Antonin VILLEMIN (2017EE17080) - Quentin BOURGUE(2017EE17079)

COL362 - Assignment 1

Database Design

Introduction

In this part of the assignment, we will try to compare different methods to insert a large amount of data.

Approaches

(a) Bulk load :

Create a .csv file with all the data necessary for the 500 000 queries. A single request using the command COPY of postgrest is finally executed in the SQL shell. This request provides the data to the database from a .csv file. To generate this file, we used a python script, included in the folder code, with the sql request and the .csv file. **The average time for this approach is 11s.**

(b) Insert statements for each tuple :

Create a .sql file with 500 000 queries by using a python script. The SQL shell will execute all the requests one-by-one. **On average, this approach needs 56s.**

(c) Programmatically, using JDBC :

We use JDBC to interact with the database through a java code. In this code, we do a loop FOR and at each iteration, we do a new INSERT in the database. Data are generated before this loop, to avoid the generation of random data to modify the measured time. **On average, this approach needs 206s.**

All the averages were made thanks to 4 different measures.

Observations

In our case, the fastest approach is the bulk load. Indeed, this method executes only one request with all the data. This request has been particularly made in order to INSERT a large amount of data. It adds all the data in a row. So, it seems logical that it is the fastest approach whereas the second approach runs 500 000 queries. Whenever a query is made, the program comes back to the sql file to see what is the next query and execute it. Finally, the method using JDBC is the slowest one. It seems also logical because all the requests are made in Java even if the database works in SQL. So, in the loop FOR, the java code will use the JDBC driver to transmit a SQL request which has been made thanks to a java string. After this request realized, the program has to come back to the java code to do the next instruction.

Conclusion

Finally, the bulk load is the fastest approach principally because it use a command designed in order to insert a large amount of data. The slowest one is the one using JDBC. JDBC is a good way to manipulate database but has to interact with the database in sql. The conversion between the two languages slows the program. Of course it was possible to do a bulk load but it wasn’t the goal of the assignment.