Rancho BioSciences

tranSMART GWAS data ETL Guide

T 1-844-4-RANCHO or T 1-760-717-7881 | www.RanchoBioSciences.com

Database changes (run only once)

1) Install kettle. Make sure, you can use sqlldr program. Unzip dump files.
2) Check all `*.sh` scripts in package folder and correct PASS, HOST and SID (skip ETL folder).
3) Unarchive *.7z files with dump data
4) Run 1.drops_n_dumps.sh (will take some time).
5) Run 2.grants.sh.
6) Update stored procedures (user tm_cz) i2b2_load_eqtl_top50, i2b2_load_gwas_top50 and i2b2_move_analysis_to_prod via sqldeveloper (sqls attached).

Uploading data

1) Update script ETL/load_analysis_batch.sh to correspond your system and dataset;

   Important: use full, NOT relative, paths.

   Check:
   a) kettle (`kitchen.sh` file) location
   b) path to `process_analysis_files.kjb` file
   c) gwasData location (sample gwasData attached)
   d) temp directory
   e) sqlldr path (usually $ORACLE_HOME/bin/sqlldr)

2) Update script nightly_processing.sh.

   Check:
   a) kettle (kitchen.sh file) location
   b) nightly_processing.kjb file location

3) Put kettle.properties (can be found in attached _kettle folder) to ~/.kettle. Update it for your database.
4) Unzip mod_MAGIC_2hrGlucose_AdjustedForBMI.tsv (can be found in `gwasData` folder); update `Load_GWAS_file_batch.ktr` kettle job in ETL folder, correct tsv file location in `Read GWAS File` step.
5) Update RASS and ETL_DIR (path to attached ETL folder, where load_analysis_bath.sh is stored) properties in `3.loader.sh` and run it. Check for table tm_cz.cz_job_audit to find out, if upload finished successfully.
6) Check 4.fmapp.sql to correspond your dataset (configured for MAGIC) and run 4.fmapp.sh.

Transmart changes

1) GwasController, function `createSOLRQueryPath`, set return to

   return "http://<SOLR_IP>:<SOLR_PORT>/solr/rwg/select?q=*".

   Set <SOLR_IP> and <SOLR_PORT> to correspond your solr installation.

SOLR configuraration

Please make sure that you configured SOLR for GWAS data.