INFO8002 Large-scale database systems
Exercise session III

Instructions

For the main question, be ready to:

- Describe the architecture of your solution and its operations.
- Define all its components and their interactions.
- Motivate your design decisions.
- Make diagrams whenever necessary.
- **State clearly your choices and assumptions.**

Question 1.

You are a project leader of a data engineering team at an Astronomical Observatory. Every night, the set of telescopes generate terabytes of optical observations that have to be stored reliably at the observatory's data center as the data will be processed during the day. Your team is responsible for designing the *write(observation)* API, and should be robust against at least 2 failures. Describe the architecture and the underlying procedure, state your assumptions.

Question 2.

After a day of processing, the data is ready to be queried by astronomers. Come up with a MapReduce computation graph to confirm the expansion of the universe. The data is partitioned among different compute nodes. Every point of light in the optical observation is classified as a STAR, GALAXY or QUASAR and has a redshift (z) attribute (i.e., Doppler redshift). Prevent as many network shuffles as possible.

- It is possible that some compute-nodes are slower then others, how do you guarantee the throughput?
- What if a compute-node crashes during the computation?