Instructions for project work Data Engineering I

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 course. By studying a new (for you) scientific dataset you will go through the processes of understanding the problem, the data, and to develop a scalable data processing backend.

Assignment

There are many sites hosting large open datasets to the public, in the hope that they will be useful for scientific analysis. In your project you should:

1. Browse the public datasets listed below, and choose a dataset to work on based on your scientific interests. Note that some datasets are very large. In that case, make sure that it is possible to extract subsets of it easily and stage them in SSC.

2. Design and implement a scalable data processing solution for the chosen dataset, demonstrating your knowledge of key concepts and tools introduced in this course. Typically, massive datasets need to be analyzed in stages, where the first step tends to consist of some form of data reduction, for example a filter, or feature extraction for downstream machine learning. In such pre-processing stages, 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 in line with the above that allows you to reuse previous analysis software or develop some relatively simple preprocessing or analysis code. Note that development of data-specific analysis code is not the main aim of the project, so the details are not that important and the analysis does not need to be advanced.

   With the dataset and analysis/preprocessing objective defined:

   a. Architect and develop a horizontally scalable data processing solution. This is the central part of the project.
   b. Demonstrate the scalability of your approach in suitable computational experiments. This is also an essential part of the project.

Recommended datasets:

1. 1000 genomes:
   https://aws.amazon.com/1000genomes/
2. Million songs dataset:
   http://millionsongdataset.com/
3. Reddit comments: http://files.pushshift.io/reddit/comments/

There is also a list of many interesting datasets here from all areas of science and technology: https://github.com/awesomedata/awesome-public-datasets

If you have a special interest and want to choose a dataset from here you need to contact Andreas and be prepared to answer the following questions:

1. What is the total size of data accessible for download?
2. What is the format of the data?
3. Have you tested loading the data in Python?

Presenting your results

Your work should be presented in the form of a written report describing your work and results.

Seminar: You are required to attend the project seminar, for which you should prepare a 5 min 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 of the data?
b. What is your tentative plan for the architecture and computational experiments? What technology will you work with, and how will you design your scalability studies?
c. Present any preliminary results.

Report: The written report *must not exceed 2500 words excluding references*, and it should be structured as follows:

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, and would there be alternatives? What are the pros and cons of the formats used?

Computational experiments

This is 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. Think carefully about suitable ways to illustrate the scalability of your solution.

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

Be consistent in what format you use for your references. For example, use the MLA (Modern Language Association) format. At http://www.ub.uu.se/ you can click on “cite this item” and choose a format.