MDS Self-Study
Winter Term 2024/25
Prep 3

For the upcoming lesson on November 4, 2024, prepare the self-study material in advance and answer the questions under Q&A. Bring your answers to the next lesson.

Self Study Material

This unit is based on the book "Mining Massive Datasets" (version 3, available online for free at `http://www.mmds.org/#ver30`), specifically

- chapters 2.1–2.2 for a recap on distributed file systems and MapReduce,
- chapters 2.3.3.–2.3.8 on implementing operators of relational algebra using MapReduce,
- chapters 2.5.1–2.5.2 on communication costs

Q&A

1. In MapReduce processing, what is a prerequisite for using a Combiner?

2. If you use a Combiner, does that mean you won't need a Reducer?

3. How many Map tasks are usually created?

4. Why is the number of Reduce tasks typically configured to be lower than the number of Map tasks?

5. In the extended relational algebra, what is the operator symbol for the grouping operator?

6. What is the communication cost of an algorithm?

7. When can communication cost *not* be used to measure the efficiency of an algorithm?

8. What is the justification for focusing on communication costs in (acyclic) MapReduce workflows?

9. What are the two reasons why we do not consider the output size?

10. Why is wall-clock time also an important criterion?

11. What is the communication cost of computing $R(A, B) \bowtie S(B, C)$ with MapReduce?