VE472 — Methods and tools for big data

Homework 1

Manuel — UM-JI (Summer 2021)

Reminders

- Write in a neat and legible handwriting or use LaTeX
- Clearly explain the reasoning process
- Write in a complete style (subject, verb, and object)
- Be critical on your results

Ex. 1 — *Processes, cgroups, and namespaces*

1. Write a short summary describing what `cgroups` are.
2. Explain the differences and similarities between `cgroups` and processes in Linux.
3. How does kernel `namespace` increase the security of the OS?

Ex. 2 — *Increasingly large dataset*

Retrieve the flight data from the course server at focs.ji.sjtu.edu.cn, port 2572, using login `motivatedstudent` and password `IwanttoworkhardinVE472`. For each of the questions below.

Task to be completed for each question

Study the year 1987, and then keep increasing the size of the dataset by adding more years until your computer cannot handle it anymore. Monitor its RAM and CPU usage as the size of the dataset increases. Then when answering a question include a graph showing how the dataset size impacts the time and memory necessary to complete the request. On the graph highlight the point where the tasks switches from "compute bound" to "I/O bound".

Note: although it is better to have optimized code and accurate answers, the goal of this exercise is to monitor the behaviour of your the computer as the size of the dataset increases. Therefore ensure the benchmarks are accurate and present clear and clean graphs.

1. Basic hardware profile.
   a) What CPU does your computer have?
   b) How much RAM does your computer have?
   c) Explain how you will monitor the RAM and CPU usage in the following questions.
2. Determine the following information:
   a) Which carrier is most commonly late?
   b) Which are the three most commonly late origins, due to bad weather?
   c) What is the longest delay experienced for each carrier?

   Hint: do not use a spreadsheet editor, you would **miserably** fail.
3. Can you discover any pattern explaining departure delays?

   Hints:
   - Test various statistical models on a specific year, and try to extend your result to more years.
   - If a single year already takes too long, start with only a few months.

- As a starting point, model the departure delay as follows.

$$DepDelay = DayOfWeek + DepTime + CRSDepTime + ArrTime + CRSArrTime + UniqueCarrier.$$

Ex. 3 — *Very basic Java*

1. Given a text file where each line is composed of three fields, first-name, name, and email, write a very short and simple Java program generating a text file where (i) the order of the lines is random and (ii) each line is composed of the previous fields in the following order: name, first-name, and email.
2. Use inheritance an polymorphism to define various types of vehicles owned by a company. The definition of the actual objects is left to your imagination. Write a short program to demonstrate your work.