CS202 Project Proposal
Performance evaluation and comparison between distributed file systems

Group 13
Chen-Yang Yu 862052273
Po-Cheng Kuo 862029279

Motivation

DFS have become an important area of information processing and it’s rapidly developing

- Access to files from multiple hosts sharing via a computer network
- For multiple users on multiple machines to share files and storage resources

Hadoop Distributed File System

- It is a distributed file system that handles large data sets running on commodity hardware
- HDFS is used to scale a single cluster to hundreds of nodes
- Provides high throughput access to application data and is suitable for applications that have large data sets.
- Highly fault-tolerant and is designed to be deployed on low-cost hardware

Hadoop Distributed File System-Mapreduce

Map

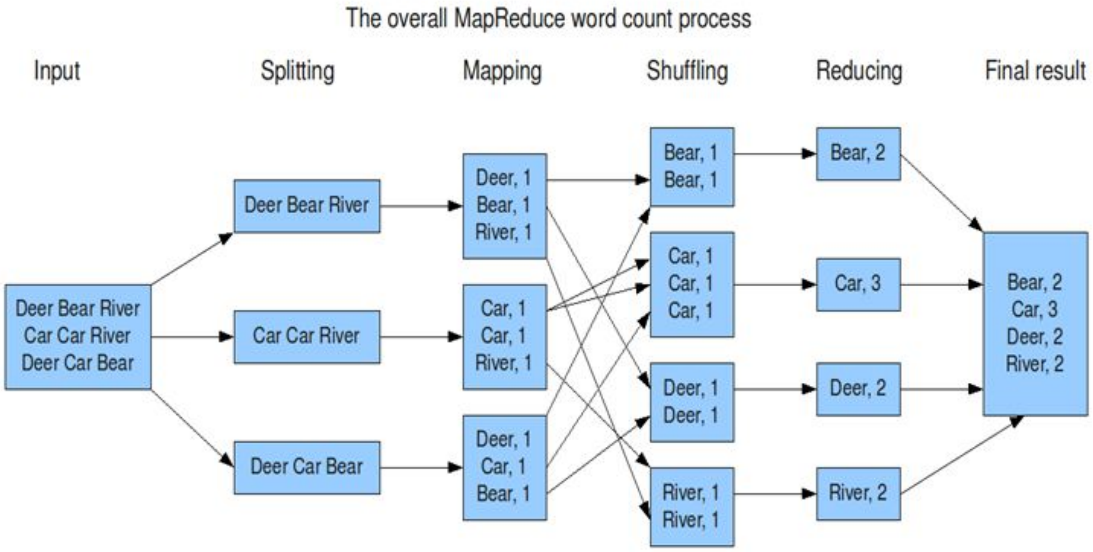
- takes a set of data and converts it into another set of data
- individual elements are broken down into tuples(key/value pairs)

Reduce

- Take output of mapper as input
- combines data tuples into a smaller set of tuples

Hadoop Distributed File System-Mapreduce

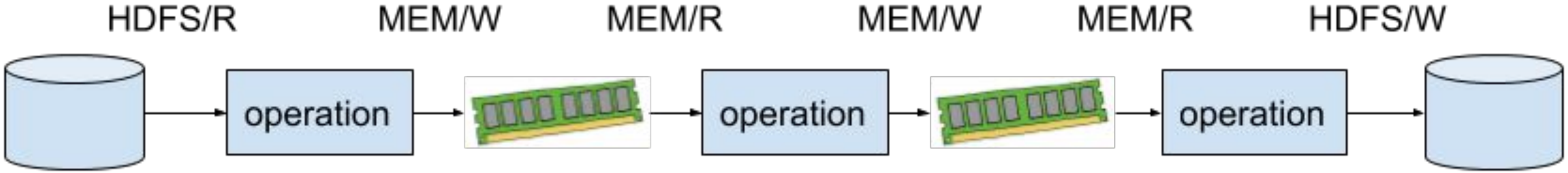
Use word count as example



Apache-Spark

When we aim to iterative computing

Solution: Resilient Distributed Datasets: RDD caching

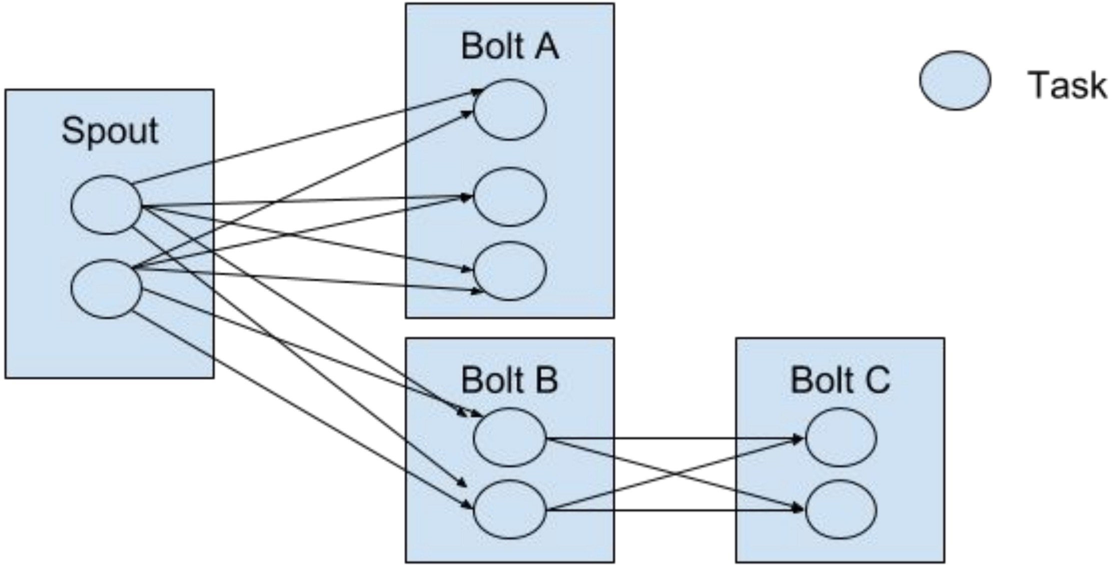


In-place memory to guarantee the locality

Apache-Storm

Build a data stream in distributed system to reduce the latency

Topology to represent data stream, composed of Spout and Bolt.



Evaluation

Design different scenario to show the advantage/ disadvantages

(1) Basic operation

Read/ Write (Append) operation

CPU bound/ IO bound program

(2) Benchmark:

Word count

Sorting

Learning algorithm: K-means/ Linear regression

Streaming Data