White Paper

Implementing SAS Access ® 9.3 with Greenplum Database

Planning, Implementation, Configuration and Best Practices

Abstract

This paper outlines recommendations for SAS software with the Greenplum Database via SAS-Access. It is designed to help customers and field personnel implement basic SAS connectivity in a Greenplum database environment.

December 2012

Copyright © 2012 EMC Corporation. All Rights Reserved.

EMC believes the information in this publication is accurate of its publication date. The information is subject to change without notice.

The information in this publication is provided “as is”. EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com.

Part Number hxxxxx

Table of Contents

[Executive Summary 4](#_Toc338751279)

[Audience 4](#_Toc338751280)

[Scope 4](#_Toc338751281)

[Greenplum Database Overview 6](#_Toc338751282)

[EMC Greenplum Data Computing Appliance 6](#_Toc338751283)

[EMC Greenplum Data Integration Accelerator 7](#_Toc338751284)

[Overview of SAP Business Objects 8](#_Toc338751285)

[Installation and Configuration Process 9](#_Toc338751286)

[Download the SAP® BusinessObjects™ server software 10](#_Toc338751287)

[Obtaining a BusinessObjects™ software license 12](#_Toc338751288)

[Download the SAP® BusinessObjects™ client tools 13](#_Toc338751289)

[Identify the appropriate ODBC driver on Powerlink 15](#_Toc338751290)

[Download the identified ODBC driver from Subscribenet 19](#_Toc338751291)

[Install the pre-requisite SAP® BusinessObjects™ application host software 23](#_Toc338751292)

[Install the BusinessObjects™ software 23](#_Toc338751293)

[Install the ODBC driver 25](#_Toc338751294)

[Validate the ODBC driver 28](#_Toc338751295)

[Solution Testing 32](#_Toc338751296)

[Creating a new 32-bit ODBC connection to Greenplum 32](#_Toc338751297)

[Starting the Information Design Tool 32](#_Toc338751298)

[Universe Design Tool 45](#_Toc338751299)

[Web Rich Client 58](#_Toc338751300)

[Performance Tuning 63](#_Toc338751301)

[Best Practices 74](#_Toc338751302)

[Conclusion 75](#_Toc338751303)

[References 76](#_Toc338751304)

# Executive Summary

The Greenplum Database is a high performance analytics platform based on a shared-nothing MPP (Massively Parallel Processing) architecture. Greenplum Database facilitates Business Intelligence and Data Analytics using commodity hardware. The database integrates with a wide variety of 3rd party vendor software products including solutions for Data Integration, Business Intelligence, Data Modeling and Data Analytics.

SAS 9.3 is the latest version of the SAS Analytics suite of software. It provides customers the ability to make business decisions through self-service access to data analysis. The software is designed to help organizations make better business decisions based on advanced statistical analytics.

The purpose of this paper is to describe the process needed to successfully implement SAS-Access in a Greenplum environment. It provides a step-by-step outline to successfully design, implement, and configure SAS connecvity with Greenplum. It also describes some basic best practices for testing the solution and ensuring success in a production system.

## Audience

This white paper is intended for EMC field facing personnel including technical consultants, solution architects, and support engineers, as well as customers who are considering integration of SAS in a Greenplum database environment.

## Scope

This document provides an overview of the steps needed to complete a successful implementation of the SAS 9.3 software in a GPDB environment. It includes the software required in the SAS environment, basic steps needed to install the SAS-Access software, the steps needed to install, configure, and validate connectivity to the Greenplum database via a supported data driver, and perform basic interoperability testing of the Greenplum/SAS solution. It also provides some best practices for the solution. This guide does not provide information on installing the Greenplum Database – this information can be found in the Greenplum Database Installation Guide found on EMC Powerlink. More detailed information on the installation of SAS can be found in the references found at the end of this document. The examples used in this document are provided to give the reader a high-level understanding of SAS functionality in a Greenplum environment.

# Greenplum Database Overview

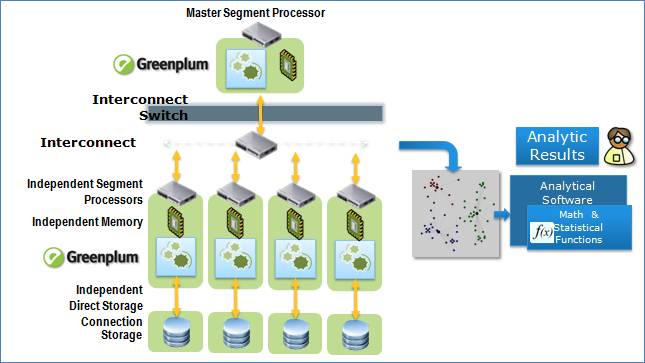
The Greenplum Database design is based on a share-nothing MPP (Massively Parallel Processing) architecture which facilitates Business Intelligence and analytical processing using commodity hardware. Data is distributed across multiple segment servers in the Greenplum Database to achieve no disk-level sharing. The segment servers are able to process queries in a parallel manner in order to promote the highest degree of parallelism and scalability.

Highlights of the Greenplum Database include:

* Dynamic Query Prioritization which provides continuous real-time balancing of the resources across queries
* Self-Healing Fault Tolerance which provides intelligent fault detection and fast online differential recovery
* Polymorphic Data Storage and Multi-storage/SSD Support which includes tunable compression and support for both row-and column-oriented storage
* Analytics and Language Support which supports analytical functions for advanced in-database analytics
* Health Monitoring and Alerting, which provide integrated email and SNMP notification for advanced support capabilities.

## EMC Greenplum Data Computing Appliance

EMC’s Greenplum Data Computing Appliance (DCA) is a purpose-build, massively parallel processing (MPP) data warehousing appliance that is created to integrate storage, database, and networking into a single enterprise-class system based on the Greenplum Database. It is built to deliver the industry’s fastest data loading speed, and can linearly expand to accommodate customers’ storage requirements for Big Data. It takes advantage of large clusters of increasingly powerful, commodity servers, storage and network switches to minimize the customers’ cost of ownership. The database software is Greenplum Database, a “shared-nothing” architecture that is optimized for fast queries and data loading, for fastest operations with the maximum degree of parallelism possible. The DCA integrates Greenplum Database, data loading and Hadoop software with compute, storage and network components, delivered racked and ready for immediate data loading and query execution.

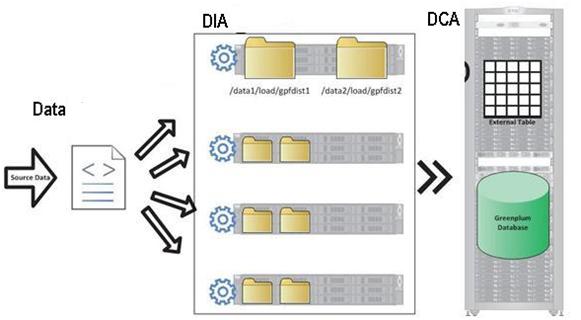


## EMC Greenplum Data Integration Accelerator

The Data Integration Accelerator (DIA) is specially built to facilitate fast data loading to the DCA. It integrates the Greenplum data loading software called gpfdist with the server, storage and networking gear into a single system. It leverages the high-speed internal communication network to deliver the data quickly to the DCA.

Each server in the DIA is preloaded with RedHat Enterprise Linux operating systems. These RedHat Linux hosts are intended to be configured and have the requisite software stacks installed so they can serve as a ‘near the DCA’ platform for data integration software.

The DIA comes in blocks of 4 servers. Each block is referred to as a module; up to 4 modules of DIA can be ordered in each rack. With each server, the Greenplum gpfdist utility is pre-loaded by default; this utility provides parallel file distribution server utility software that is used with read-only external tables for fast, parallel data loading of text files into a Greenplum database.



# Overview of SAS Data Integration Tools

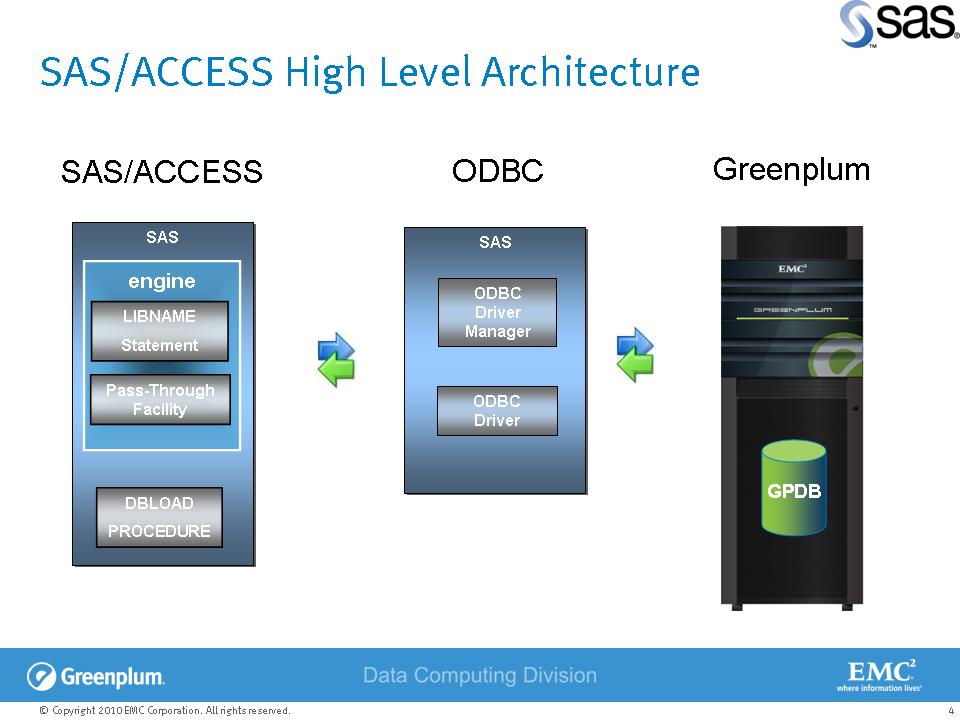
SAS Data Integration tools provide capabilities for enterprise data access and processing across systems and platforms. The key components of SAS Data Integration are Enterprise Data Integration, Data Quality, and Data Access Engines. This paper focuses on the Data Access Engines and specifically SAS/ACCESS for GreenPlum.

SAS/ACESS interfaces are out-of-the-box solutions written specifically to run on client/server database systems where a database engine supplies data to the local application. In this paper, all of the examples uses SAS as the client application and GreenPlum database as the data provider.

Access to the database requires sending database commands from the SAS client to the database engine to be executed on the server. Key features of the SAS Data Integration tools are:

* Data Integration – allows the user to treat data as a resource to be viewed and used in various SAS programs regardless of data format.
* SQL Support – SAS provides both a named “library” (LIBNAME) and Pass-Through SQL as the basic means of data integration.
* Bulk-loading – moves data from SAS into third-party data stores
* Temporary Table Support – creates tables that can be accessed by multiple SAS processes.
* Metadata Integration – maintains database metadata within the SAS Metadata repository
* Data Integrity and Security – enables encoding of database passwords for login authentication.

In general, it is usually easier to use SAS/ACCESS ODBC to access the database. This requires setting up and ODBC datasource on the client system pointing to a supported ODBC driver. In the case of GreenPlum, this is the Data Direct ODBC driver for GreenPlum database. This only needs to be done once for each client workstation, after which the GreenPlum database will be accessible.



# Installation and Configuration Process

Installation and configuration of SAS/ACCESS with Greenplum Database requires the following high-level steps:

1. Insure the SAS Foundation software is installed in the environment.
2. Obtain a license for the SAS/ACCESS software. (Procedure for this)
3. Download the SAS/ACCESS software if it was not included in the original depot) and install. (Procedure/Screen shots needed)
4. Download the Data Direct Driver for SAS/ACCESS and GreenPlum and install.
5. Install the ODBC driver and map a DSN for the GreenPlum database.
6. Install the SAS/ACCESS software on the SAS Foundation Server.
7. Start SAS Explorer.
8. Test a LIBNAME statement from SAS to the GreenPlum database.
9. Test a Pass-Through SQL statement from SAS to the GreenPlum database.

The following sections in this document describe each of these steps in more detail.

## Download and Install the SAS® ™ server software

Before you can use the SAS/ACCESS for Greenplum features, you must have a valid SAS license and SAS software depot for installation. The full install process is outside of the scope of this document. Please reference the *Configuration Guide for SAS 9.3 Foundation for Unix Environments*  or the *Configuration Guide for SAS 9.3 Windows Environments* product documentation from the SAS web site at: <http://support.sas.com/documentation/installcenter/>

## Confirm the appropriate ODBC driver is included with the SAS software

Connectivity is established between the SAS® Foundation/Enterprise™ server and the Greenplum database via an ODBC driver. This should be a Progress DataDirect ODBC driver.

**Pre-Installation Checklist**

1. Verify SAS-ACCESS License is Active.
2. Gather Connectivity information for Greenplum Database
   1. Greenplum database host name or IP
   2. Greenplum Login/Password
   3. Port Number GreenPlum is listening on
   4. Database Name to connect to.
3. Verify network connectivity between SAS server and Greenplum server with ping from/to each system.
4. If you plan on using Bulk Loading, verify Segment Servers can ping SAS Server
5. Is a firewall enabled? If so, configure for read/write access for SAS Server and GreenPlum DB server.
6. Configure /etc/hosts entries on SAS Server for GreenPlum database server and on GreenPlum server for SAS server(s). This will be required on the SAS server to use the LIBNAME statement.
7. If SAS is on Linux, configure environment variables:
   1. ODBCHOME (points to SAS Access installation directory)
   2. ODBCINI=$ODBCHOME/odbc.ini.
   3. LD\_LIBRARY\_PATH=$LD\_LIBRARY\_PATH/$ODBCHOME/lib
8. On the GreenPlum database server, configure pg\_hba.conf to allow trusted local and host connections for the SAS server(s).
9. On the GreenPlum database server, grant privileges to the user id in the odbc.ini file (if not gpadmin). Recommendation would be to have a SAS user with access to a specific database/schema for reading data and a “sandbox” database/schema/workspace for SAS users to create their own temporary data sets.

NOTE: do we need to mention that SAS can be configured with a direct connect option in the LIBNAME or Pass-Through sections? While ODBC is certainly the easiest way to connect, direct may offer better performance and flexibility in connection that using ODBC and eliminate having to install drivers on a lot of machines.

## ODBC driver configuration

This section explains the key configuration settings in the “odbc.ini” file that need to be modified to validate connectivity between the SAS server and the GreenPlum database. A DSN will need to be configured for GreenPlum, regardless of whether this is a Windows or Linux environment. The DSN in our example is called “gplum”. Thee “driver” parameter points to the ODBC driver lib. This is the library that the driver manager will dynamically load when SQLConnect or SQLDriverConnect is called for that DSN. If this points to the wrong place, the DSN will not work.

The “Description” parameter” documents the product we are using to access the GPDB. You will also need to provide the following parameters: database name (Database), local hostname or IP address of the GreenPlum database master server (HostName), GreenPlum User Id (logonID), password for the GPDB user (Password), and the port number GreenPlum is listening to (PortNumber). This last parameter is typically 5432, but my vary in your environment.

The appropriate odbc.ini DSN section for a Linux environment is shown below:

[gplum]

Driver=/GPodbc/lib/S0gplm60.so

Description=SAS ACCESS to Greenplum

Database=sastest

HostName=localhost.localdomain

LogonId=gpadmin

Password=changeme

PortNumber=5432

**“odbc.ini parameters**

In Windows, you can test the connectivity of the configured ODBC DSN by pressing the “Test Connectivity” button on the last configuration screen. (need windows screen shots for ODBC set up)

In Linux, execute the “demodbc” program located in <odbc home>/demo directory. The command syntax is:

**Demodbc [-uid <username>] [-pwd <password>] [-xml] <datasource/DSN name>**

This script attempts to connect to a table named “emp”, which does not exist in the default installation of GPDB. You will receive an error but this indicates the database was connected to even though the database was not found.

**Connectivity Test**

This section presents how to establish connectivity between SAS and the GreenPlum database. SAS/ACCESS provides two methods of relational database interface:

* “Pass-Through” SQL – direct connection to the database and execution of SQL statements.
* Named Library (“LIBNAME”) – reusable connection parameters and database access via a SAS procedure (PROC)

Some background is required to understand these two approaches. It is important to understand the relative similarities and differences between PROC (procedural) SQL and DATA step (libname) programming. While SQL was developed to talk to relational databases, SAS was developed to manage and analyze flat files. The following table shows equivalence of the terms used in SAS and SQL.

**Terms Mapping for SAS and SQL Equivalences**

|  |  |
| --- | --- |
| **SAS** | **SQL** |
| Data Set | Table |
| Observations | Rows |
| Variables | Columns |
| Merge | Join |
| Extract | Query |

Both Pass-Through and LIBNAME statements can be executed from the command line (sas <filename.sas>). Output will be stored in a <filename>.lst file and errors will be logged to a <filename>.log file. (Have to confirm this for SAS 9.3….) In SAS 9.3, output options were changed to have all data written to a browser as the default, so the output will be in a .htm file. Adding the following lines at the head of your script (.sas) file will emulate the 9.2 behavior:

ods graphics off;

ods html close;

ods listing;

To use the SAS environment, either start SAS Enterprise Menu from the Start Menu in Windows or, using an X-Term window, start sas on a Linux system with sas &. This will open a set of windows. The Query Window (verify!) will be where you enter the procedures listed below. Using the three commands above, output will appear in an Output window in SAS, while errors will appear in a sub-frame of the Query window. If you do not enter the commands listed above, query results will be returned to a browser window.

**Connectivity Test Using Pass-Through SQL**

Pass-Through uses SAS/Access to connect to the database and to send statements directly for execution. This will be an easier test for those familiar with SQL as it uses standard SQL to communicate with the database. A SAS SQL procedure does the following:

1. Establish a connection with the database using a CONNECT statement and terminates the connection with a DISCONNECT statement.
2. Once the connection is established, the EXECUTE statement can be used for non-query statements (i.e., INSERT, UPDATE, DELETE) or data can be retrieved with a SELECT statement using the CONNECTION TO component in the FROM clause of the statement.

The format of Pass-Through SQL is basically this:

PROC SQL;

# required statement for SAS

CONNECT TO greenplm AS mydb (DSN info)

# “mydb” can be any identifier – used to reference the database connection

# “greenplm” is a REQUIRED identifier for SAS/ACCES to talk to GPDB

%PUT &sqlxmsg;

# SAS syntax to capture SQL error messages in a variable

EXECUTE (INSERT/UPDATE/DELETE) BY mydb

# Note the reference to the database connection “mydb”

# OR for SELECTS

SELECT \* (or <field list>) FROM mydb

(SELECT \* (or <field list>) from <table>);

%PUT &salxmsg;

# Grab any error text from the query

DISCONNECT from mydb;

QUIT;

Note that SAS SELECT’s in Pass-Through mode look very much like a select from a sub-query. If you don’t write it this way, the SELECT will execute but no data will be returned to the SAS output screen. It can almost be thought of as: SAS is reading from a pipe (the database connection parameter) which is executing the actual SELECT on the database.

Example Pass-Through procedures are below:

<insert scripts and screenshots here>

**Connectivity Testing Using SAS Native Connections (LIBNAME)**

The LIBNAME statement extends the SAS global LIBNAME statement to enable you to assign a “libref” to GPDB. This lets you reference a GPDB object directly in a DATA step or SAS procedure. You can associate a SAS libref with a database, schema, server, or group of tables and views. This approach is much more likely to be familiar to existing SAS users.

A libref is a shortcut or nickname for the database where the tables and views are stored. The libref can be any SAS name with the limitation that name can be no longer than eight characters long. The libref contains three required options:

* Libname <libref name>
* Engine Name: this must be “greenplm”
* Connection Options: database/user/DSN information

An example of a LIBNAME would be:

**Libname mydb greenplm uid=gpadmin pwd=changeme dsn=gplum**

Where the libname is “mydb”, the engine name is “greenplm” and the connection options specify the user, passwd, and the ODBC DSN containing the database/schema information. An ODBC DSN does not have to be used; you can specifiy a direct connection option with the database hostname/IP address, a port option, a database option and, optionally, a schema option.

When a LIBNAME is created, it will be stored in a SAS Library for reuse. The LIBNAME can be de-refrenced within a calling script/procedure with this syntax:

**Libname mydb CLEAR;**

Some examples of LIBNAME scripts are shown below:

<insert procedures and screenshots here>

# Bulk Loading with SAS and GreenPlum

The SAS bulk-loading facility allows the import of data to SAS and the export of data from SAS. This has been well documented in the white paper **BULKLOADING USING GREENPLUM’s GPFDIST IN A SAS PROGRAM.** There are two addendum’s to that white paper that need to be mentioned here.

1. GreenPlum can import an exported SAS Data Set (file extension: sas7bcat) which is a compressed file exported by SAS. When called to import this, SAS will unpack the file and convert it to a pipe-delimited format suitable for loading into the GreenPlum database with gpfdist.
2. For truly high-speed loading of files with GPFDIST, the parameter BL\_USE\_PIPE=YES should be included in the parameter section of the SAS procedure, particularly if importing exported SAS data sets. As SAS unpacks the file, it will send the pipe-delimited data to a named pipe being read by gpfdist.

This is what a procedure would look like in SAS to used named pipes:

proc sql;

create table mydblib.emp (

bulkload=YES

bl\_host=SASENV /\*replace with the hostname where\*/

/\* gpfdist is running \*/

bl\_port='8085'

bl\_protocol='gpfdist'

**bl\_use\_pipe=YES**)

as select \* from Sasuser.emp;

quit;

There is an important caveat to this last item: this only works on systems that support named pipes (i.e., Linux). The BL\_USE\_PIPE option is NOT currently supported by SAS on Windows systems. If put in a procedure on a Windows system, the procedure will return an error that this parameter is not currently supported on this platform. If taken out of the procedure, the procedure will execute, but it will first unpack the SAS data to a pipe-delimited ASCII text file in the directory specified by the environment variable **GPLOADHOME** ( need to verify I have this environment variable name correct ), then will pass this file to gpfdist for loading. This can be a VERY SLOW process if the data set to be imported is very large (> 1 GB).

# Best Practices

This section describes a number of considerations and best practices for optimizing a SAP® BusinessObjects™ and Greenplum database environment.

I need some help on this section. A couple of thoughts I had were:

Know your users. Long-time SAS users may prefer the LIBNAME/Data Step procedures listed here; users used to writing SQL may prefer Pass-Through.

Run bulkloads on Linux systems.

Give the SAS users a “sandbox” database/schema fully readable/writable by SAS users as a place for them to create and work on their own data sets when using GreenPlum database.

Direct connect versus ODBC? What is best for users? LIBNAME’s can be re-used for connectivity, but there are security implications in putting a user/password in a library that can be read by anyone, Can the user and password be parameterized so that when called users put in their own userid/password? ODBC is going to involve driver installation on (potentially) a lot of user machines but may be slightly more secure.

Other thoughts on what we should recommend here?

# Conclusion

SAP® BusinessObjects™ is one of the more ubiquitous data integration and business intelligence platforms available today.

In this white paper, processes for installing, configuring, testing, and troubleshooting SAP® BusinessObjects™ in a Greenplum environment were discussed.

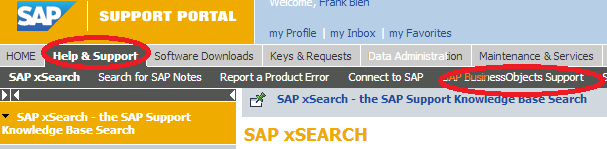
# References

1. SAP Business Objects 4.0 prerequisites

<https://websmp105.sap-ag.de/~sapidb/011000358700001237042010E/xi4_install_win_en.pdf>

[SAP login required]

1. SAP Documentation



<https://websmp202.sap-ag.de/support-welcome>

[SAP login required]

1. General reference SAP information

<http://www.scribd.com/doc/91155588/Features-of-SAP-Business-Objects-4>

1. Business Objects installation, configuration and running reports

<http://guozspace.wordpress.com/2012/04/28/sap-business-objects-bi-4-0-installation-steps/>

1. Best practices for SAP BOBJ upgrade

<http://wiki.sdn.sap.com/wiki/display/BOBJ/BI+4.0+SP04+Update+-+Best+practices%2C+Known+Issues+and+Possible+Solutions>

1. EMC Greenplum Database, DCA, and DIA Documentation

<http://powerlink.emc.com>

[EMC Power Link login required]

1. Greenplum internal Confluence page on ODBC drivers (32-bit versus 64-bit)

<http://confluence.greenplum.com/display/~mekas1/ODBC+32-bit+and+64-bit+windows+registry+settings>

[Greenplum internal login required]