Map Reduce
A Flexible Data Processing Tool
What is Map Reduce

- A programming model for processing and generating large datasets
- Parallel, distributed algorithm on a cluster
- Consists of two major parts: map() and reduce()
“Map” step

- The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes.
- Sub-problems may be divided again, forming a multi-level tree structure.
- Worker nodes solve the problem and then pass the result back to the master node.
“Reduce” step

- The master node collects the answers to all the sub-problems
- Master node combines the answers to all the sub-problems and combine them to form the output, which is the answer to the original problem
Advantages of Map Reduce compared to traditional parallel database

- Highly effective
- Fault-tolerant - Failure in the middle of a multi-hour execution does not require restarting the job from scratch
- Works well in a heterogeneous system
- Good framework for complicated functions