Creating new database and running ETL(Kettle) Job

Below are the steps to create a new database. The steps are after installing and setting up the Greenplum (Community edition or enterprise edition). And all the steps below are for Linux environment.

1. Login to the GP database and create a new empty database using the below command.  
   create database database\_name;
2. Copy the ETL folder from GitHub to /home/gpadmin/ folder.
3. Modify the parameter DB\_NAME in home/gpadmin/ETL/CREATE\_NEW\_DATABASE/Checkbook\_DB\_SetUp.sh file to the database name created in the 1st step.
4. Run the below command from home/gpadmin/ETL/CREATE\_NEW\_DATABASE/ folder to create tables, procedures, initial reference/static data and Trends data. After running the below command verify if the database is successfully created or not.  
   nohup sh Checkbook\_DB\_SetUp.sh &
5. If the ETL folder from GitHub is not copied to /home/gpadmin then replace /home/gpadmin in the below files to the folder where the ETL folder is copied.  
   a) ETL/CREATE\_NEW\_DATABASE/Checkbook\_DB\_SetUp.sh  
   b) ETL/CREATE\_NEW\_DATABASE/Scripts.sql  
   c) ETL/CREATE\_NEW\_DATABASE/ScriptsForReferenceTables.sql

d) ETL/CREATE\_NEW\_DATABASE/Trends.sql

e) ETL/KETTLE\_JOB/PreProcessing\_DataFiles/master\_preprocess.sh

f) ETL/KETTLE\_JOB/Solr/getSolrCount.sh

Below are the steps to process test data using the ETL (Pentaho Kettle) Job.

1. Download the pentaho data integration from the below link in the same server where the Greenplum is installed. Kettle requires the Sun Java Runtime Environment (JRE) version 1.5 <http://sourceforge.net/projects/pentaho/files/Data%20Integration/4.2.0-stable/>
2. Extract the tar file into folder of your choice, say /usr/local/ folder. It will create a folder called data-integration. So now the path of the kettle software is /usr/local/data-integration
3. Copy the ETL/.kettle folder to the system user’s home directory.
4. Modify the below parameters with the correct values in the user\_home\_directory/.kettle/kettle.properties file  
     
   NYC\_EMAIL\_SERVER=MAIL\_SERVER\_NAME\_HERE

NYC\_EMAIL\_PORT=25

NYC\_EMAIL\_FROM=FROM\_EMAIL\_ID\_HERE

NYC\_EMAIL\_TO=TO\_EMAIL\_ID\_HERE

NYC\_EMAIL\_CC=CC\_EMAIL\_ID\_HERE

NYC\_FMS\_GROUP\_EMAIL\_CC=CC\_EMAIL\_ID\_HERE

NYC\_OASIS\_GROUP\_EMAIL\_CC=CC\_EMAIL\_ID\_HERE

CHECKBOOK\_DB\_HOST=DB\_HOSTNAME\_HERE

CHECKBOOK\_DB\_USER=DB\_USER\_HERE

CHECKBOOK\_DB\_PASS=DB\_PASSWORD\_HERE

CHECKBOOK\_DB\_NAME=DB\_NAME\_HERE

CB\_FILES\_SOURCE\_DIR=/home/gpadmin/ETL/SOURCE\_DATA/

CB\_FILES\_DEST\_DIR=/home/gpadmin/ETL/DEST/

CB\_FILES\_GPDIST\_DIR=/home/gpadmin/ETL/GPFDIST/datafiles/

CB\_FILES\_BACKUP\_DIR=/home/gpadmin/ETL/BACKUP/

1. Modify the below solr properties and ETL/KETTLE\_JOB/Solr/getSolrCount.sh file with the correct hostname, port and solrCoreName values  
     
   NYC\_SOLR\_DELETE\_PENDING\_CONTRACTS\_DEV=http://hostname:port/solrCoreName/update?stream.body=%3Cdelete%3E%3Cquery%3Econtract\_status:pending%3C/query%3E%3C/delete%3E&commit=true

NYC\_SOLR\_DELETE\_RECORDS\_DEV=http://hostname:port/solrCoreName/dataimport?command=delta-import&clean=false&jobID=replace\_job\_id

NYC\_SOLR\_INCREMENTAL\_INDEXING\_DEV=http://hostname:port/solrCoreName/dataimport?command=full-import&clean=false&jobID=replace\_job\_id

NYC\_SOLR\_FULL\_INDEXING\_DEV=http://hostname:port/solrCoreName/dataimport?command=full-import&clean=true&jobID=0

NYC\_SOLR\_CHECK\_INDEX\_STATUS\_DEV=http://hostname:port/solrCoreName/dataimport/

NYC\_SOLR\_RECORDS\_COUNT\_DEV=http://hostname:port/solrCoreName/select/?q=\*%3A\*&version=2.2&start=0&rows=10&indent=on

1. Make sure gpfdist utility is running on 8081 port and pointing to /home/gpadmin/ETL/GPFDIST/ directory as below.  
   gpfdist -d /home/gpadmin/ETL/GPFDIST/ -p 8081 -l /home/gpadmin/ETL/GPFDIST/log
2. The source files that needs to be processed by ETL job should be in the folder home/gpadmin/ETL/SOURCE\_DATA/. And the filenames should match with actual\_pattern column of etl.ref\_file\_name\_pattern table.
3. The layout of the files should be based on the file layout documents of ETL\Checkbook\_File\_Layouts\Checkbook\_File\_Layouts \_20120717 folder.
4. Below is the command to be run for processing the test files that are in home/gpadmin/ETL/SOURCE\_DATA/ directory .  
   /usr/bin/nohup /bin/sh /usr/local/data-integration/kitchen.sh -file /home/gpadmin/ETL/KETTLE\_JOB/NYC\_Master\_OS -norep -level:Minimal > /usr/local/data-integration/NYCETLMaster.out &