MapReduce: Simplified Data Processing on Large Clusters

Jarett Miller

Alan Labouseur

25 November 2013

Dean, Jeffrey and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters". OSDI 2004. (2004) Web. 20 Nov. 2013.

Overview

- Consists of 2 functions, map and reduce
- Map- filters/sorts a certain data set
- Reduce- merges and summarizes data provided by map input
- There are many functional/real world uses for MapReduce, in fact, it is used by Google, Facebook, and Amazon
- Used for large sets of data
- Allows users with little or no experience to run parallel operations
- “Scalable”, meaning it can be modified to work on different sized systems

Implementation

- Achieves high performance on large clusters of commodity PCs (clusters can consist of hundreds or thousands of machines)
- The input data is processed by the map and reduce functions across these machines, each of them processing a part of the data
- If worker machines do not respond to an initial ping for a certain length of time, the job is passed on to another machine
- Final output is saved to inexpensive IDE disks
- Many different implementations exist depending on the environment

Analysis

- Quite a simple process, which is easy to use and implement
- Process time is dependent upon amount of data, but could prove to be extremely useful to many small companies as well as large ones mentioned in the paper, such as Google
- Bandwidth of a network is vital to the transfer of the data/info and can affect efficiency

Advantages/Disadvantages

Advantages

-Easy to use

-Uses relatively cheap hardware

-Parallel processes data of all sizes, handling failures exceptionally well along the way

-MapReduce is scalable

-map-reduce, opposed to file systems more effectively decouples computation from storage than key-value servers do, and is much easier to scale out

Disadvantages

-Time efficiency can be altered depending on the size of the data

-Not suitable if a lot of data passes through the network

-Parallelization may not be the necessary solution in many situations

Real-world Use Cases

- Google- at Google, MapReduce replaced ad hoc programs, and now updates their large-scaled indexes and run various types of analysis. It also is responsible for the data generated in web searches
- Used for distributing pattern-based searches, distributed sorting, graph computations, etc.
- Amazon's site uses elastic MapReduce in order to process searches, and create suggestions based upon users search history