VE472 — Methods and tools for big data

Challenge 1

Manuel — UM-JI (Summer 2022)

- Read research articles
- Better understand big data tools
- Rewarded by a bonus on the final grade

Topics and tasks

Select at least one of the following topics, read the original research article, and compare it to how the idea was implemented in practice. Comment on those choices.

- Drill, based on the Dremel paper [4].
- HDFS, based on the Google file system paper [2].
- MapReduce, based on the Mapreduce paper [1].
- Spark, based on the Spark paper [5].
- Hadoop's resource management based on the dominant resource fairness paper [3].

Reward

Each topic can bring up to two marks that will be added to the final grade, after the curve has been decided. The quality of the research work will be assessed through a slide presentation and a short question and answer discussion.

References

[1] Jeffrey Dean and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters". In: *Commun. ACM* 51.1 (Jan. 2008), pp. 107–113. ISSN: 0001-0782. DOI: 10.1145/1327452.1327492.

[2] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. "The Google File System". In: *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*. SOSP '03. Bolton Landing, NY, USA: ACM, 2003, pp. 29–43. ISBN: 1-58113-757-5. DOI: 10.1145/945445.945450.

[3] Ali Ghodsi et al. "Dominant Resource Fairness: Fair Allocation of Multiple Resource Types". In: *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation*. NSDI'11. Boston, MA: USENIX Association, 2011, pp. 323–336. URL: http://dl.acm.org/citation.cfm?id=1972457.1972490.

[4] Sergey Melnik et al. "Dremel: Interactive Analysis of Web-scale Datasets". In: *Commun. ACM* 54.6 (June 2011), pp. 114–123. ISSN: 0001-0782.

[5] Matei Zaharia et al. "Spark: Cluster Computing with Working Sets". In: *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*. HotCloud'10. Boston, MA: USENIX Association, 2010, pp. 10–10.