MapReduce: Simplified Data Processing on Large Clusters

Matthew Musich
November 25, 2013

Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." Google, Inc., 2004. 23 Nov. 2013.

Main Idea

- MapReduce is a programming model and implementation technique
- Processes and generates results of large sets of data
- Used on large clusters of commodity machines in order to process data quickly

Implementation

- MapReduce programs use 2 functions:
  - **Map:** processes data inputs into a key pair that outputs pairs of intermediate keys
  - **Reduce:** merges the intermediate values stored from the map function based on the intermediate keys and creates an output

- The inputted data is processed by the map and reduce functions across many machines that process a small portion of the data

Analysis

- It was a good place to start for parallel machine based processing, but there can be improvements in optimization

- There are still limits of time when it comes to processing immense amounts of data, especially for what Google processes.

- Network bandwidth might not be the best way to transfer the information

- There might be better options in physical disks directly reading and writing rather than being sent over the network.

Advantages & Disadvantages

- Advantages
  - The key system can keep track of analytics on the live data
  - Linear scaling with the simple addition of cheap commodity hardware
  - Simple programming model
- Disadvantages
  - Not always easy to implement everything as a MR program
  - A lot of data shuffling over the network
  - Not efficient for large amounts of short process data

Real World Uses

- Fast and efficient processing of log files of any type
  - Etsy uses Amazon's Elastic MapReduce to calculate user behavior and search recommendations
  - Nokia collecting and analyzing with MapReduce, vast amounts of data from their mobile phone network
- Manage transactions of data from data storage systems
  - NetApp uses MapReduce to parallel process transaction data from their main systems for diagnostic purposes