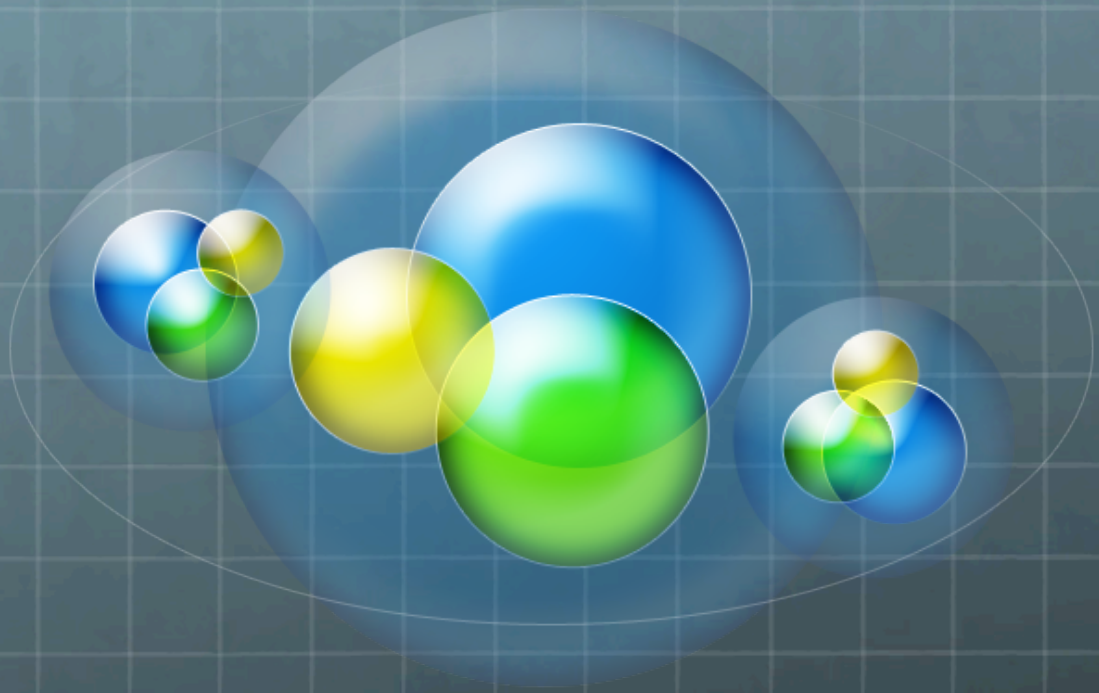

MapReduce

Skylar Senning

11/24/13

Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." OSDI 2004. (2004) Web. 16 Nov. 2013.

Main Ideas

- MapReduce: is a programing model created to read and generate large sets of data
- Uses a large number of computers (clusters) to process and generate results from data sets
- Easy to distribute work across nodes
- “Allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distibuted system.”(pg.1)

Implementation

1. Splits the input into N pieces of 16-64 megabytes, then starts copeis of the programs on groups of computers

2. One copy is the “Master”, which assigns work to the other computers. The master finds idle workers and assigns tasks

3. Workers read contents and parse out, map functions buffered into memory. Locations are passed back to master, who then forwards information to reduce workers

4. Reduce workers reads all intermediate data, and then sorts data. If the intermediate data is too large to fit in memory, then a external sort is used.

5. Reduce workers sorts for keys, and then pases the key and information to the reduce function.

6. User program is woken when all map and reduce tasks have been completed

Analysis

- Having the master pass all the information seems like a large amount of storge
- Encapsulation makes implementation far easier
- Fault deterrent by skipping records that could have errors
- Simple solution to dealing with large amounts of complex data without the need for complex code.

Advantages

- Highly Scalable
- Processing can be performed in parallel on multiple nodes (servers)
- Useful in a wide range of applications
- Fault Tolerant
- Backups and Checkpoints

Disadvantages

- It is challenging to code map functions
- A lot of overhead to implement
- Slow workers significantly delay completion time
- Map/Reduce function sometimes fail
- A lot of data passing through the network

Real World Usage

- Data Mining
  -Information on customers, trends, statistics
  -Used by Google
- Data Sorting
  -Large amounts of data can be sorted quickly
- Finance
  -Sales trends (amazon, walmart)
- Facebook used MapReduce until 2011 until the amount of data become too large