

Summary

- **Large-scale data** poses new technical problems for:
 - **storage** \Rightarrow distributed file systems.
 - **computations** \Rightarrow MapReduce programming model.
 - Split the data in chunks.
 - Map workers all execute the same operation on a chunk and return a key-val pair.
 - Reduce workers process all key-val pairs with the same key at once.
- **Algorithmic costs** of MapReduce:
 - **Communication costs** vs. **computation costs**.
 - **Reducer size** and **replication rate**.
- Extensions of MapReduce: **Spark** and **TensorFlow**.
- MapReduce for **machine learning**.
- **Link analysis** with **PageRank**.