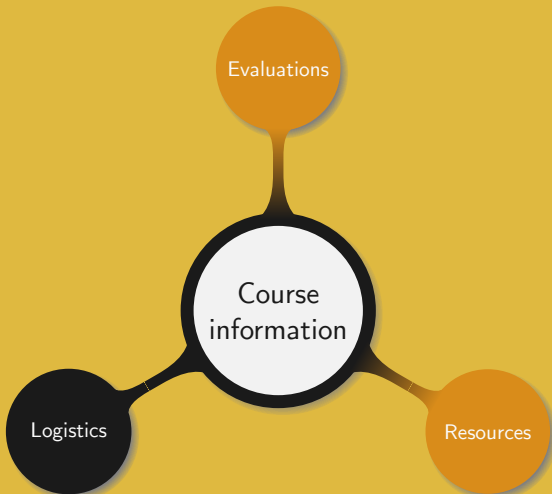


Methods and tools for big data

0. Course information

Manuel – Summer 2022



Teaching team:

- Instructor: Manuel (charlem@sjtu.edu.cn)
- Teaching assistants:
 - Yangyang (wangyangyang@sjtu.edu.cn)
 - TBD (TBD)

Important rules:

- When contacting a TA for an important matter, CC the instructor
- Prepend [VE472] to the subject, e.g. Subject: [VE472] Grades
- Use SJTU jBox service to share large files (> 2 MB)

Never send large files by email

Course arrangements:

- Lectures:
 - Tuesday 16:00 – 17:40
 - Thursday 16:00 – 17:40
- Labs: Wednesday 18:20 – 20:40

Office hours:

- Anytime on Piazza
- On appointment

Primary goals:

- Understand how big data sets are analysed in practice
 - Be able to use Hadoop
 - Learn how to work in the Hadoop ecosystem
- Be able to performed advanced data analysis on large data sets
 - Get good foundations on big data analysis
 - Be able to design, implement, and use advanced algorithm in Spark

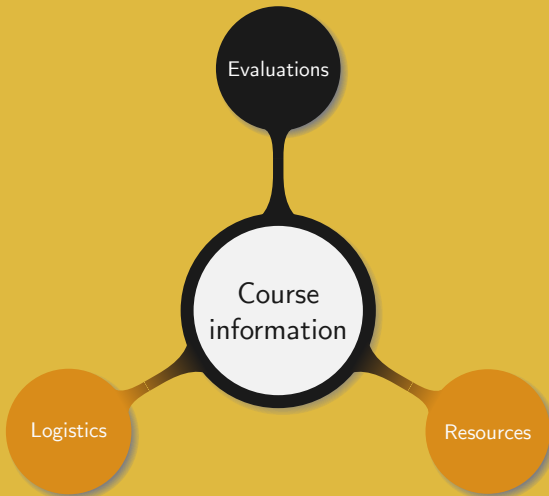
Be able to analyse any given dataset, regardless of there size

Learning strategy:

- Course side:
 - ① Understand the new issues appearing as datasets grow
 - ② Be able to setup a Hadoop cluster and use it
 - ③ Understand why traditional algorithms fail on big data
 - ④ Be able to implement advanced algorithms for big data
- Personal side:
 - ① Derive algorithms for big data
 - ② Use and work “inside” Hadoop, Drill, and Spark
 - ③ Relate known strategies to new problems
 - ④ Perform extra research

Detailed goals:

- Understand the basic logic behind Hadoop
- Have a general knowledge of the Hadoop ecosystem
- Be familiar with the basic Hadoop components: HDFS, YARN, and MapReduce
- Understand the structure of Drill and Spark
- Be able to work in Hadoop and “extend” its functionalities
- Know what tool to use for common specific purposes related to the study of big data
- Be familiar with common dimension reduction techniques
- Understand the limitations when facing “real” big data
- Be able to run basic data analysis on big data



Homework:

- Total: 5 or 6
- Content: basic Hadoop, algorithms, Spark

Labs:

- Total: 12
- Content: guided sessions to setup and work with Hadoop, and Spark

Projects:

- Total: 1
- Content: analysis of some big dataset

Challenge:

- Total: 1
- Content: compare theory and practice in Hadoop and Spark implementations

Grade weighting:

- Midterm exam: 15%
- Final exam: 15%
- Quizzes: 20%
- Projects: 30%
- Homework: 10%
- Labs: 10%

Assignment submissions: -10% per day, not accepted after 3 days

Grades will be curved with the median in the range $[[B, B+]]$

General rules:

- Not allowed:
 - Reuse the code or work from other students or groups
 - Reuse the code or work from the internet
 - Share too many details on how to complete a task
- Allowed:
 - Reuse part the course or textbooks and quoting the source
 - Share ideas and understandings on the course
 - Provide hints on where or how to find information

Documents allowed during the exams:

- Midterm: none
- Final: a single A4 paper sheet with original handwritten notes

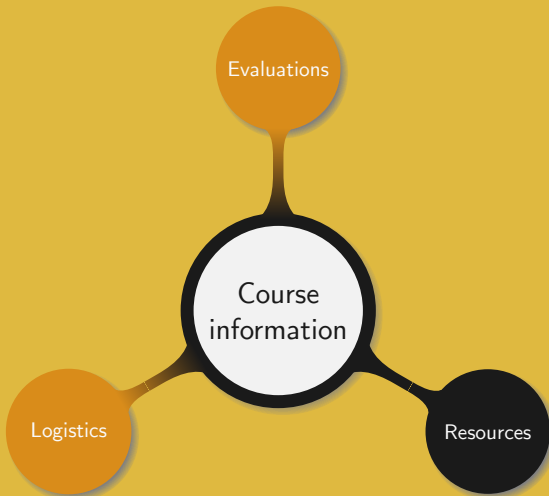
Group works:

- Every student in a group is responsible for his group's submission
- If a student breaks the Honor Code, the whole group is guilty

Contact us as early as possible when:

- Facing special circumstances, e.g. full time work, illness
- Feeling late in the course
- Feeling to work hard without any result

Any late request will be rejected

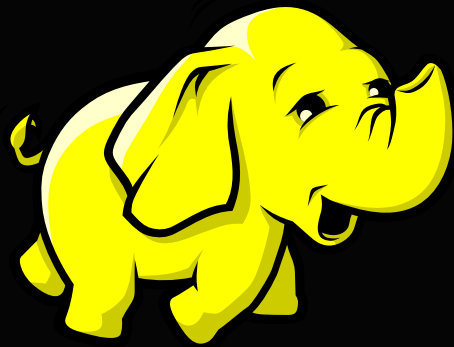


Information and documents available on the Canvas platform:

- Course materials:
 - Syllabus
 - Lecture slides
 - Homework
 - Labs
 - Projects
- Course information:
 - Announcements
 - Notifications
 - Grades
 - Polls

Useful places where to find information:

- *Hadoop the definitive guide*
- *Spark the definitive guide*
- *Machine learning, an algorithmic perspective*
- *Introduction to Data Mining*, by Tan et al..
- *Mining of Massive Datasets*, by Leskovec et al.. by White
- Search information online, i.e. $\{\text{websites} \setminus \{\text{non-English websites}\}\}$



Thank you!