Hadoop
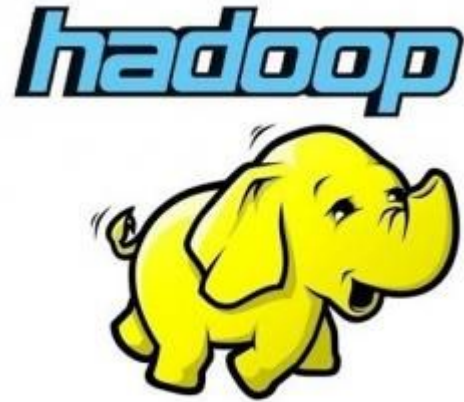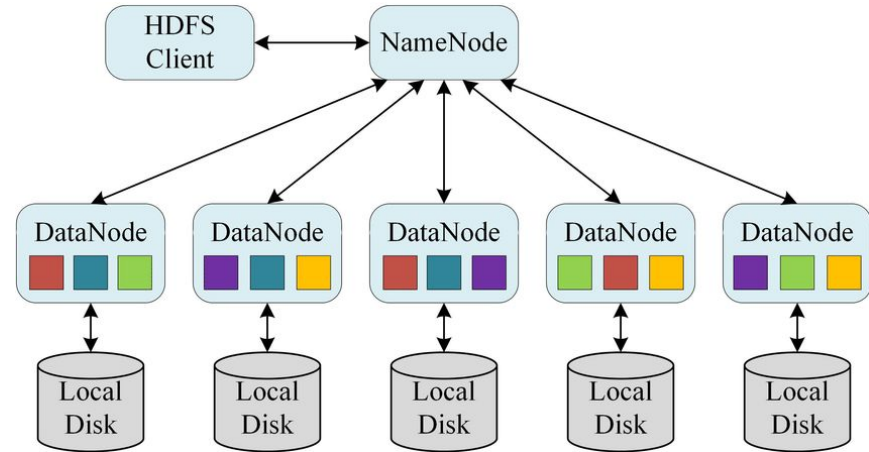
Hadoop

- Library that used to process big data
- Open source
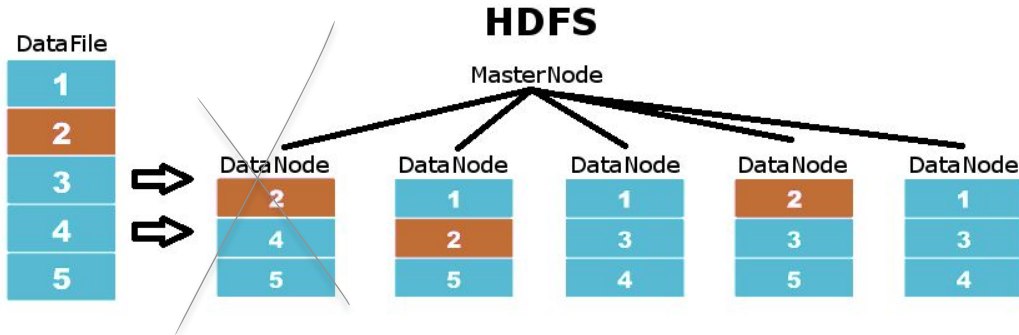- Run process with multiple machines

HDFS

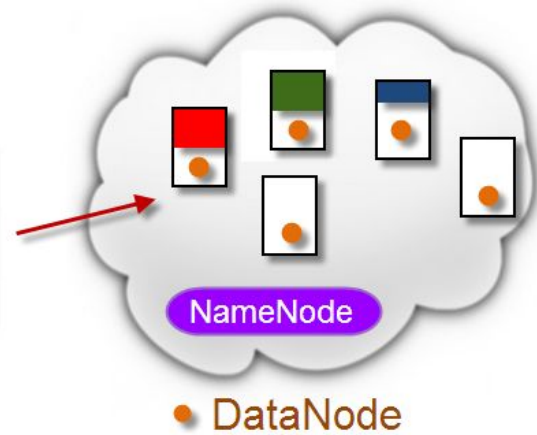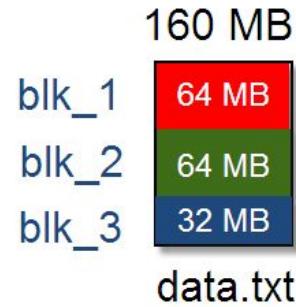
- HDFS is used to store the data
- Hadoop distribute the file when load

HDFS Blocks

- Data is distributed with bulk mode
- Replication factor
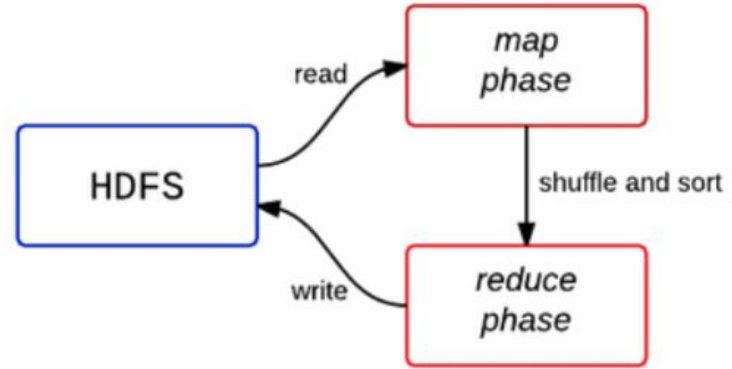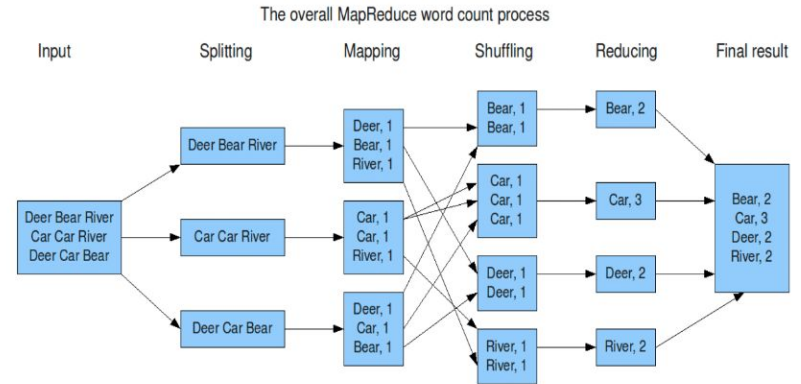
Map Reduce

- Map reduce is a library to process data
- Java , Pig and Hive are used to implement

Map Reduce

- Split : Data is splitted to 64 mb blocks
- Map: Map with key(word) and values(1)
- Shuffle : Group the same words
- Reduce : Count by keywords

The overall MapReduce word count process

| Input | Splitting | Mapping | Shuffling | Reducing | Final result |

Finish