MapReduce: Simplified Data Processing on Large Clusters, A Comparison of Approaches to Large-Scale Data Analysis, and Michael Stonebraker on his 10-Year Most Influential Paper Award at ICDE 2015

All papers and videos obtained from Labouseur.com

Danny Mulick
Database Management
10/20/2016

MapReduce: Simplified Data Processing on Large Clusters

- This paper was written to exemplify the characteristics and process of MapReduce programming
- MapReduce is a model for programming that allows for the processing of large amounts of data.

Implementation

- This is done through the use of parallel machine instances running the same function to speed up processing, under the control of one master machine.
- Once all data is processed, another set of machine instances aggregate the processed data and return information about it

My analysis

- My thoughts on this are that it is the next logical step in the evolution of programming, and it is a wonderful idea.
- Since the utilization of the assembly line, humans have been splitting up tasks to make them easier as more workers are able to work on them.
- That is much like this method, as we are taking a large project, categorizing it, and setting many machines to work on it simultaneously.

Main ideas of second paper

- A Comparison of Approaches to Large-Scale Data Analysis is written to show the differences between some of the leading methods of analysing big data.
- Two approaches to this task are to use MapReduce or Parallel DBMSs
- The systems discussed within the paper are Hadoop, DBMS-X and Vertica
- The paper highlights the differences in the methods of loading data

Implementation

- Each of the systems noted in the paper take in segments of data, and process them in some fashion, usually in differing stages that return one final result based off of analysis of the data.
- Multiple systems, or instances, of the functions run on virtual nodes to accelerate the processing of the data, as opposed to solely one system running the functions

Analysis for Second Paper

- Each method for analyzing big data stores has its own benefits and detriments.
- It seems that all the methods still return the same result, the only difference really seems to be in the process of obtaining the metadata
    - MapReduce changes the way the data is processed, by splitting it up into two different functions
    - DBMSs utilize the SQL language to obtain the data, then to write out tables of metadata to be viewed

Comparison of the Two Papers

- The MapReduce program style is more reliable than the DBMS style. Should a task or instance fail in the MapReduce process, only one section has to be rerun, if any. For DBMS, the entire query must be restarted
- Both have pre-built functions that allow the user to jump right in and get to work with their task

Stonebreaker Main Ideas

- Spoke about how the future of data storage will be column stores replacing row stores
- Talks about the up and coming fields of data science and analytics
- “One size fits none” mentality
  - No adaptability among differing systems

Advantages/Disadvantages of MapReduce

- Advantages
  - MapReduce can effectively process large amounts of data faster than a traditional DBMS
  - Cost effective due to utilizing virtual machines instead of single instances on every physical machine
  - Little effort needed to set up each instance, beyond installing and setting up the directories for the libraries
- Disadvantages
  - Not universal, meaning that each instance has to be specifically written to handle each task