Lab Course: Distributed Data Analytics

0. Overview

Mohsan Jameel

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

Outline

0. Organizational Stuff

1. Lecture Overview

Outline

Exam and Credit Points (1/2)

- The course gives 6 ECTS
- requires 180h student effort, the duration of the course is 14 weeks.
  1. 4h/week (in the lab)
  2. 9h/week (own time for solving exercise sheets)
  3. (4 + 9) h/w * 14 w = 180h

- There will be a weekly exercise sheet.
- You will get approximately 6 to 7 days in-between the date of release and the date of submission.
- The grading of this course will be based on solutions submitted in each individual lab.
  - There will be no written exam at the end of term

Exam and Credit Points (2/2)

- The course can be used in
  - Data Analytics MSc
  - IMIT and AINF MSc. / Informatik / Gebiet KI & ML
  - Wirtschaftsinformatik MSc / Business Intelligence
- Register yourself at LSF (POS module) and learnweb.

Exercises

- There will be a weekly exercise sheet with 3 questions uploaded **every Thursday** to our webpage or learnweb (3114).
- Solutions to the exercises can be submitted until **next Friday 23:59 Berlin Time**
- Exercise Sheets will be graded in next Lab (including viva)
- Labs Group 2 **every Monday 14:00–18:00**, C-147
- Labs Group 1 **every Thursday 10:00–14:00**, C-147
- Each lab exercise will carry equal weight-age towards the final mark.

Exercise Submission Format

Each Excercise will consists of three questions

- Q1: Implement a given problem using parallel/distributed computing concepts. [10 Marks]
  - Need to provide a working code and report
- Q2: Show performance gains/improvement over serial program [5 Marks]
  - Graphs or tables showing speedup curves or execution time
  - explanation of the graphs/tables
- Q3: Solve problem with state-of-the-art library [5 Marks]
  - Graph comparing state-of-the-art and your code
  - comparison of execution time (etc)

Exercise Submission Format

- You should submit a single .zip or .tar file (Please dont use other formats). Name your file as LASTNAME_exNO.zip i.e. Bob_ex01.zip
  1. A report as a .pdf document (LASTNAME_exNO.pdf i.e. Bob_ex01.pdf), which contains all the graphs and outputs along with the explanation of the results. The report in a word document format will not be accepted.
  2. All your code that are required to complete a task must be in .py or .ipynb (LASTNAME_exNO_qNO.pdf i.e. Bob_ex01_q1a.py etc) format. Your code should be self explanatory and well commented.

Exercise Checking

- Each student will submit an individual solution. (no group submissions)
- All submissions should be made through the learnweb (course code 3117).
- No late submission, missing a lab will result in 0 points.
- Need to present the solution in the next lab possible.
- Points will be awarded based on your submitted report and code.
- To obtain maximum mark, you also have to present your solution in the lab a couple of times.
- A question answer session (Lab viva) will be conducted for a random sample of students.
- Write your own code/solution. Do not copy it.

Outline

Syllabus

Thu. 09.04.	(1)	Introduction and Distributed Computing with MPI I
Thu. 16.04.	(2)	Distributed Computing with MPI II
Thu. 23.04.	(3)	Distributed Computing with MPI III
Thu. 30.04.	(4)	TensorFlow I
Thu. 07.05.	(5)	TensorFlow II
Thu. 14.05.	(6)	TensorFlow III
Thu. 21.05.	(7)	TensorFlow III
Thu. 28.05.	(8)	Apache Spark I
Thu. 04.06.	(9)	Apache Spark II
Thu. 11.06.	(10)	Apache Spark III
Thu. 18.06.	(11)	Distributed Machine Learning Algorithm I
Thu. 25.06.	(12)	Distributed Machine Learning Algorithm II
Thu. 02.07.	(13)	Distributed Machine Learning Algorithm III