

Lab Course: distributed data analytics

0. Overview

Nghia Duong-Trung, 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

0. Organizational Stuff

1. Lecture Overview

Exam and Credit Points

- ▶ The course gives 6 ECTS
- ▶ 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
- ▶ 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 (3117).
- ▶ 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 **every Monday 14:00–18:00**,
- ▶ Each lab exercise will carry equal weight-age towards final grade.

Exercise Submission Format

Each Exercise will consists of three questions

- ▶ Q1: Implement a given problem using parallel/distributed computing concepts. [10 Marks]
 - ▶ Need to provide a working code.
- ▶ Q2: Show performance again/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 Checking

- ▶ All submission should be made through 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 code (50-70%) and answer to viva questions (30-50%)
- ▶ Lab viva missed for more than one week will result in points based on submission.
- ▶ Write your own code/solution. Do not copy it from Internet.
- ▶ Each student should have its own solution and submission.

Outline

0. Organizational Stuff

1. Lecture Overview

Syllabus

- Thu. 10.04. (1) Introduction and Multicore Programming I
- Thu. 17.04. (2) Multicore Programming II
- Thu. 24.04. (3) Distributed Computing with MPI I
- Thu. 01.05. (4) Distributed Computing with MPI II
- Thu. 08.05. (5) Distributed File System and Apache Hadoop HDFS
- Thu. 15.05. (6) Apache Hadoop Map Reduce Family
- Thu. 22.05. (7) NoSQL Databases I
- Thu. 29.05. (8) NoSQL Databases II
- Thu. 05.06. (9) Apache Spark I
- Thu. 12.06. (10) Apache Spark II
- Thu. 19.06. (11) Distributed Machine Learning Algorithm I
- Thu. 26.06. (12) Distributed Machine Learning Algorithm II
- Thu. 03.07. (13) Distributed Machine Learning Algorithm III