GenoMetric Query Language (GMQL)
Quick Start

Genomic Computing Group

Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano

September 21, 2015

1 GMQL Working Modes

GMQL has two working modes that differ in processing capabilities.

1. **MAPREDUCE mode**
   The power of GMQL is its ability to run complex, big data queries on huge input datasets in a cluster of machines. In the `MAPREDUCE` mode, GMQL uses the cloud computing Hadoop Mapreduce engine for processing and the Hadoop Distributed File System (HDFS) to distribute and maintain the data in a cluster. To run it, an Hadoop installation is required.

2. **LOCAL mode**
   In the `LOCAL` mode the GMQL toolkit uses only the resources (RAM, CPU and storage) available on the computer that runs GMQL; it works on the local file system (LFS) for all file operations, without duplicating or distributing any sample data file. The `LOCAL` mode allows easy installation and testing of the toolkit, but it has limited performance, sufficient just for processing a limited data size.

2 Dependencies

JAVA JDK 7.

- You can download JDK 7 from:
  `http://www.oracle.com/technetwork/java/javase/downloads/`
- Untar the java package into a specific folder, for example:

  ```
  /home/user1/java7/
  ```

- Then, add the following lines to the path:

  ```
  export JAVA_HOME=/home/user1/Java7/jdk1.7.0_25/
  export PATH=/home/user1/Java7/jdk1.7.0_25/bin:$PATH
  ```

Racket v5.3 or later.

- You can download Racket from:
  `http://mirror.racket-lang.org/installers/5.3/racket/racket-5.3-bin-x86_64-linux-debian-squeeze.sh`
- Then, add the following lines to the path:

  ```
      export RACKET_HOME=/racket/folder
      export PATH=$PATH:$RACKET_HOME/bin/
  ```

Hadoop (required only for `MAPREDUCE mode`).

- For Hadoop V 1.x see for example:
  `http://hadoop.apache.org/docs/r1.2.1/single_node_setup.html`
- For Hadoop V 2.x (Yarn) see for example:
  `http://hadoop.apache.org/docs/stable/`

3 Setting the Environment Variables - `LOCAL` mode

To use the default values for the environment variables, append the content of the file `GMQLPackage/conf/GMQL-env.sh`, located in the `conf/` folder, to the `~/.bashrc` or `~/.bash_profile` file, located in the home folder of each user who will use GMQL. Do not run `conf/`.[1]

To change the default values of the environment variables:

- Set the directory path for java JDK (the default is: `/usr/lib/jvm/java-7-oracle/`):

  ```
      export JAVA_HOME=/home/user1/Java7/jdk1.7.0_25/
  ```

- Set the GMQL home, i.e. the location where all the local data of the GMQL repository, as well as the control and configuration data, are located (the default is: `/home/yourUserName/gmql_repository`):

[1] Before running the installation, make sure to set the configurations in both GMQL-env.sh and .bashrc and not one of them

```
export GMQL_HOME=~/gmql_repository
```

- Set the environment variables for the Apache Pig installation, where Apache Pig V0.15.0 is automatically installed (the default is: `/userHomeFolder/pig`):

  ```
  export PIG_HOME=~/pig
  ```

  To learn more about Pig environment variables, see the guide at: `http://pig.apache.org/docs/r0.15.0/start.html`.

- Set the execution mode (the default is: `LOCAL`):

  ```
  export GMQL_EXEC=LOCAL
  ```

4 Setting the Environment Variables - `MAPREDUCE` mode

- Besides setting the execution mode to `MAPREDUCE` (i.e. `export GMQL_EXEC=MAPREDUCE`), set the following environment variables [2]

  ```
  export HADOOP_HOME=/usr/local/hadoop
  export HADOOP_COMMON_HOME=$HADOOP_HOME
  export HADOOP_CONF_DIR=$HADOOP_HOME/conf
  export HADOOP_CLASSPATH=$GMQL_HOME/utils/lib/*:$HADOOP_CLASSPATH
  ```
  [3] [4]

- Set the GMQL home on the HDFS, where storing the sample data (the default is: `/user/`; note that here we mean exactly ”user”, not the user name):

  ```
  export GMQL_DFS_HOME=/user/
  ```

- Add the variables to the system path:

  ```
  export PATH=$PATH:$JAVA_HOME/bin:$
            HADOOP_HOME/bin:$PIG_HOME/bin:$GMQL_HOME/bin
  ```

- Finally, Open the file `GMQLPackage/GMQL/gmqlc/configurations.rkt`, change the value in line 16 `hdfs://localhost:9000/` to be equal to Hadoop configuration valiable `fs.defaultFS` value in `core-site.xml`

[2] This environment variables should be set for all the users of the system.

[3] Make sure to set the configurations directory.

[4] Class not found exception might be raised in case of not setting HADOOP“˙CLASSPATH properly.

5 Installing GMQL

1. Go to the folder `.../GMQLPackage/`
2. Run the installer, i.e. `./install.sh`
   and follow the instructions on the screen to install GMQL.
3. Run the following command to register your user to the GMQL repository:

   ```
   repositoryManagerV1 RegisterUser
   ```

   In case of multi-users of the GMQL system, each user must run this command from his/her environment.

   The output of this command should look like:

   ```
   INFO:  Local Folders Creation ...
   INFO:  Folder, /home/gql_repository/data/username/indexes/ true
   INFO:  Folder, /home/gql_repository/data/username/datasets/ true
   INFO:  Folder, /home/gql_repository/data/username/metadata/ true
   INFO:  Folder, /home/gql_repository/data/username/schema/ true
   INFO:  Folder, /home/gql_repository/data/username/results/ true
   INFO:  Folder, /home/gql_repository/data/username/queries/ true
   INFO:  Folders are created
   ```