DLI Accelerated Data Science Teaching Kit

Lecture 10.3 - MapReduce Overview

How Hadoop Scales Up Computation?

- Uses master-worker architecture, and a simple computation model called MapReduce.

- A simplified way to think about it

  1. Divide data and computation into smaller pieces; each machine works on one piece

  2. Combine results to produce final results

MapReduce: Simplified Data Processing on Large Clusters

http://static.usenix.org/event/osdi04/tech/full_papers/dean/dean.pdf

How Hadoop Scales Up Computation?

More technically...

1. **Map phase**
   Master node **divides** data and computation into smaller pieces; each worker node (“mapper”) works on one piece independently in parallel
2. **Shuffle phase** (automatically done for you)
   Master **sorts and moves** results to “reducers”
3. **Reduce phase**
   Worker nodes (“reducers”) **combines** results independently in parallel

DLI Accelerated Data Science Teaching Kit

Thank You