Install Guide for vRODBC

Version	1.0
Date	09-27-2013
Author	Shreya Prasad <sprasad@vertica.com>
Description	This documents provides instructions on how to Install, configure and use vRODBC.

1. About vRODBC

vRODBC is an ODBC client which provides database connectivity to the R environment. vRODBC is an HP Vertica enhanced version of the RODBC open-source ODBC client. vRODBC provides significantly faster data transfer rates from HP Vertica to systems running R-instances than RODBC for many types of data.

vRODBC provides significantly better data load performance than RODBC for data sets including numeric values, floating point numbers and boolean values. For Character data types, you can also achieve better data load performance with vRODBC if you load the data in your R-environment as "`characters`" rather than "`factors`". Note that you must set the R-env option `stringsAsFactors` to `FALSE` to obtain better character data load performance.

2. Software Requirements

vRODBC requires following pre-requisites softwares to be Installed and configured, before Using vRODBC.

- R
- unixODBC- a Linux Driver Manager
- HP Vertica ODBC Driver

HP Vertica requisites that unixODBC must be installed before proceeding with Installation of HP Vertica ODBC Driver.

3. Installation of pre-requisite softwares

Note: You must have root or sudo privileges to perform the Installation.

Below are the steps to Install of the pre-requisite softwares:

1. **Obtaining and Installing R**

   R is available from The Fedora Project's EPEL (Extra Packages for Enterprise Linux) RPM repository.

   1. Download the EPEL yum repository configuration file from http://epel.mirror.freedomvoice.com/6/i386/epel-release-6-8.noarch.rpm

   2. Install the EPEL configuration file:

   ```
   sudo rpm -Uhv epel-release-6-8.noarch.rpm
   ```

   3. Install R package:

   ```
   sudo yum install R
   ```

2. Obtaining and Installing unixODBC

1. Download latest version of UnixODBC unixODBC-xxx.tar.gz from www.unixodbc.org.

2. Unpack the package

```
tar -xvf unixODBC-xxx.tar.gz
```

3. To Install, run commands:

```
cd unixODBC-xxx
sudo ./configure -prefix=/usr
sudo make
sudo make install
```

4. unixODBC is now Installed.

3. Obtaining and Installing HP Vertica ODBC Driver

Note: If HP Vertica Analytics Platform is already installed in your Linux system, then there is no need to follow the below steps to download and Install ODBC driver. ODBC client drivers are already installed by the Server .rpm files. The user just need configure the drivers. See section X to find instructions on how to configure client drivers.

1. Log in to myVertica portal http://my.vertica.com

2. Go to Downloads section to locate 64-bit HP Vertica 6.1.2 Linux ODBC driver package and download it. Package name: `vertica-odbc-6.1.2-xx.x86_64_linux.tar.gz`

3. Login as root to the machine.

4. If the directory `/opt/vertica/` does not exist, create it:

   ```
   mkdir -p /opt/vertica/
   ```

5. Copy the downloaded file to the `/opt/vertica/` directory. For example:

   ```
   cp vertica-odbc-6.1.2-xx.x86_64_linux.tar.gz /opt/vertica/
   ```

6. Change to the /opt/vertica/ directory and uncompress the file

   ```
   cd /opt/vertica
   tar vzxf vertica-odbc-6.1.2-xx.x86_64_linux.tar.gz
   ```

7. Two folders are created which contains the necessary files for HP Vertica ODBC Driver
   A. `/opt/vertica/include.` Contains the header file
   B. `/opt/vertica/lib64`. Contains the library file

8. Installation of HP Vertica ODBC Driver is complete. However, User must configure the ODBC Driver before you can use it. See Section 4. *Post Driver Installation Configuration* for details on how configure ODBC Driver.

4. Post ODBC Driver Installation Configuration

User must configure the ODBC Driver before you can use it. This includes creating configuration files and setting mandatory environment variables.

- User must create 2 Configuration files as below:

  a) **odbc.ini.** It defines the Data Source Names(DSNs) that tells the ODBC how to access the HP Vertica Database.

  Create a file namely, odbc.ini in the following format:

  ```
  [Test]
  Description = vRODBC Test
  Driver = /opt/vertica/lib64/libverticaodbc.so
  Database = testdb
  Servername = host01
  UserName = dbadmin
  Password = password
  Port = 5433
  ConnSettings =
  Locale = en_US
  ```

  Configuration Options.

  1. **ODBC Data Source Section.** The name of this section is the DSN Name, It is called while using vRODBC to identify and connect to specific HP Vertica database.
  2. **Description.** Additional information about the data source.
  3. **Driver.** This specifies the location of the ODBC Driver which user installed in Step (2).
  4. **Database.** Name of the database running on the Server.
  5. **Servername.** Name of the Server where HP Vertica is installed.
  6. **UserName.** Either the database superuser (same name as database administrator account) or a user that the superuser has created and granted privileges.
  7. **Password.** Password of the specified UserName.
  8. **Port.** The port number on which HP Vertica listens for ODBC connection .
  9. **ConnSettings.** Can contain SQL commands separated by a semicolon. These commands can be run immediately after connecting to the server.
  10. **Locale.** The default locale used for the session.

b) **vertica.ini.** It defines some HP Vertica-specific settings required by the drivers .

Create a file namely, vertica.ini in the following format:

```
[Driver]
DriverManagerEncoding = UTF-16
ODBCInstLib = /usr/lib/libodbcinst.so
ErrorMessagesPath = /opt/vertica/lib64
LogLevel = 0
LogPath = /tmp
```

`ODBCInstLib` and `ErrorMessagesPath` are mandatory settings and should be set in order for the ODBC driver to work correctly.

- Once Configuration files are created, User must set environment variables `$VERTICAINI` and `$ODBCINI` to the absolute file path of `vertica.ini` and `odbc.ini` files respectively. Set the environment variable as and add it to shell startup files (for example, in the .bashrc file of dbadmin):

```
export VERTICAINI=<Absolute path of vertica.ini>
export ODBCINI=<Absolute path of odbc.ini>
```

- Finally, ensure that Operating System/s "lib" search path, for example /usr/lib is added to shell's library search path variable LD_LIBRARY_PATH.

Installalation and Configuration settings are complete for pre-requiste softwares for using vRODBC.

5. Installation of vRODBC

Note: You must have root or sudo privileges to Install vRODBC.

To Install vRODBC:

1. Download the vRODBC tar file from the FTP site indicated on your cover page.
2. Install vRODBC

```
sudo R CMD INSTALL vRODBC_0.2.0-xxx.tar.gz
```

3. vRODBC is Installed and ready to use.

6. Test vRODBC

This section provides sample code that user can run once vRODBC and its pre-requisite packages are Installed and Configured successfully.

Database Setup. Let there be a HP Vertica Server on machine `host01` which has a database `testdb`. This information is set up under ODBC DSN Name `[Test]` in *odbc.ini* configuration file.

Let `Testdb` have table customer_info defined as follows:

Column
id
CustType
Feature1
Feature2
Feature3

Running vRODBC:

1. Start a R-session.

   ```
   $ R
   ```

2. Load vRODBC library

   ```
   > library(vRODBC)
   ```

3. Create a vRODBC Connection to database defined in `Test` ODBC DSN.

   ```
    > connect <- odbcConnect("Test")
   ```

4. Issue Query to the database. Result set is saved in segment variable.

   ```
   > segment <- sqlQuery(connect, "select * from customer_info")
   ```

5. Check the result set

   ```
   > segment
     id CustType feature1 feature2 feature3
   1 22        A        7   10.273   0.2746
   2 61        A       11    3.567    0.812
   3 32        D       24   18.320   0.0279
   ```

6. Close vRODBC connection.

   ```
   > odbcClose(connect)
   ```

7. Performance of vRODBC

vRODBC has been developed to achieve improved data load timings from HP Vertica to an R-instance. It is used in extensively in Package HPDGLM, a distributed Generalized Linear Model based on distributedR enviornment, to load data faster from HP Vertica to distributedR using its function `dataLoader()`. See HPDGLM-Manual and HPDLGM-UserGuide for further details on HPDGLM and its dataLoader() functionality. See Distributed-R-Manual and Distributed-R-UserGuide for further details on distributedR software.

Table 1. shows DataSet load timings from HP Vertica to a distributedR cluster. To load data from HP Vertica, DataLoader() in HPDGLM Package with vRODBC ODBC client is used.

- Load timings are presented across 3 data sizes – <500 million rows * 7 columns>, <1 billion rows * 7 columns> and <2.5 billion rows * 7 columns>. All columns are of Numeric type (interger values and floating point values)

- Data set is loaded on a 7-node distributedR cluster

- Data load timings are in minutes

DataSet size: # rows (size)	7-node distributedR cluster
500 Million rows (38.5 GB)	3.8
1 Billion (77 GB)	6.4
2.5 Billion (161.7 GB)	15.8

Table 1. vRODBC Performance with HPDGLM.dataLoader()