MapReduce & Parallel DBMS

By: Clayton Enright

Sources:

Dean, Jeffery and Ghemawat, Sanjay. “MapReduce: Simplified Data Processing on Large Clusters”.
Pavlo, Andrew; Paulson, Erik; Rasin Alexander; et. al. “A Comparison of Approaches to Large-Scale Data Analysis”.

MapReduce

- Allows for quick and efficient processing of large data sets
- The system partitions and distributes input data for the user and runs the analysis in parallel across a large cluster of independent machines
- Simple, only two main functions, ‘Map’ and ‘Reduce’
- ‘Master’ copy of program distributes Map and Reduce functions to machines evenly
- Map outputs are stored on local disks where they are then received by Reduce functions and rolled up into final output

Implementation of MapReduce (Google)

- Google's implementation allows for high scalability and it is very easy for their programmers to pick up and begin using
- "Large clusters of commodity PCs connected [through] switched Ethernet" (Dean & Ghemawat)
- Allows for fault tolerance to avoid problems caused by malfuncitoning individual units
- Ability to skip "bad records" in large data sets where it is possible to overlook a few records
- Gets rid of the need for a new specialized program each time a user needs to process data in a certain way (allows for simple expression while hiding the more complex parts of load balancing and parallelization)

Analysis of MapReduce

- Easier to set up and start using than parallel DBMS
- Effective for use in large companies where many programmers may not have experience with parallel systems
- Allows for many users to analyze large data sets quickly and to come to conclusions and higher quality decision-making information much faster

Comparison to Parallel DBMS

- MapReduce simplifies the intricacies of implementing parallel DBMS solutions by hiding many of the load-balancing and parallelization functions
- Parallel DBMS can run many analysis tasks much faster than MapReduce implementations like Hadoop.
- Parallel DBMS are streamlined when used directly, but can be challenging to grasp and take a bit longer to code

Advantages & Disadvantages

- MapReduce is a relatively simple model with only two main functions, Map and Reduce
- Allows user to structure data in any way he or she may see fit (or no structure at all)
- Easy to learn and start using effectively
- More coding may be required, low-level language
- Output not generated as quickly
- Unstructured data can create more work and take longer, can also lead to corrupted data output