Creating GP databases and running ETL (Kettle) Jobs to load data.

Below are the steps to create new databases. 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. Copy the source/database/GREENPLUM folder from GitHub to /home/gpadmin/ folder.
2. If the source/database/GREENPLUM folder from GitHub is not copied to /home/gpadmin then replace /home/gpadmin in the below files to the folder where the GREENPLUM folder is copied.  
   a) GREENPLUM/Checkbook/CREATE\_NEW\_DATABASE/Checkbook\_DB\_SetUp.sh  
   b) GREENPLUM/Checkbook/CREATE\_NEW\_DATABASE/Scripts.sql  
   c) GREENPLUM/Checkbook/CREATE\_NEW\_DATABASE/ScriptsForReferenceTables.sql

d) GREENPLUM/Checkbook/CREATE\_NEW\_DATABASE/Trends.sql

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

f) GREENPLUM/Checkbook/KETTLE\_JOB/Solr/getSolrCount.sh

1. GREENPLUM/Checkbook\_ogent/CREATE\_NEW\_DATABASE/Checkbook\_EDC\_DB\_SetUp.sh
2. GREENPLUM/Checkbook\_ogent/CREATE\_NEW\_DATABASE/Scripts.sql
3. GREENPLUM/Checkbook\_ogent/CREATE\_NEW\_DATABASE/ScriptsForReferenceTables.sql

j) GREENPLUM/Checkbook\_ogent/KETTLE\_JOB/PreProcessing\_DataFiles/master\_preprocess.sh

k) GREENPLUM/Checkbook\_ogent/KETTLE\_JOB/Solr/getSolrCount.sh  
l) GREENPLUM/Checkbook\_ogent/KETTLE\_JOB/DUMP\_AND\_RESTORE/get\_oge\_contracts\_and\_restore.sh

m) GREENPLUM/Checkbook\_ogent/KETTLE\_JOB/DUMP\_AND\_RESTORE/get\_oge\_fms\_data\_dump\_and\_restore.sh

n) GREENPLUM/Checkbook\_ogent/KETTLE\_JOB/DUMP\_AND\_RESTORE/verify\_errors.sh

1. Login to the GP database and create new database1 using the below command.  
    create database database\_name1;
2. Modify the parameter DB\_NAME in home/gpadmin/GREENPLUM/Checkbook/CREATE\_NEW\_DATABASE/Checkbook\_DB\_SetUp.sh file to the database name (database\_name1) created in the above (3rd) step.
3. Run the below command from home/gpadmin/GREENPLUM/Checkbook/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 &
4. Login to the GP database and create new database2 using the below command.  
    create database database\_name2;
5. Modify the parameter DB\_NAME in home/gpadmin/GREENPLUM/Checkbook\_ogent/CREATE\_NEW\_DATABASE/Checkbook\_EDC\_DB\_SetUp.sh file to the database name (database\_name2) created in the above (6th) step.
6. Run the below command from home/gpadmin/GREENPLUM/Checkbook\_ogent/CREATE\_NEW\_DATABASE/ folder to create tables, procedures, initial reference/static data. After running the below command verify if the database is successfully created or not.  
   nohup sh Checkbook\_EDC\_DB\_SetUp.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 GREENPLUM/.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 (database\_name2)

CHECKBOOK\_DB\_NAME\_OGE=DB\_NAME\_HERE (database\_name2)

CB\_FILES\_SOURCE\_DIR=/home/gpadmin/GREENPLUM/Checkbook/SOURCE\_DATA/

CB\_FILES\_DEST\_DIR=/home/gpadmin/GREENPLUM/Checkbook/DEST/

CB\_FILES\_GPDIST\_DIR=/home/gpadmin/GREENPLUM/Checkbook/GPFDIST/datafiles/

CB\_FILES\_BACKUP\_DIR=/home/gpadmin/GREENPLUM/Checkbook/BACKUP/

CB\_FILES\_SOURCE\_DIR\_OGE=/home/gpadmin/GREENPLUM/Checkbook\_ogent/SOURCE\_DATA/

CB\_FILES\_DEST\_DIR\_OGE=/home/gpadmin/GREENPLUM/Checkbook\_ogent/DEST/

CB\_FILES\_GPDIST\_DIR\_OGE=/home/gpadmin/GREENPLUM/Checkbook\_ogent/GPFDIST/datafiles/

CB\_FILES\_BACKUP\_DIR\_OGE=/home/gpadmin/GREENPLUM/Checkbook\_ogent/BACKUP/

1. Modify the below solr properties   
   NYC\_SOLR\_DELETE\_PENDING\_CONTRACTS=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=http://hostname:port/solrCoreName/dataimport?command=delta-import&clean=false&jobID=replace\_job\_id

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

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

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

NYC\_SOLR\_RECORDS\_COUNT=http://hostname:port/solrCoreName/select/?q=\*%3A\*&version=2.2&start=0&rows=10&indent=on  
NYC\_SOLR\_DELETE\_RECORDS\_OGE=http://hostname:port/solrCoreName/update?stream.body=%3Cdelete%3E%3Cquery%3Eagency\_type:oge%3C/query%3E%3C/delete%3E&commit=true

NYC\_SOLR\_FULL\_INDEXING\_OGE=http://hostname:port/solrCoreName/dataimport?command=full-import&entity=contracts\_oge&entity=spending\_oge&clean=false&jobID=0

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

1. Modify GREENPLUM/Checkbook/KETTLE\_JOB/Solr/getSolrCount.sh and GREENPLUM/Checkbook\_ogent/KETTLE\_JOB/Solr/getSolrCount.sh files with the correct hostname, port and solrCoreName values.
2. Modify GREENPLUM/Checkbook\_ogent/KETTLE\_JOB/DUMP\_AND\_RESTORE/get\_oge\_contracts\_and\_restore.sh and GREENPLUM/Checkbook\_ogent/KETTLE\_JOB/DUMP\_AND\_RESTORE/get\_oge\_fms\_data\_dump\_and\_restore.sh files with the correct database name and connection parameters.
3. Make sure gpfdist utility is running on 8081 and 8082 ports as below.   
   gpfdist -d /home/gpadmin/GREENPLUM/Checkbook/GPFDIST/ -p 8081 -l /home/gpadmin/GREENPLUM/Checkbook/GPFDIST/log

gpfdist -d /home/gpadmin/GREENPLUM/Checkbook\_ogent/GPFDIST/ -p 8082 -l /home/gpadmin/GREENPLUM/Checkbook\_ogent/GPFDIST/log

1. The source files that needs to be processed by ETL jobs should be in the below folders. And the filenames should match with actual\_pattern column of etl.ref\_file\_name\_pattern table  
    home/gpadmin/GREENPLUM/Checkbook/SOURCE\_DATA/ 🡪 To process test files in database\_name1  
   home/gpadmin/GREENPLUM/Checkbook\_ogent/SOURCE\_DATA/ 🡪 To process test files in database\_name2
2. The layout of the files should be based on the file layout documents of GREENPLUM/Checkbook/Checkbook\_File\_Layouts/Checkbook\_File\_Layouts \_20120717 folder.
3. Below is the command to be run for processing the test files that are in home/gpadmin/GREENPLUM/Checkbook/SOURCE\_DATA/ directory to database\_name1.   
   /usr/bin/nohup /bin/sh /home/gpadmin/GREENPLUM/Checkbook/KETTLE\_JOB/process\_data\_using\_etl\_for\_database\_name1.sh &
4. Below is the command to be run for processing the test files that are in home/gpadmin/GREENPLUM/Checkbook\_ogent/SOURCE\_DATA/ directory to database\_name2.

/usr/bin/nohup /bin/sh /home/gpadmin/GREENPLUM/Checkbook\_ogent/KETTLE\_JOB/process\_data\_using\_etl\_for\_database\_name2.sh &