Institut Polytechnique de Paris
DataAI

Professor: Ioana Manolescu
Teaching assistant: Madhulika Mohanty[1]

Architectures for Big Data: Lab 3 - AWS

January 8, 2024

Submission deadline: Monday, January 22, 2024 at 23:55:00 (Paris time).

1 Lab organization[2]

1.1 Cloud setup

We will be using the Free Tier by Amazon Cloud for this assignment. Please set up an account here: `https://portal.aws.amazon.com/billing/signup#/start`. Note that you would need to provide a credit/debit card details but it will not be charged as long as you use the free tier service.

1.2 Setting up Amazon Cloud RDS

We will be using the Amazon Cloud RDS (Relational Database Service). Start with setting up a MySQL database instance for use from under the "Databases" tab here:`https://console.aws.amazon.com/rds/home?p=rdsms&cp=bn&ad=c`. Choose the Free Tier, enable Public Access and select a custom VPC security group.

You will find general instructions on setting up the DB and accessing it here: `https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Welcome.html`. The exact set of instructions to be used for this assignment can be found in the slides provided for this lab.

Note: You will have to select only permitted instances with 20GB space to be able to use the Free tier. More details on what is available under Free Tier can be found here: `https://aws.amazon.com/rds/free/`. The number of database instances that you create defines the consumption factor for your free hours, so create new database instances with caution. For example, if you have x free hours for one database instance; you will have only $\frac{x}{2}$ hours with 2 database instances and so on.

2 Assignment tasks

You will use the **synthetic, relational data**, part of the **TPC-H** benchmark. The goal of the lab session is to load, run and measure the TPC-H query benchmark on MySQL in Amazon Cloud.

Task 1 You will find the details of the benchmark here: `http://tpc.org/tpc_documents_current_versions/current_specifications5.asp`. TPC data can be generated in various sizes ("scale factors"). For this assignment, we will work with scale factors 0.1 and 1 and the datasets can be downloaded from Moodle under `Lab-3 files`. Basically, the scaling factor 1 dataset is 10 times as large as the dataset with scaling factor 0.1.

You are now ready to use the cloud. Connect to the cloud database created before using the instructions on the slides.

We will now load the datasets, starting with scaling factor 0.1.

Hint: Try to perform all the tasks with a scaling factor of 0.1 before moving on to 1. Delete the previous database before loading another dataset. This will help save cloud space.

[1] madhulika.mohanty@inria.fr

[2] This lab material is based on the previous edition of the course (with changes) courtesy of Ioana Manolescu, Angelos Anadiotis and Pawel Guzewicz.

1. The schema of the dataset can be found in the file named `dss.ddl`. Run this file which has a sequence of SQL commands to create the tables.

2. The integrity constraints and indexes are defined in the file `dss.ri`. Run this file to build them.

3. Follow the instructions for loading the data from a file on MySQL here: `https://dev.mysql.com/doc/refman/8.0/en/load-data.html` to populate the tables on your database instance. The data for each table is in a file named as ⟨*table-name*⟩*.tbl*.

   Hint: Since there are integrity constraints on the tables, there is a particular order in which the tables are to be loaded. You can infer this order by looking at the constraints file `dss.ri`. In order to avoid access restrictions, navigate to the directory containing the table files and launch the mysql client from there.

Task 2 Now, we will run the benchmark queries on the cloud. You will find them in a folder titled `queries`.

1. For each TPC-H scale factor, measure the running time of every query.

2. Gather the results in a graphic(plot) using a spreadsheet, a notebook etc.

   - For the 0.1 scale factor dataset, this would give one graph with each query on the x axis, and the running time on the y axis.
   - Increase the scale factor to 1, re-run the queries on it and discuss the scalability of every query with respect to its query plan. You can use the `EXPLAIN` command (`https://dev.mysql.com/doc/refman/8.0/en/explain.html`) to get information about the query plan.

3. How does the system scale for the two datasets? Discuss the trend you observe amongst various queries and why.

Task 3 Repeat the same task on MySQL on your local machine. How does this compare with the cloud? Report your observations.

3 Deliverables:

You will need to deliver :

- **Code:** anything you had to write yourself (any scripts for running queries, generating graphs, etc.)
- **Report:** A report on the outcome of the tasks in the assignment. It should include a structured description of the steps you used to load the data, query it in the respective setting, and noting the runtimes. In particular, any details characterizing what you used (version of every software, virtual machine configuration, indexes if any, etc.) must be present in the report. It should also have the results and conclusions of the scalability analysis.

3.1 Submission guidelines

Please follow the submission rules and guidelines available at Moodle for Lab-1.