Cloud Computing with Map-Reduce and Hadoop

Author: Dwight D. Anderson Mentors: Wenjun Zeng, Qia Wang

Objectives

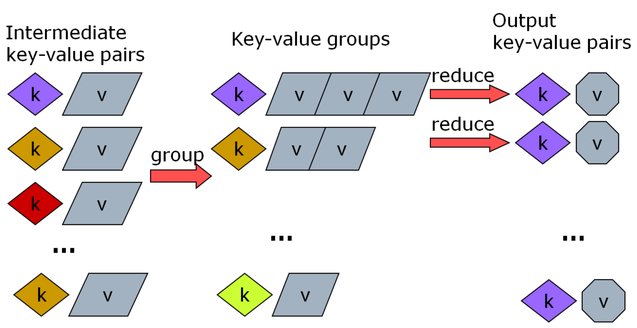
- Cloud Computing
- Parallel processing
- Map-Reduce algorithm
- Hadoop environment

Map-Reduce

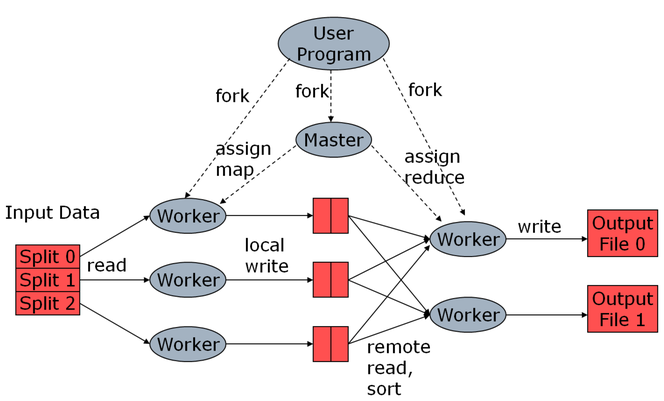
- Simple data-parallel programming model designed for scalability and fault-tolerance
- Pioneered by Google
- Popularized by Hadoop project

The Map and Reduce Step

map(k,v) → list(k1,v1)
reduce(k1, list(v1)) → v2



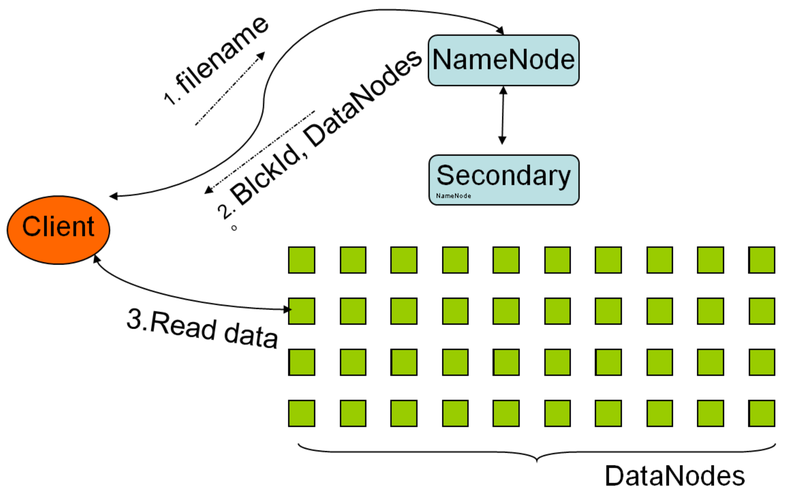
Distributed Execution Overview

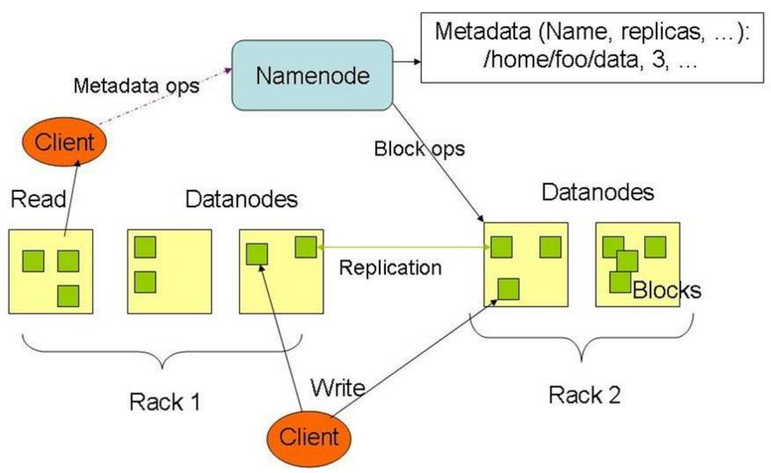


The Hadoop Environment

- Distributed, highly fault-tolerant file system
- Handling large data sets
- Communication protocols are build on the TCP/IP model.
- HDFS: Hadoop Distributed File System

HDFS Architecture

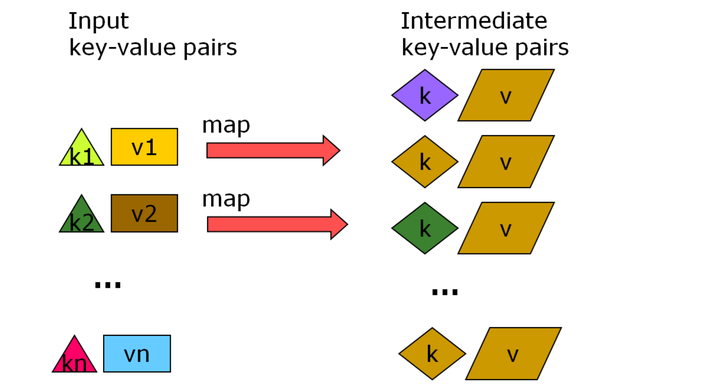




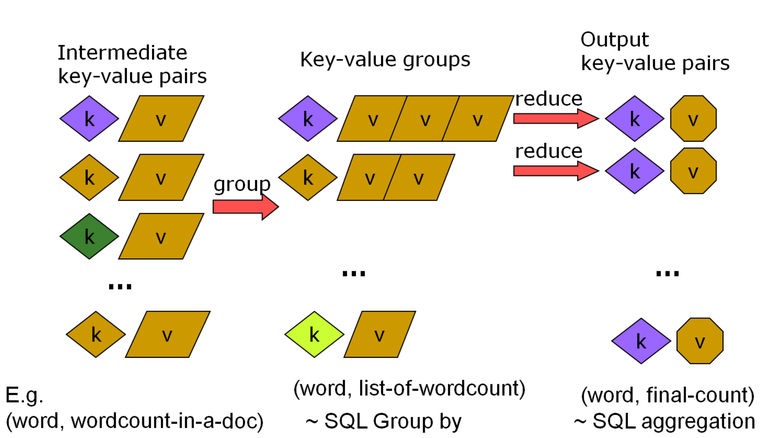
NameNode : Maps a file to a file-id and list of MapNodes
DataNode : Maps a block-id to a physical location on disk

Implementation: Word Count

The Map Step



The Reduce Step



Verification

- All the tasks are simulated within a single thread
- Results are verified correctly

Future Work

- Run the task with multi-threads or multi-servers to compare performance
- Apply Map-Reduce to more complicated data-mining algorithms