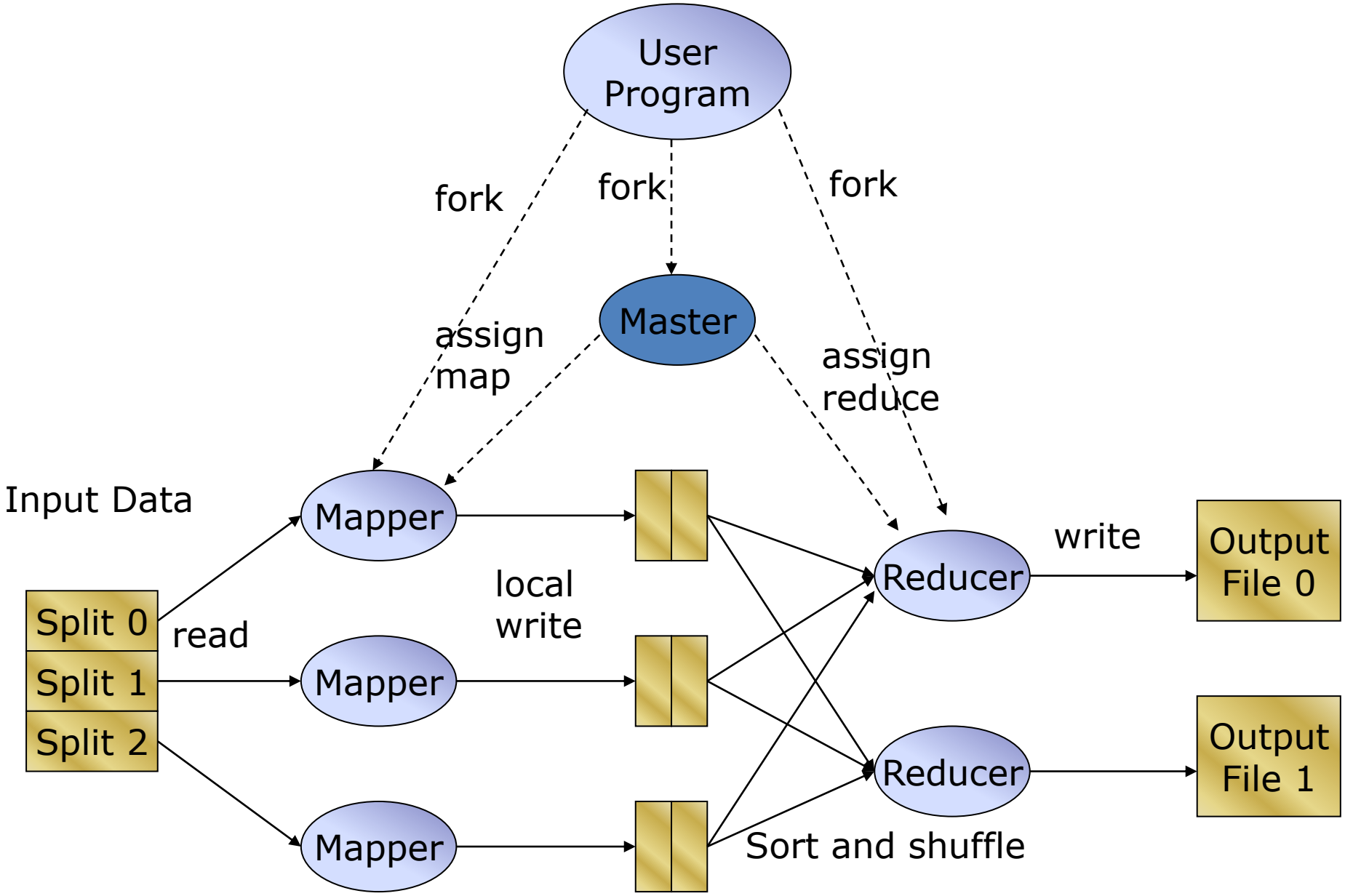
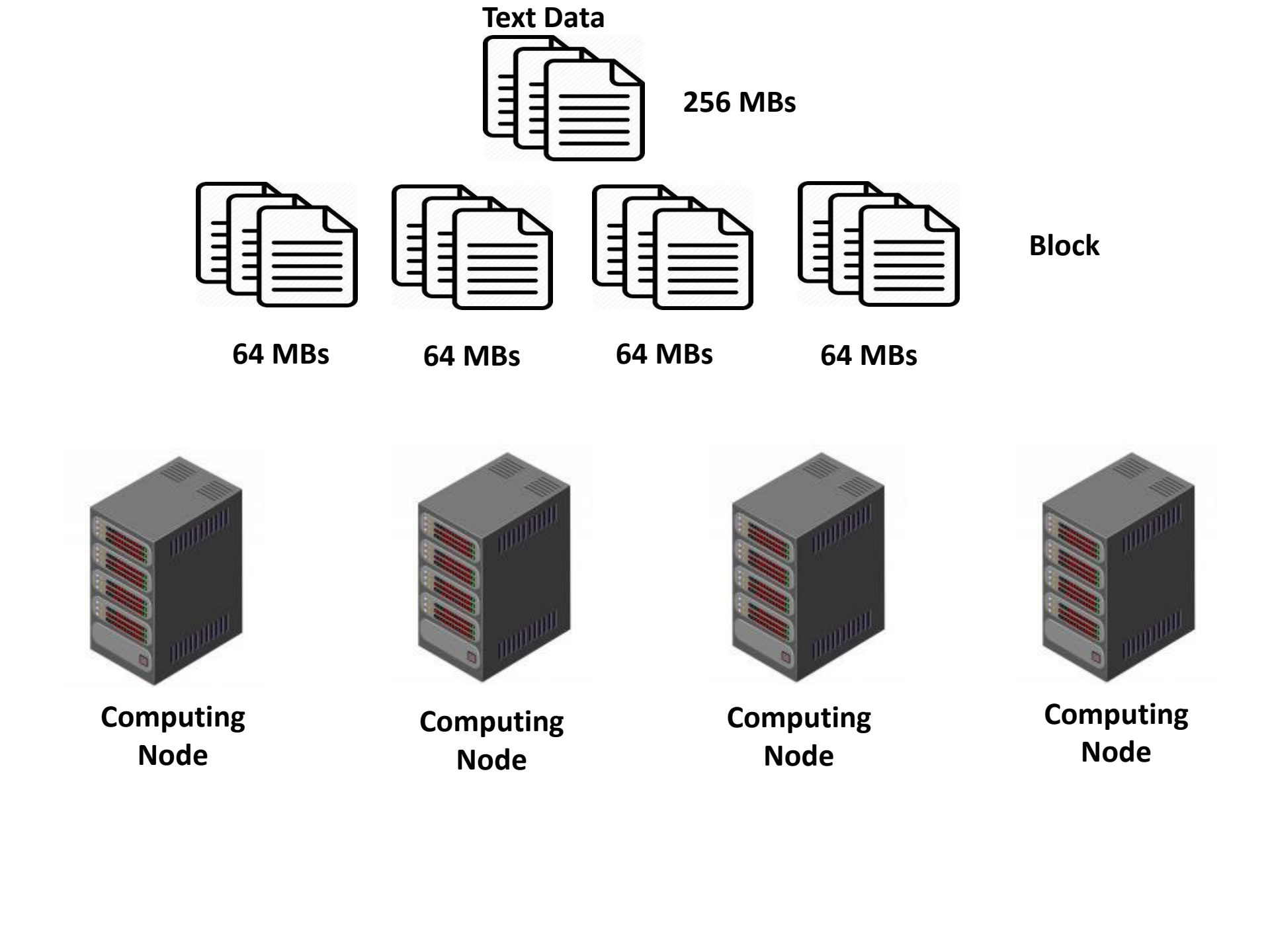
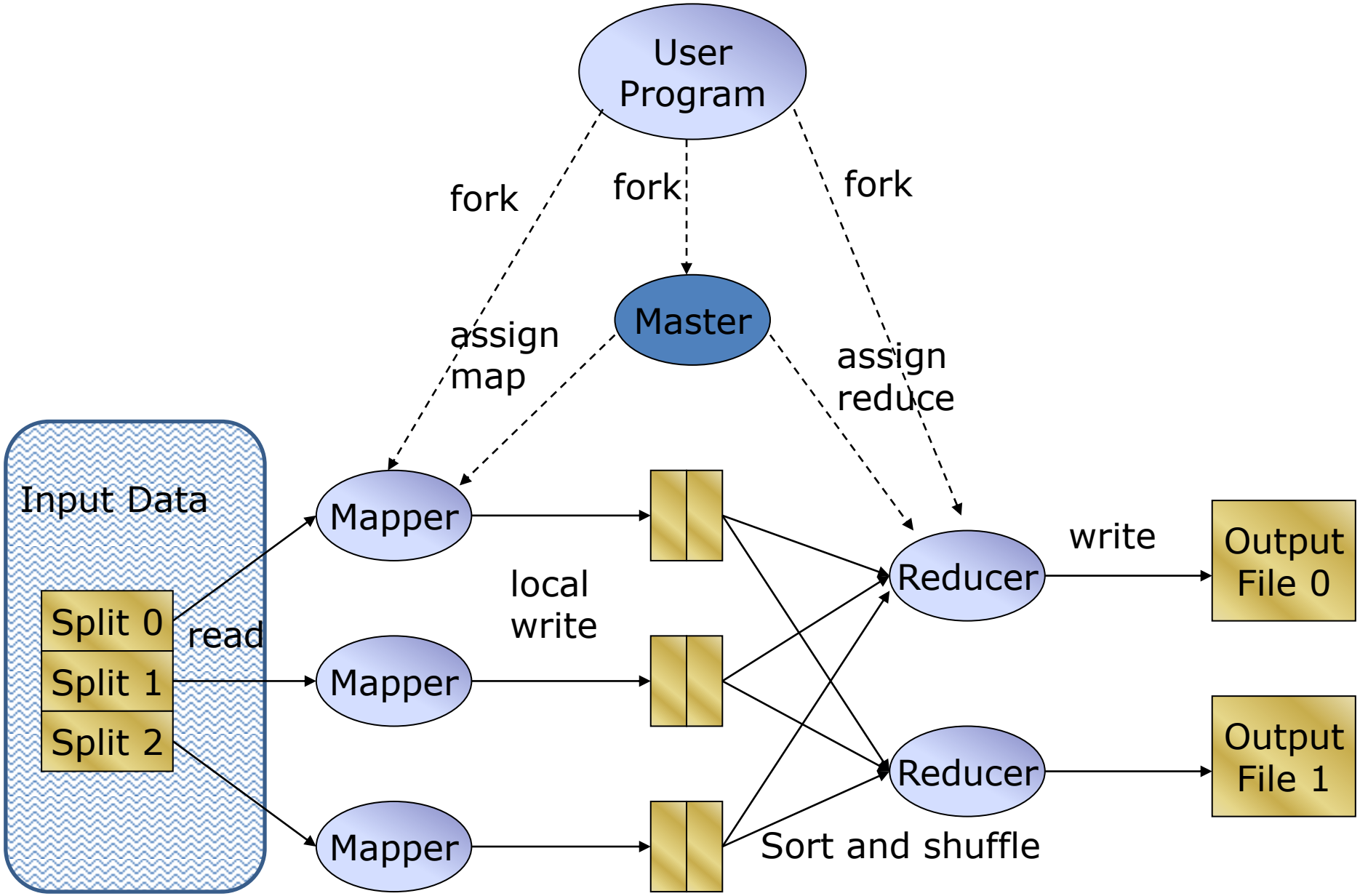
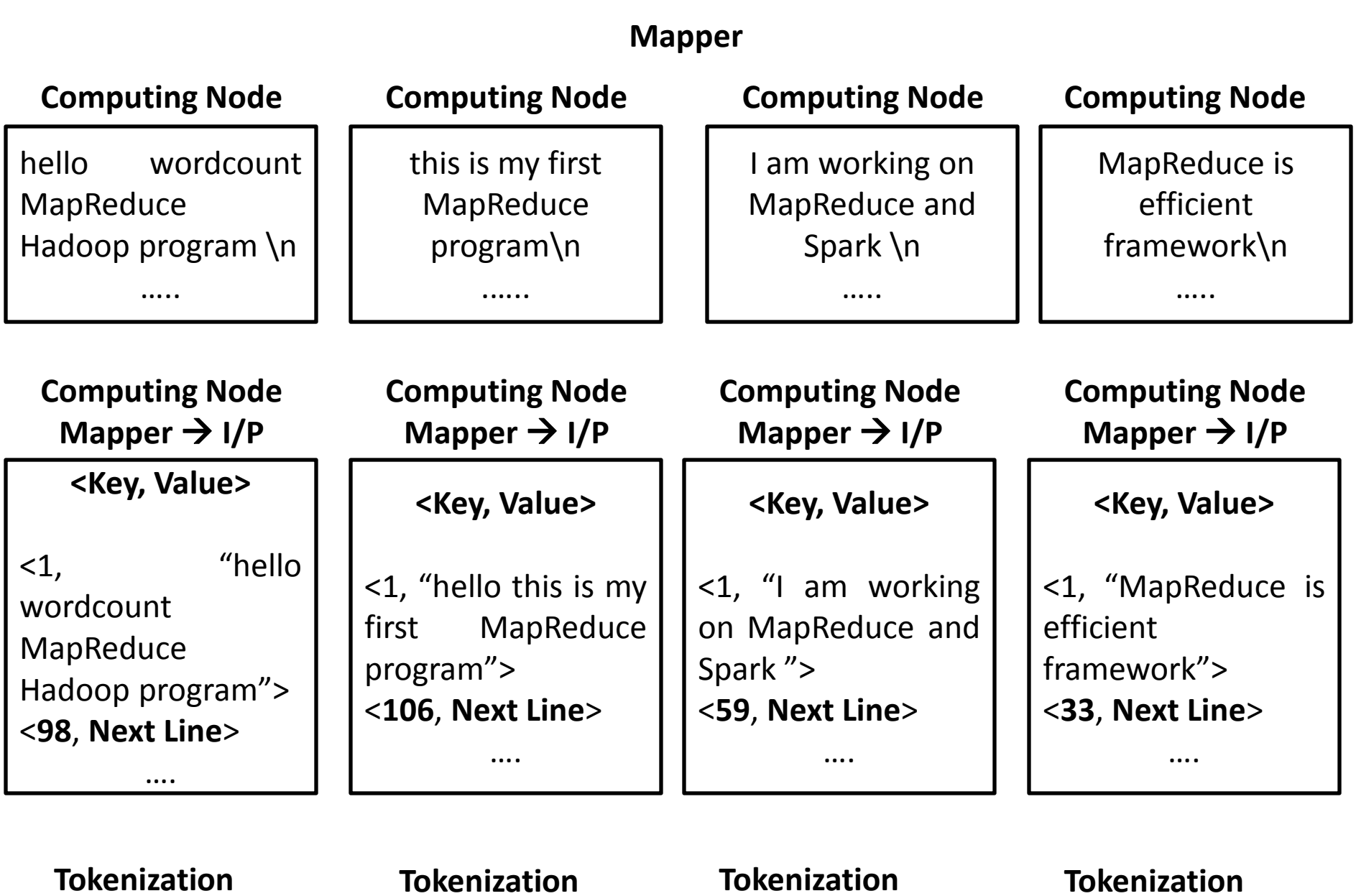
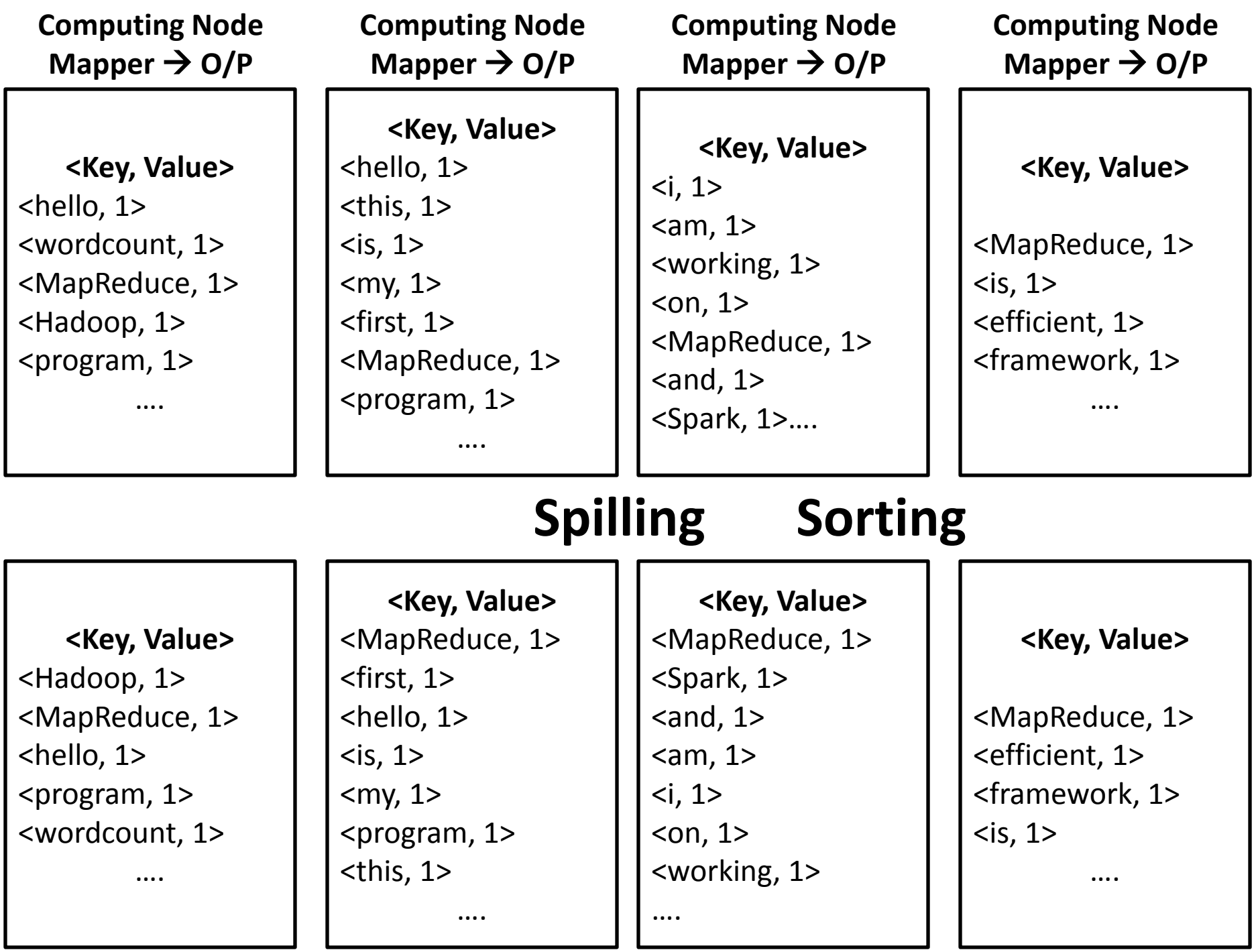
MapReduce

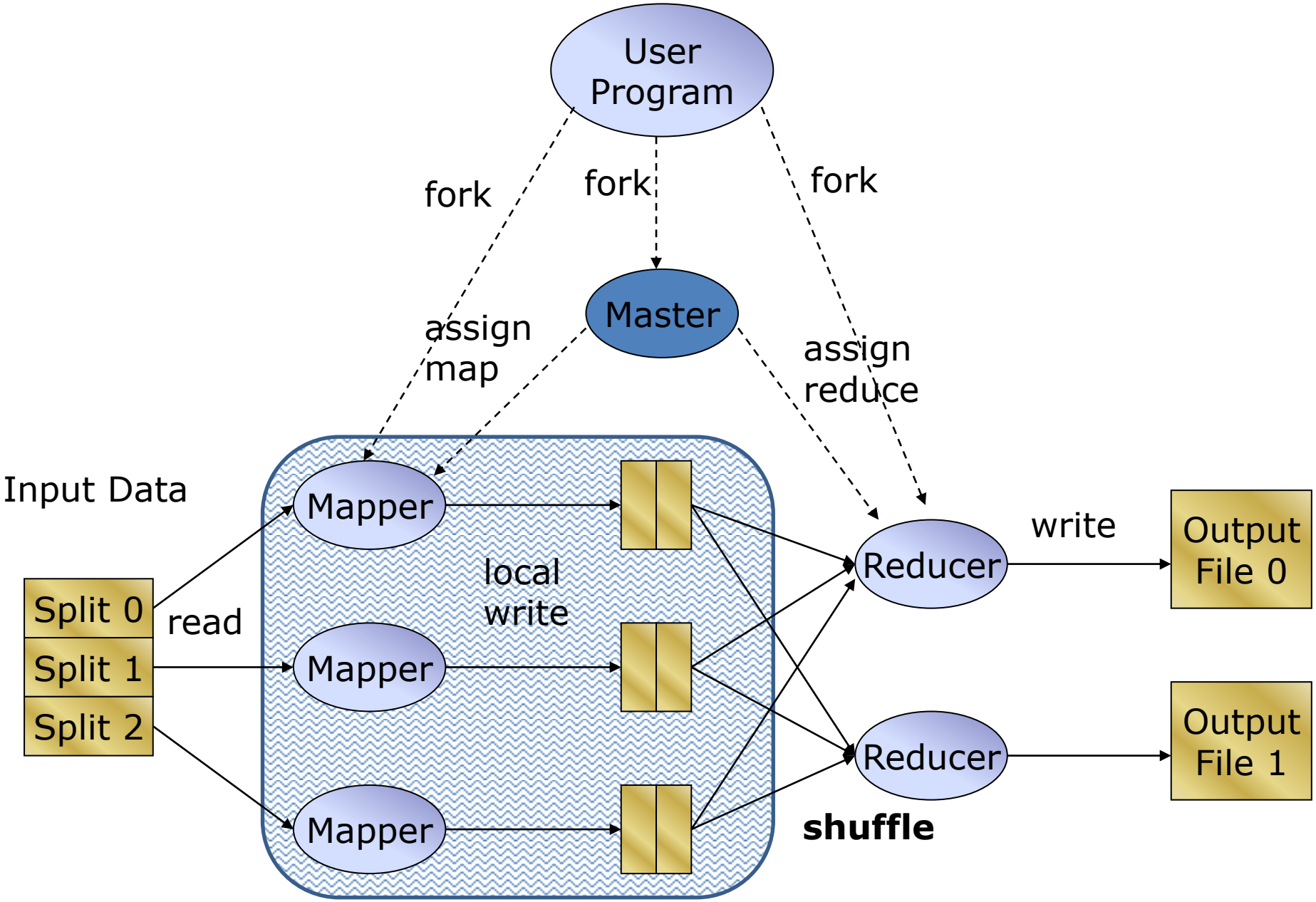

User Program
fork
fork
fork
Master
assign map
assign reduce
Input Data
Mapper
Split 0
Split 1
Split 2
read
Mapper
Mapper
local write
Reducer
Reducer
write
Output File 0
Output File 1
Sort and shuffle

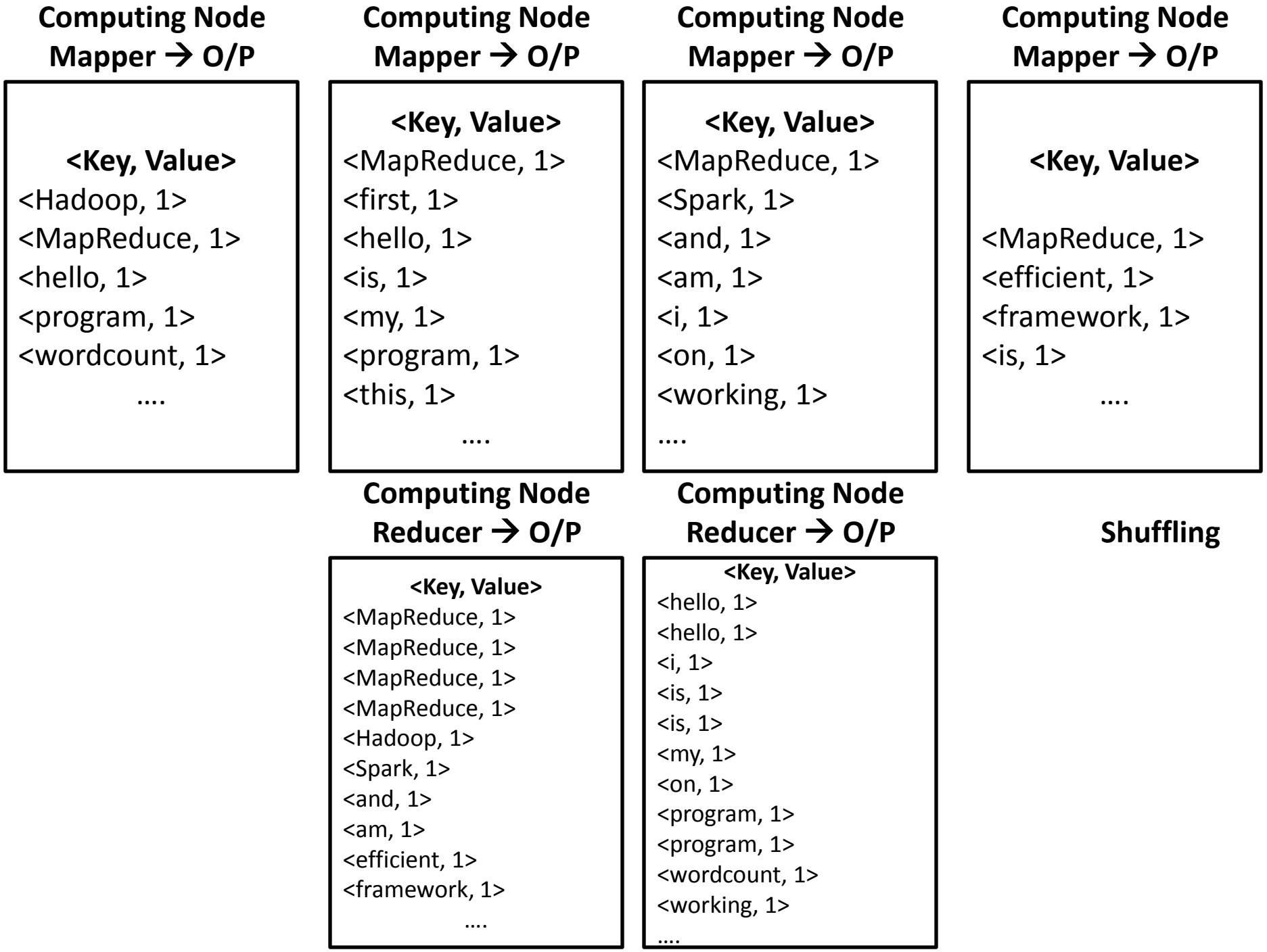

Text Data
256 MBs
Block
64 MBs
64 MBs
64 MBs
64 MBs
Computing Node
Computing Node
Computing Node
Computing Node


User Program
fork
fork
fork
Master
assign map
assign reduce
Input Data
Split 0
Split 1
Split 2
read
Mapper
Mapper
Mapper
local write
Reducer
Reducer
write
Output File 0
Output File 1
Sort and shuffle


Mapper
Computing Node
Computing Node
Computing Node
Computing Node
hello wordcount MapReduce Hadoop program \n .....
this is my first MapReduce program\n .....
I am working on MapReduce and Spark \n .....
MapReduce is efficient framework\n .....
Computing Node Mapper → I/P
Computing Node Mapper → I/P
Computing Node Mapper → I/P
Computing Node Mapper → I/P
<Key, Value>
<1, "hello wordcount MapReduce Hadoop program">
<98, Next Line>
....
<Key, Value>
<1, "hello this is my first MapReduce program">
<106, Next Line>
....
<Key, Value>
<1, "I am working on MapReduce and Spark ">
<59, Next Line>
....
<Key, Value>
<1, "MapReduce is efficient framework">
<33, Next Line>
....
Tokenization
Tokenization
Tokenization
Tokenization


Computing Node
Mapper → O/P
<Key, Value>
<hello, 1>
<wordcount, 1>
<MapReduce, 1>
<Hadoop, 1>
<program, 1>
….
Computing Node
Mapper → O/P
<Key, Value>
<hello, 1>
<this, 1>
<is, 1>
<my, 1>
<first, 1>
<MapReduce, 1>
<program, 1>
….
Computing Node
Mapper → O/P
<Key, Value>
<i, 1>
<am, 1>
<working, 1>
<on, 1>
<MapReduce, 1>
<and, 1>
<Spark, 1>….
Computing Node
Mapper → O/P
<Key, Value>
<MapReduce, 1>
<is, 1>
<efficient, 1>
<framework, 1>
….
Spilling
Sorting
<Key, Value>
<Hadoop, 1>
<MapReduce, 1>
<hello, 1>
<program, 1>
<wordcount, 1>
….
<Key, Value>
<MapReduce, 1>
<first, 1>
<hello, 1>
<is, 1>
<my, 1>
<program, 1>
<this, 1>
….
<Key, Value>
<MapReduce, 1>
<Spark, 1>
<and, 1>
<am, 1>
<i, 1>
<on, 1>
<working, 1>
….
<Key, Value>
<MapReduce, 1>
<efficient, 1>
<framework, 1>
<is, 1>
….


User Program
fork
fork
fork
Master
assign map
assign reduce
Input Data
Mapper
Mapper
Mapper
local write
Split 0
Split 1
Split 2
read
Reducer
Reducer
write
Output File 0
Output File 1
shuffle


Computing Node
Mapper → O/P
<Key, Value>
<Hadoop, 1>
<MapReduce, 1>
<hello, 1>
<program, 1>
<wordcount, 1>
….
Computing Node
Mapper → O/P
<Key, Value>
<MapReduce, 1>
<first, 1>
<hello, 1>
<is, 1>
<my, 1>
<program, 1>
<this, 1>
….
Computing Node
Mapper → O/P
<Key, Value>
<MapReduce, 1>
<Spark, 1>
<and, 1>
<am, 1>
<i, 1>
<on, 1>
<working, 1>
….
Computing Node
Mapper → O/P
<Key, Value>
<MapReduce, 1>
<efficient, 1>
<framework, 1>
<is, 1>
….
Computing Node
Reducer → O/P
<Key, Value>
<MapReduce, 1>
<MapReduce, 1>
<MapReduce, 1>
<MapReduce, 1>
<Hadoop, 1>
<Spark, 1>
<and, 1>
<am, 1>
<efficient, 1>
<framework, 1>
….
Computing Node
Reducer → O/P
<Key, Value>
<hello, 1>
<hello, 1>
<i, 1>
<is, 1>
<is, 1>
<my, 1>
<on, 1>
<program, 1>
<program, 1>
<wordcount, 1>
<working, 1>
….
Shuffling

Computing Node
Reducer → O/P

<Key, Value>
<MapReduce, 1>
<MapReduce, 1>
<MapReduce, 1>
<MapReduce, 1>
<Hadoop, 1>
<Spark, 1>
<and, 1>
<am, 1>
<efficient, 1>
<framework, 1>
….

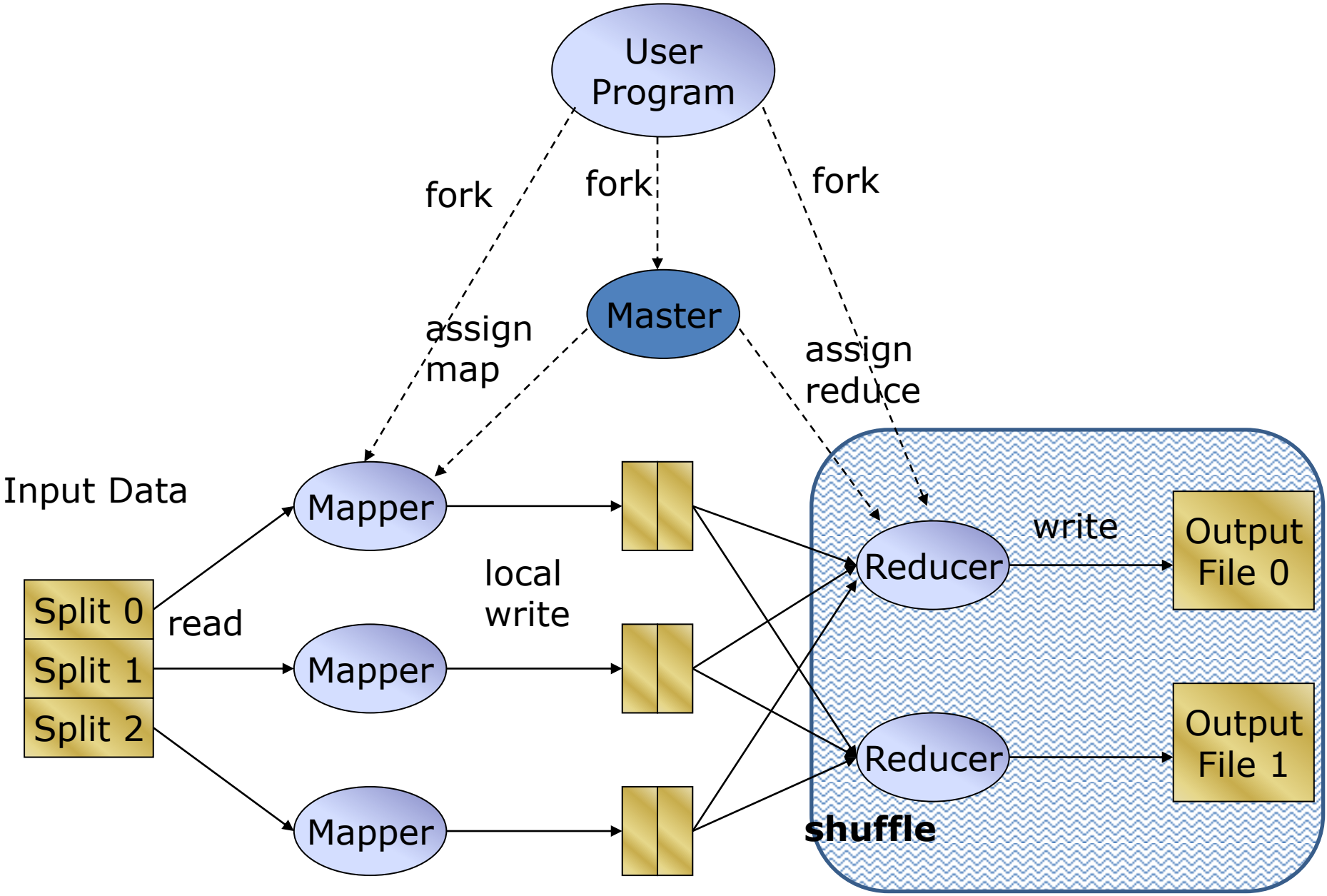
Computing Node
Reducer → O/P

<Key, Value>
<hello, 1>
<hello, 1>
<i, 1>
<is, 1>
<is, 1>
<my, 1>
<on, 1>
<program, 1>
<program, 1>
<wordcount, 1>
<working, 1>
….

<Key, Value>
<MapReduce, 4>
<Hadoop, 1>
<Spark, 1>
<and, 1>
<am, 1>
<efficient, 1>
<framework, 1>
….

<Key, Value>
<hello, 2>
<i, 1>
<is, 2>
<my, 1>
<on, 1>
<program, 2>
<wordcount, 1>
<working, 1>
….

HDFS


User Program
fork
fork
fork
Master
assign map
assign reduce
Input Data
Mapper
Mapper
Mapper
Split 0
Split 1
Split 2
read
local write
Reducer
Reducer
write
Output File 0
Output File 1
shuffle

- No of Mapper and Reducer
- Key Parameter in MapReduce Design
  - Key-Value pair
  - Load Balance b/t Mapper and Reducer