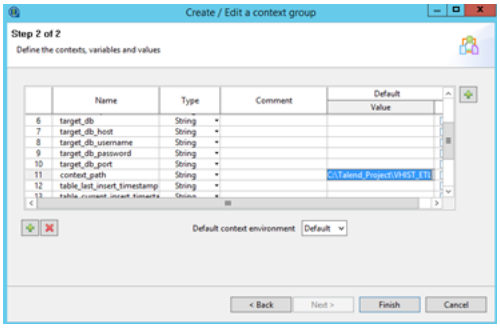


The QuickStart VHist ETL jobs are now available under **Job designs**.

4. Under Contexts, select Global_Context



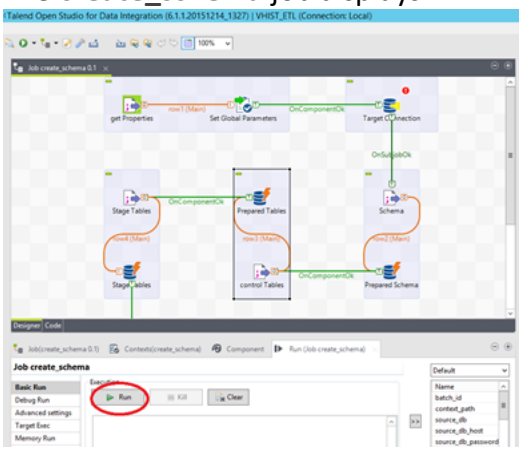
5. Set the context_path to <*VHIST_Job_Location*>/VHIST_ETL. Leave the other variables blank.



6. Click Finish to complete the configuration of the QuickStart VHist ETL project.
   The VHist ETL jobs are now available for **Edit**, **Execute**, **Build** and **Export**.

Create the Data Warehouse

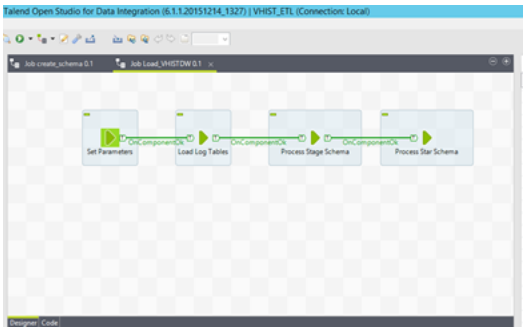
1. Under **Job Designs**, select **create_schema**.
2. Click the run icon.
3. The **create_schema** job displays.



Populate the Data Warehouse

The Load_VHISTDW job loads the stage and star schemas. Run this job for the initial and incremental loads.

1. Under **Job designs**, select **Load_VHISTDW**.
2. Click the run icon.
   The load process displays.



Create the ETL Job in Batch Mode

Once you have installed, configured, and tested the development deployment, you can use Talend Data Integration to build the ETL job as a jar file. The Build functionality in Talend Data Integration lets you create standalone scripts for Windows batch and Linux shell deployments. To build the jar file, follow the instructions in the Talend documentation.

The following instructions for Batch Mode deployment assume that you have already built the ETL job using Talend Data Integration:

1. Create the data warehouse (a one-time task):
   Run this command, providing the parameter, <*VHIST_Job_Location*

   ```
   sh <VHIST_Job_Location>/VHIST_ETL/setup/setup_schema.sh
        <VHIST_Job_Location>
   ```

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

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

2. Populate the data warehouse:
   Run this command, specifying the parameter <*VHIST_Job_Location* >:

   ```
   sh <VHIST_Job_Location >/VHIST_ETL/setup/load_VHISTDW_run.sh
         VHIST_Job_Location >
   ```

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

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

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.

To schedule incremental loads:

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_Job_Location >/VHIST_ETL/setup/load_VHISTDW_run.sh
              <VHIST_Job_Location >
```

3. Restart crontab with the following command:

```
$>service crond restart
```

Cautions 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 Talend.

  > Tip If you are using the Community Edition of Vertica, your license allows up to one terabyte of free storage. If you already have a licensed installation of Vertica, you

> can build 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 Talend.

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

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

The default heap size for Talend is:

```
java -Xms256M -Xmx1024M
```

To increase the heap size for Talend:

1. Edit the file sh

   ```
   <VHIST_Job_Location >/VHIST_ETL/jobs/Load_VHISTDW/Load_VHISTDW_
   run.sh
   ```

2. In the following statement, increase the values for -Xms and –Xmx to values that are reasonable for your environment.

   ```
   java -Xms256M -Xmx1024M
   ```

See the Talend documentation for more information.

Find More Information

- Talend
- Vertica Integration with Talend: Tips and Techniques
- Vertica VHist ETL Overview
- Vertica System Tables
- Vertica Community Edition
- Vertica User Community
- Vertica Documentation