Cloud computing – a.y. 2023/2024

Final Project - Analyze letter count through Hadoop

Preliminaries - Consider a Hadoop cluster installed in **pseudo-distributed** mode (a.k.a. single-node cluster). For installation, the student needs to follow the tutorial at: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html. Besides, they may find useful information on the tutorial that I showed in class for the fully distributed installation mode.

Input data - Download and unzip the snippets.zip input archive. The `snippets` folder contains 782 text files having name `lineXXX`, where XXX runs from 000 to 781. Each input file contains a text on multiple lines, including numbers, punctuation, etc.

The Job - The goal of this project is to use Hadoop to count the number of times each letter appears within the input documents. In this analysis, consider only the 26 standard characters from the English alphabet and **ignore any special character**, such as: -_.!?'$@''#.

The student must demonstrate full knowledge of MapReduce and Hadoop and include a **Combiner** logic in their project.

The student must run **simple experiments** to:

- Report letter count from the considered input dataset
- Report simple statistics on MapReduce execution, e.g., execution time, memory consumption, impact of combining

Optionally, the student may consider In-Mapper combining or increase the number of Reducer tasks.

The student must write a **brief** (few pages) **report** describing their algorithm design through pseudocode, the experimental settings and the obtained results.

FAQs on the Cloud computing project and final exam

1. **When and how to upload the final project?**
   - The final project (code + report) must be uploaded through the **Google form** at: https://forms.gle/ycfMLuy5qaMsgG1q8
   - Each student has to submit their project as a .zip file named "Single{NameSurname}_Session{DD-MM-YYYY}.zip"
   - Project upload must be performed one week before the official session day, at the latest. For example, if the session is scheduled on the 12/06/2024, the student has time to upload their project by 04/06/2024 (included). **Projects that are uploaded after the deadline will be considered for the next session**
   - Once the project is uploaded, **no further modifications** are allowed to it

2. **What about the project discussion?**
   - Project discussion will take place on the official session day or few days before, depending on the amount of projects submitted in that session. In the second case, I will inform you on the exact date and time
   - Project discussion will consist of **two parts**:
     - For the first part (duration 15 minutes – hard time), the student has to prepare a presentation (either through a separate PowerPoint or directly commenting the project report) of the MapReduce pseudocode, comment the experimental settings and obtained results, and show an execution demo of their work on the cluster
     - The second part (no duration pre-established) will be a detailed discussion of Java code
   - The student can choose to discuss the project either in English or Italian

3. **What about the oral exam?**
   - The overall exam consists of Hadoop project discussion and the oral exams for my part and Prof. Vallati's one
   - The student can take the different parts of the exam in **different sessions and in any order**
   - Each part of the exam will have a **validity of one year**. After that deadline, that part must be repeated by the student. For example, if you discuss my oral part on 12/06/2024, this will be valid until 11/06/2025
   - Oral exams will take place on the **official session days** only
   - The student can choose to take their oral exam either in English or in Italian