Data management - Project

Marian Aldescu - DS
marian.aldescu@etu.univ-grenoble-alpes.fr

December 22, 2020

For this project, I used my personal computer with the following attributes:

- Software: OS:Ubuntu 16.04, python3.7, spark-3.0.1
- Hardware: i5-4210U CPU @ 1.70GHz, 4 cores, 2 threads/core, 8GB RAM

1— *What is the distribution of the machines according to their CPU capacity?*
Answer:
In order to extract the CPU capacity, I use the machine_events file. Since there can be multiple events with the same machine_ID, I filtered the entries to have a list with no machine_ID duplicates, then I counted how many machines correspond to each CPU type.

Elapsed time: 2.262s.

CPU	Machines
1	29.1
0.25	123
0.5	11632
1	796
Unknown	32

Table 1: Machines distribution over CPU capacity

2— *What is the percentage of computational power lost due to maintenance (a machine went offline and reconnected later)?*
Answer:
For a machine, I will consider the lost computational power as the time interval in which the machine is down('DOWN' time): after an event 'Remove'. I will compute the total time during a machine is running('UP' time) and from this we can easily obtain the total 'DOWN' time.

During the trace, a machine is usually passing through a succession of alternate operations 'Add' and 'Remove'('Update' operations do not change anything), therefore the total up can be computed as a sum of $\Delta t_n = time_R - time_A$, where $time_R$, $time_A$ are successive timestamps of the same machine(A=Add, R=Remove).

$$totalTime_{UP} = time_{R1} - time_{A1} + time_{R2} - time_{A2} + ... \quad (1)$$

To compute $totalTime_{UP}$, I use first a **map** operation to associate a timestamp for each machine, and the timestamp is negative if the event has the type 'Add'. Then, with a **reduceByKey** I compute the sum from Eq.(1) for each machine_ID.

An important detail is that the last event for each machine has indefinite action time, hence I will consider the stop trace time, as the timestamp of the last event from the dataset.

After running the computations, I obtained the following percentages:

Running time $\approx 96.1\%$
Lost time $\approx 3.9\%$