INFO8002 Large-scale data systems
Exercise session V

Instructions

For the main questions, 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

For feedback, feel free to e-mail joeri.hermans@doct.uliege.be

Question 1.

You are responsible for the design of a data processing pipeline which is looking for the longest word on the internet. Your bots will periodically push the scraped webpages to your servers. At the end of every day, your systems will have to process these webpages. Discuss the storage architecture (i.e., sharding), and your MapReduce operations to process said data. Think about all necessary involved steps. Consider the scenario when worker nodes file. How will this be detected, how does the system know what tasks have to be rescheduled? Does the full job have to be recomputed from scratch, or can it be restarted from a certain checkpoint. How would you implement such a mechanism?

Guidelines

It is recommended to design your system in the following steps:

1. What is the distributed system model you are assuming?
2. Identify all actors in your system. Provide a short description.
3. Describe the general architecture of your system.
4. For every operation, provide a short description.
5. Draw sketches of your execution (it can be in MS Paint or something, no problem).

Question 2.

Discuss the straggler problem. Why does it happen? What can be done to alleviate it, and at what cost?