Instructions for project work for LDSA 2017

The project aims at giving you first hand experience of the challenges involved in approaching a new data engineering problem, while providing an opportunity to deepen the practical experience with some of the tools that has been introduced in the class. By studying a new (for you) scientific dataset you will go through the processes of understanding the problem, the data, and using it, develop a scalable data processing backend.

Assignment

The Open Science Data Cloud (OSDC) is an initiative in cloud computing and data analysis, and they host large, public datasets of scientific interest. You should:

1. Browse the public datasets here https://www.opensciencedatacloud.org/publicdata/?commons_type=General and here: https://www.opensciencedatacloud.org/publicdata/?commons_type=Environmental and choose a dataset to work on based on your scientific interests. Also, consider the size of the data, whether it is possible to extract subsets of it easily and stage in SSC, etc.

2. Design a scalable data processing solution for the chosen dataset, demonstrating your knowledge of key concepts and tools introduced in this course. Typically, massive datasets needs to be analyzed in stages, where the first step tend to consists som some form of data reduction, for example a filter or feature extraction. In this stage, all data needs to be accessed and the output of the operation is a subset or reduced dataset more amenable to interactive analysis. By reading up on previous use of the data, propose some form of computational experiment that allows you to:

    a. Reuse previous analysis software or develop some simple code. Not that this is not the main aim of the project, so the details are not that important.
    b. Drive the development of a horizontally scalable data processing solution. This is a central part of the project.
    c. Demonstrate the scalability of your approach in a computational experiment. This is also an essential part of the project.

3. Write a report describing your work and results.

The formal presentation of the project work consists of the following:

1. Attend the project seminar, for which you should prepare a 10min presentation. The presentation should not cover the final results of the project, but rather answer the following questions:

    a. What dataset have you chosen to work with, and what is the scientific background.
    b. What is your tentative plan for the computational experiments. What technology will you work with, and how will you design your scalability studies?
    c. Any preliminary results.

2. Write and submit a project report.

The written report should not exceed 2500 words, and it should contain the following sections:

Title

Background
Description of the scientific area/problem that gave rise to the dataset. Place the dataset in context, providing sufficient references for the reader to understand the importance and significance of the data. What kind of analyses have been done in the literature?

Data format
Describe the data format(s) used in the dataset. Put them in context: why were the specific formats chosen, would there be alternatives? What are the pros and cons of the formats used?

Computational experiments
The main section of the report. Describe and motivate the choice of tools and the distributed system that you designed. Describe how you have designed your scalability experiments, and present the results.

Discussion and conclusion
Here you can discuss the outcome of the experiments and the experiences gained. Was your chosen approach suitable. What worked well and what could be improved?

References