

# 4. Map Reduce & HDFS 명령어

## 목 차

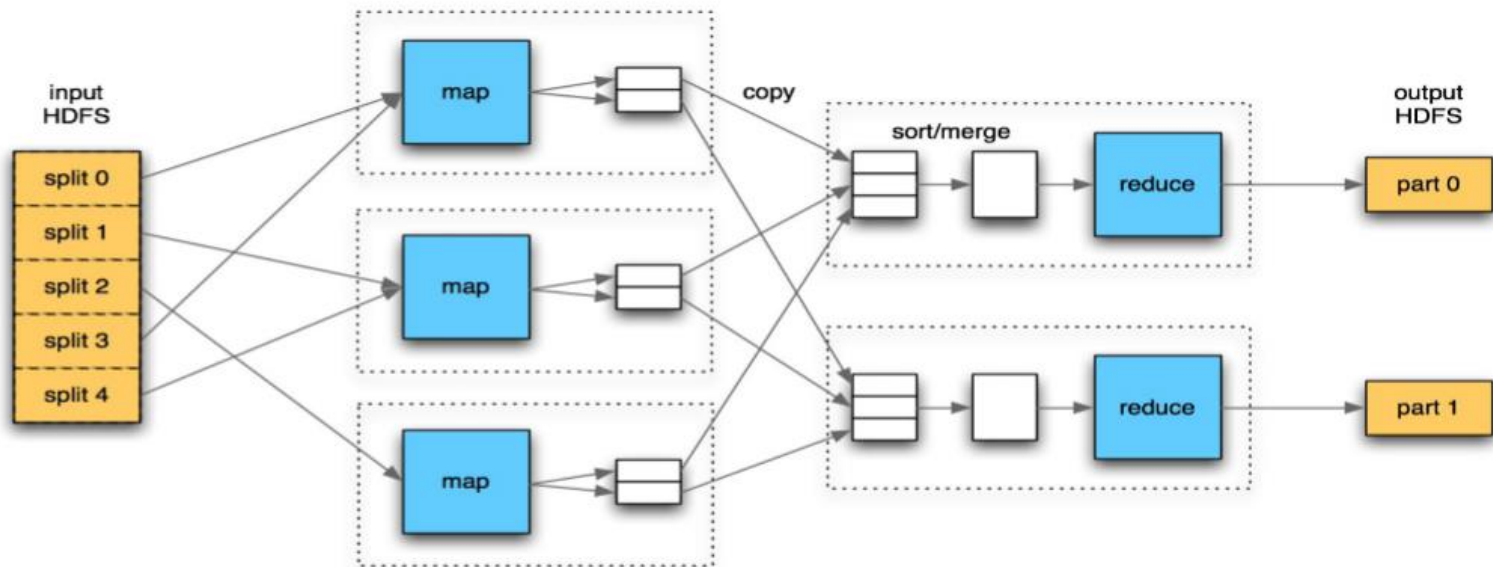
1. Map Reduce 개요
2. Hadoop/Yarn/Historyserver 시작
3. HDFS 명령어
4. Map Reduce Word Count 실습
5. Hadoop/Yarn/Historyserver 종료

# 1. MapReduce 개요

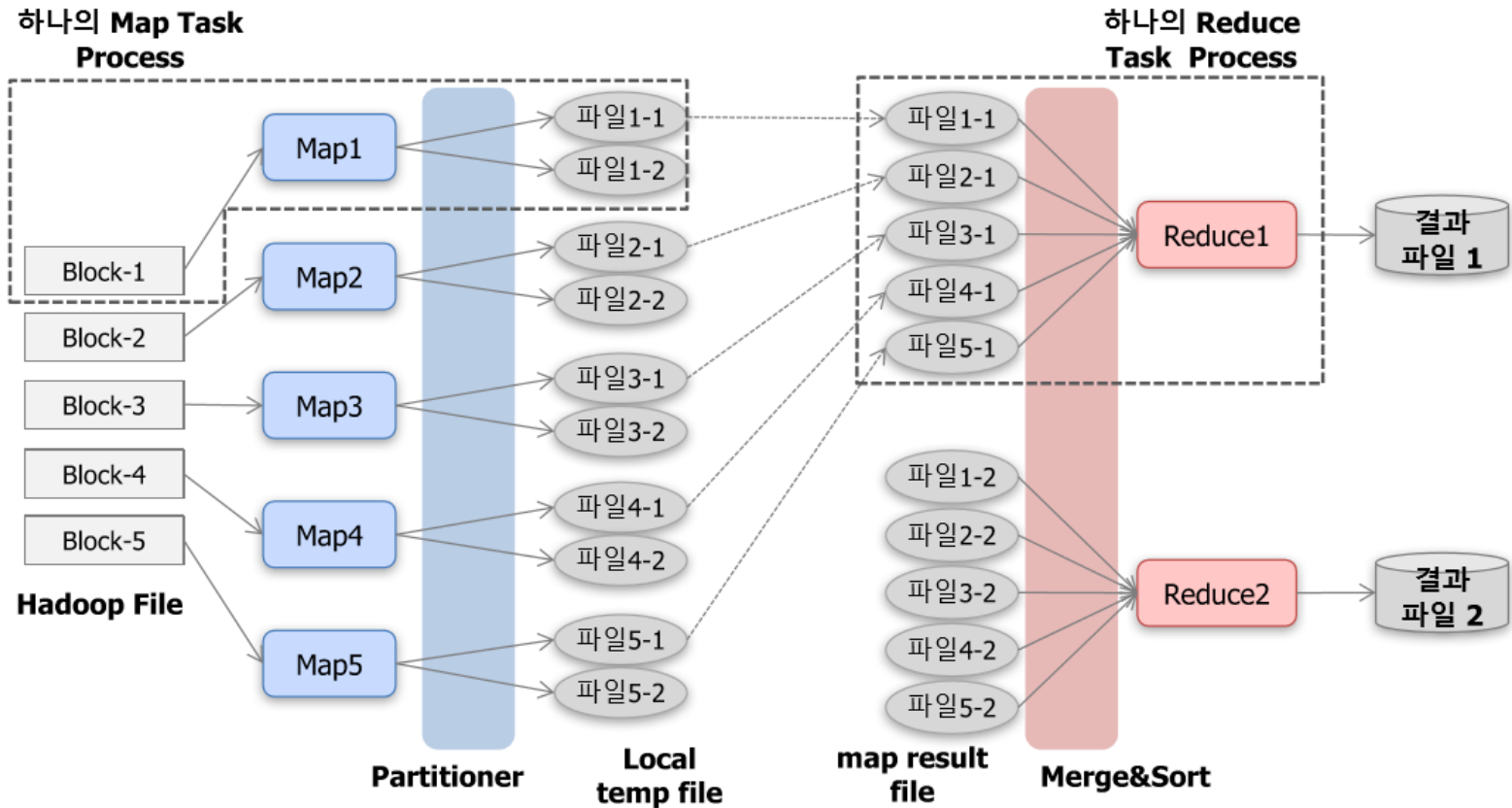
- HDFS 파일 대상 분산배치분석 지원 프레임워크
- 애플리케이션 구현 시 데이터 전송, 분산 처리, 내고장성 등의 복잡한 처리 담당
- 맵(Map)과 리듀스(Reduce) 두 단계 처리
  - ✓ 맵 : 입력 파일 한 줄 읽기 → 데이터 변형
  - ✓ 리듀스 : 맵의 결과 집계(Aggregation)
- 애플리케이션 예
  - ✓ Word Counter

# 1. MapReduce 개요

- 효과적인 분산 컴퓨팅을 위한 프로그래밍 모델
- Unix Pipeline 과 유사한 동작 방식

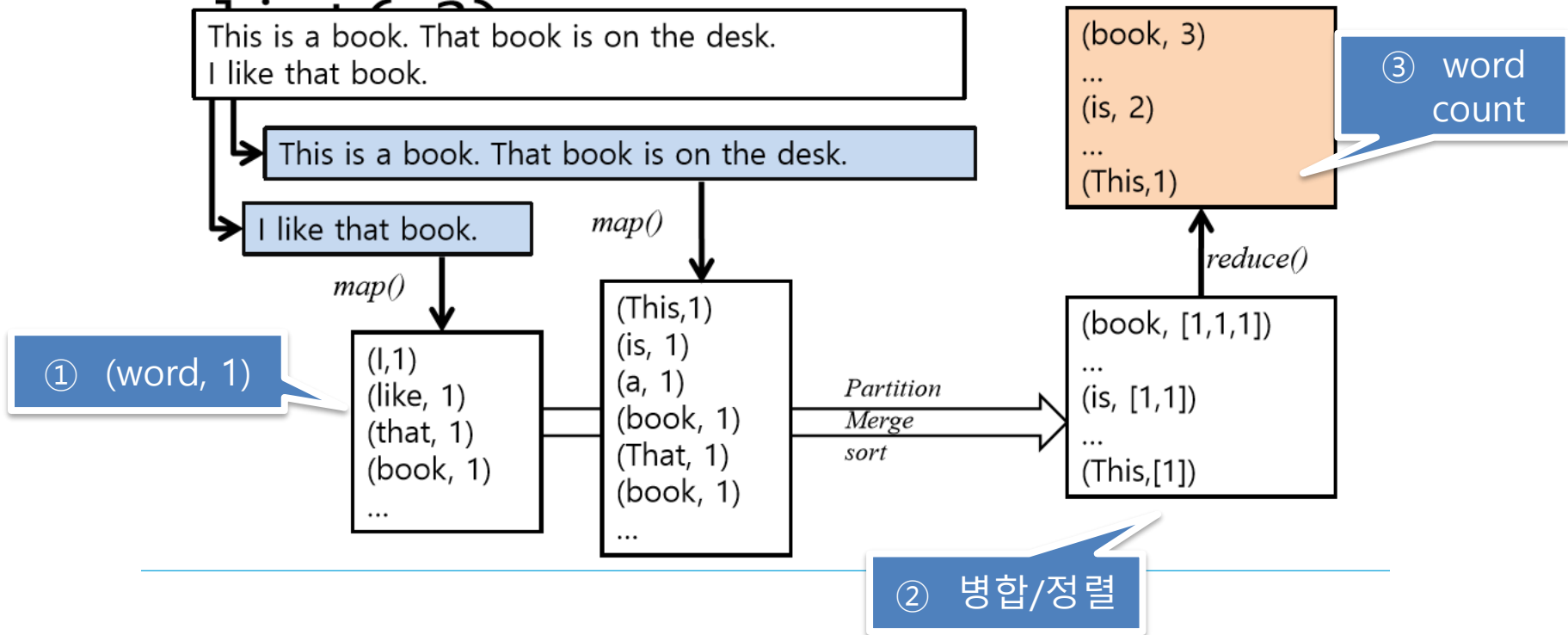


# MapReduce 데이터 흐름



# MapReduce Sample

- `map (k1,v1) → list(k2,v2)`
- `reduce (k2, list (v2)) →`

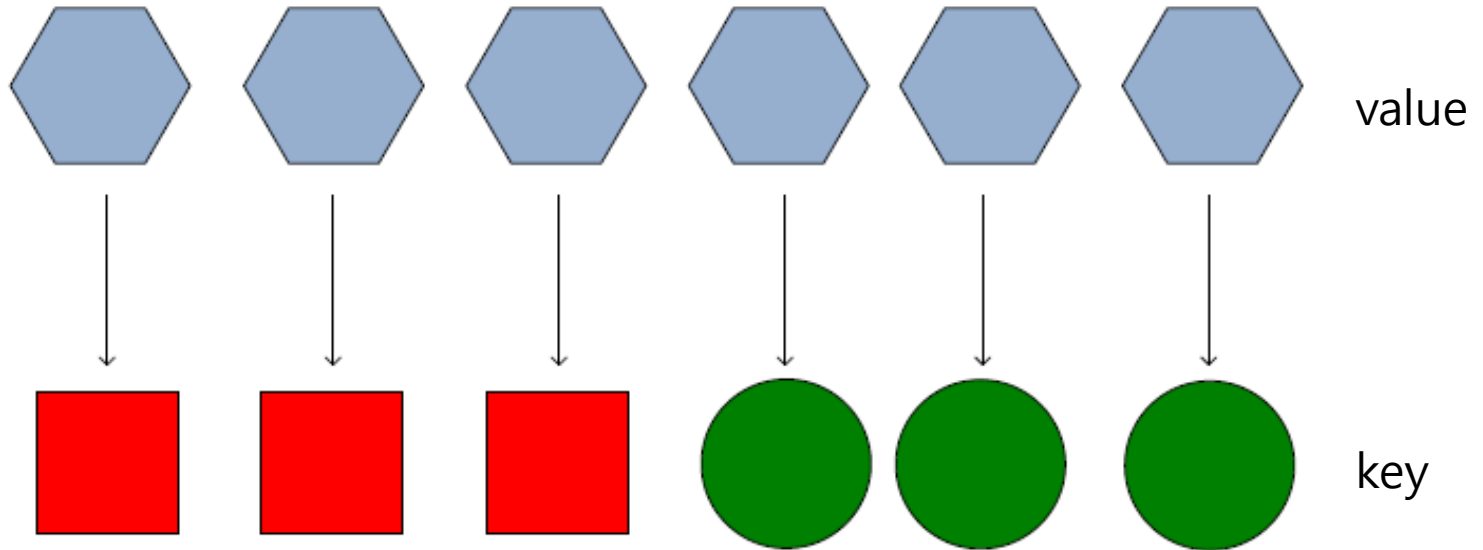


## 1) map

- 데이터 소스로부터 레코드(파일의 라인이나 DB의 Row 등)들을 읽어서 (key, value) 쌍으로 map 함수에게 보냄
- 맵 함수는 입력 레코드를 받아서 하나 이상의 (Key, Value) 형식의 중간데이터를 만들어내서 로컬 파일시스템에 저장(local temp file)

# map

```
map (in_key, in_value) ->  
    (out_key, intermediate_value) list
```



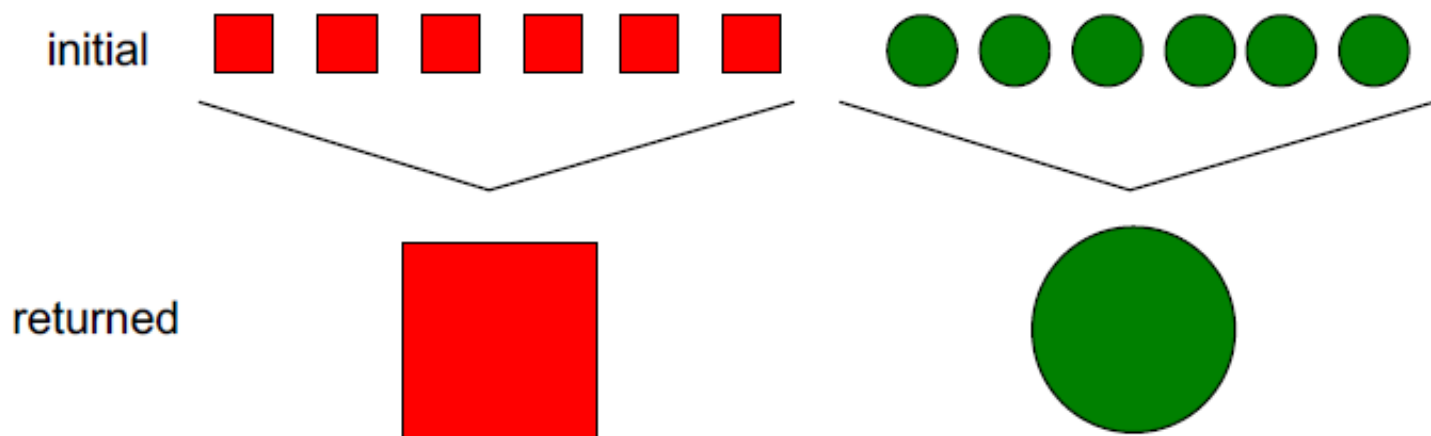
## 2) reduce

- map 단계가 끝난 후에는 동일한 output key 를 가진 모든 중간 값들은 리스트로서 결합됨
- reduce 함수에는 전달된 output key 와 중간값들의 리스트를 가지고 최종 결과 값을 만듬 (보통 key, value 쌍)

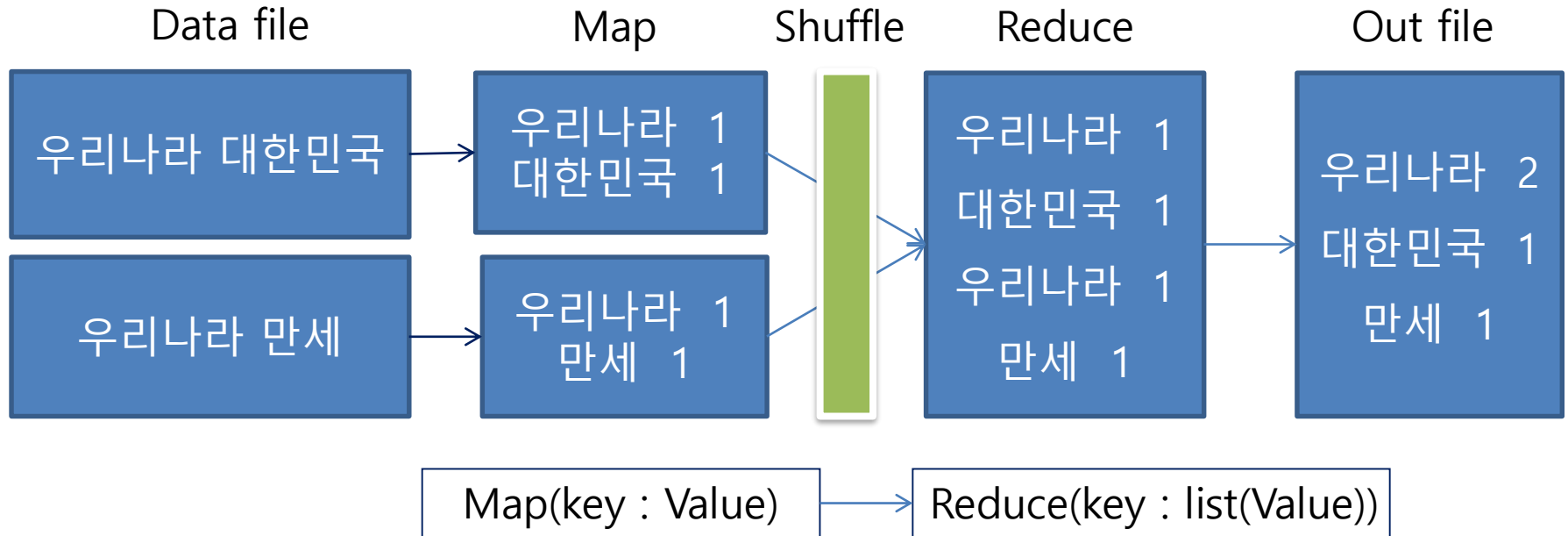


# reduce

`reduce (out_key, intermediate_value list) ->`  
`out_value list`



# Word Counter 애플리케이션 예



- ✓ Map : 입력 파일 한 줄 읽기 → 데이터 변형
- ✓ Shuffle : Map의 중간 데이터를 Reduce 단계로 전달(파티셔닝, 병합, 정렬)
- ✓ Reduce : Map의 결과 집계

## 2. Hadoop/Yarn/Historyserver 시작

맵리듀스, 하이브, 스파크 등  
의 애플리케이션은 안에서  
작업 실행

- 1. Hadoop/Yarn/Historyserver 시작

```
[hadoop@master ~]$ start-all.sh # 하둡/얀 시작
```

```
[hadoop@master ~]$ mr-jobhistory-daemon.sh start historyserver # 데몬 실행
```

\* Hadoop 상태 확인

```
[hadoop@master ~]$ jps
```

### 3. HDFS 명령어

명령어 형식) \$**hdfs** **dfs** -명령어 <인수>

명령어	기능
hdfs dfs -cat	HDFS의 특정 파일 내용 보기
hdfs dfs -put	로컬 시스템의 파일을 HDFS에 업로드
hdfs dfs -get	HDFS 파일을 로컬 시스템으로 다운로드
hdfs dfs -cp	HDFS 파일을 목적지로 복사
hdfs dfs -ls	파일과 디렉터리를 조회한다.
hdfs dfs -mkdir	디렉터리 생성
hdfs dfs -rmdir	디렉터리 삭제
hdfs dfs -rm -R	디렉터리+파일 동시 삭제
hdfs dfs -rm	파일 삭제
hdfs dfs -count	파일/디렉터리 이름, 파일 수, 디렉터리 수 출력
hdfs dfs -chmod	파일과 디렉터리에 대한 접근 권한 변경

# 4. Map Reduce Word Count 실습

## 1) Word Count 데이터 파일 준비

```
hadoop@master:~/hadoop-2.7.1
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
# /test 디렉터리 생성
[hadoop@master ~]$ hdfs dfs -mkdir /test

# ./NOTICE.txt 파일을 test 디렉터리에 복사
[hadoop@master ~]$ hdfs dfs -put ./hadoop-2.9.2/NOTICE.txt /test
[hadoop@master ~]$ hdfs dfs -cat /test/NOTICE.txt
```

Hadoop 명령어 형식)  
\$**hdfs dfs** -명령어 <인수>

1. HDFS 디렉터리(test) 만들기
2. HDFS 파일(NOTICE.txt) 올리기
3. HDFS 파일 내용 보기

## 2) Word Count 실행

hadoop jar /디렉토리/\*.jar 파라미터 대상파일 출력디렉토리

```
[hadoop@master ~]$ hadoop jar /hadoop-2.9.2/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar wordcount /test/NOTICE.txt /output
```

```
17/04/10 16:44:29 INFO input.FileInputFormat: Total input paths to process : 1
17/04/10 16:44:29 INFO mapreduce.JobSubmitter: number of splits:1
17/04/10 16:44:30 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1491803758644_0001
17/04/10 16:44:32 INFO impl.YarnClientImpl: Submitted application application_1491803758644_0001
17/04/10 16:44:32 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1491803758644_0001/
17/04/10 16:44:32 INFO mapreduce.Job: Running job: job_1491803758644_0001
17/04/10 16:45:40 INFO mapreduce.Job: Job job_1491803758644_0001 is in state: COMPLETED
: false
```

Map & Reduce 작업 상태

```
17/04/10 16:45:41 INFO mapreduce.Job: map 0% reduce 0%
17/04/10 16:46:08 INFO mapreduce.Job: map 100% reduce 0%
17/04/10 16:46:24 INFO mapreduce.Job: map 100% reduce 100%
17/04/10 16:46:26 INFO mapreduce.Job: Job job_1491803758644_0001 completed successfully
17/04/10 16:46:28 INFO mapreduce.Job: Counters: 49
    File System Counters
        FILE: Number of bytes read=173
        FILE: Number of bytes written=231359
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
```

hadoop@master:~/hadoop-2.7.1

파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)

```
Combine output records=11
Reduce input groups=11
Reduce shuffle bytes=173
Reduce input records=11
Reduce output records=11
Spilled Records=22
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=175
CPU time spent (ms)=1150
Physical memory (bytes) snapshot=277086208
Virtual memory (bytes) snapshot=4200316928
Total committed heap usage (bytes)=137498624

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=101
File Output Format Counters
  Bytes Written=123
```

[hadoop@master hadoop-2.7.1]\$

### 3) Word Count 결과보기

```
hadoop@master:~/hadoop-2.7.1
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)

[hadoop@master ~]$ hdfs dfs -ls /output
-rw-r--r-- 3 hadoop supergroup 0 2017-04-10 16:46 /output/_SUCCESS
-rw-r--r-- 3 hadoop supergroup 123 2017-04-10 16:46 /output/part-r-00000

[hadoop@master ~]$ hdfs dfs -cat /output/part-r-00000
Apache 1
Foundation 1
Software 1
The 1
This 1
by 1
developed 1
includes 1
product 1
software 1
[hadoop@master hadoop-2.7.1]$
```

워드 카운터 실행 결과

❖ HDFS에서 로컬 파일 시스템으로 파일 복사

```
hdfs dfs -get /output/part-r-00000 ~/hfile/word_count.txt
```



http://localhost:8088

All Applications

localhost:8088/cluster



## All Applications

### Cluster

[About](#)

[Nodes](#)

[Node Labels](#)

[Applications](#)

[NEW](#)

[NEW SAVING](#)

[SUBMITTED](#)

[ACCEPTED](#)

[RUNNING](#)

[FINISHED](#)

[FAILED](#)

[KILLED](#)

[Scheduler](#)

Tools

### Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	...
1	0	0	1	0	0 B	24 GB	0 B	0	24	0	3

### Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime	...
<a href="#">application_1491803758644_0001</a>	hadoop	word count	MAPREDUCE	default	Mon Apr 10 16:44:31 +0900 2017	Mon Apr 10 16:46:24 +0900 2017	...

Showing 1 to 1 of 1 entries

Master - VMware Workstation 12 Player (Non-commercial use only)

Player ▾ | [Icons] | ko ▾ (화) 15:04 [Icons]


프로그램 ▾ 위치 ▾ Firefox 웹 브라우저 ▾

Application application\_1502171388400\_0001 - Mozilla Firefox

Application application\_1... x +

localhost:8088/cluster/app/application\_1502171388400\_0001 | 🔍 검색 | ☆ | 📁 | 📧 | ⬇️ | 🏠 | ☰

Logged in as: dr.who



# Application application\_1502171388400\_0001

▼ Cluster

- About
- Nodes
- Node Labels
- Applications
  - NEW
  - NEW\_SAVING
  - SUBMITTED
  - ACCEPTED
  - RUNNING
  - FINISHED
  - FAILED
  - KILLED
- Scheduler

► Tools

Kill Application

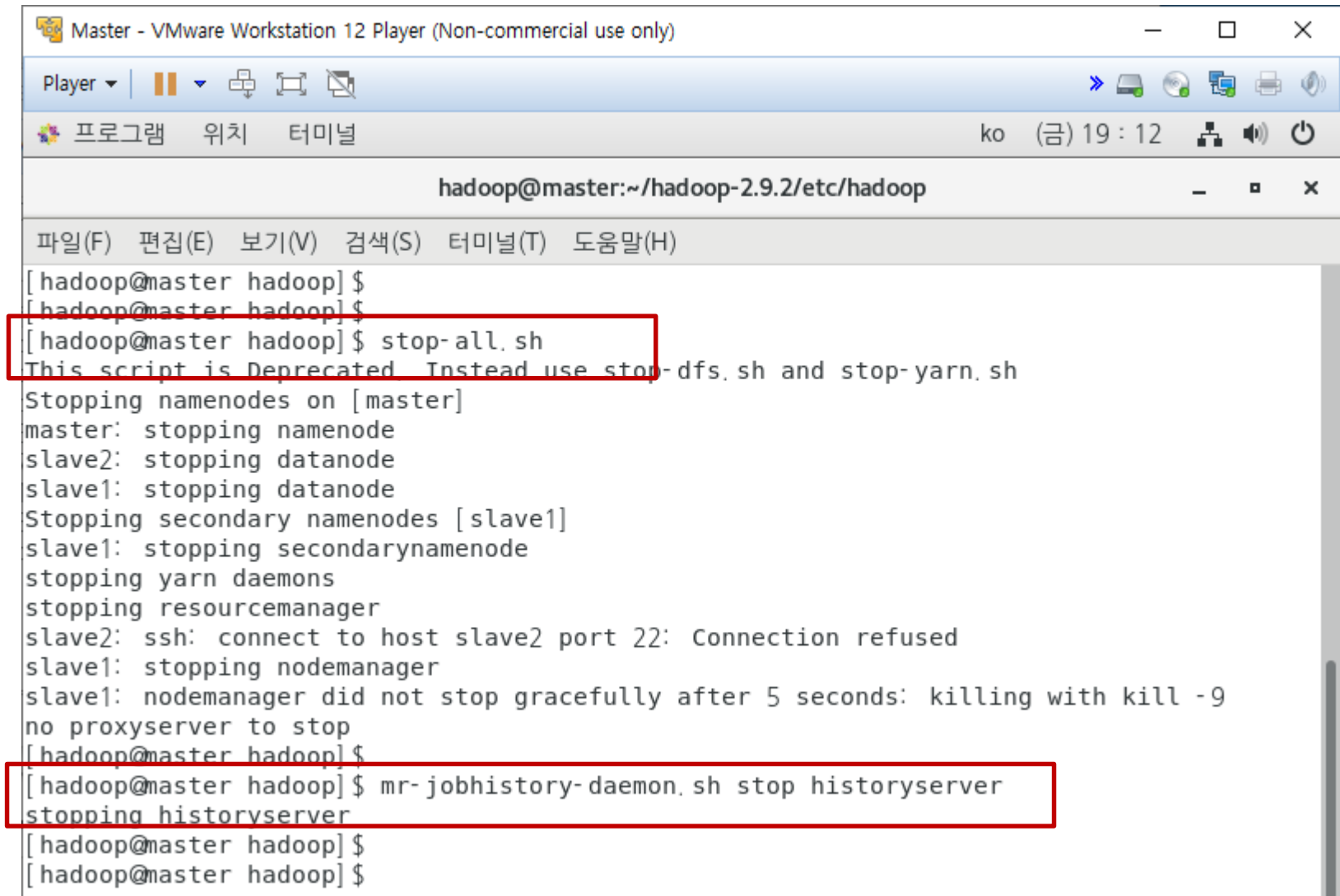
Application Overview	
User:	hadoop
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
YarnApplicationState:	FINISHED
FinalStatus Reported by AM:	SUCCEEDED
Started:	화 8월 08 14:59:48 +0900 2017
Elapsed:	1mins, 39sec
Tracking URL:	<a href="#">History</a>
Diagnostics:	

---

Application Metrics	
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current	0

hadoop@master:~/hadoop-2.7.1 | Application application\_1502171388400\_0001 | 1 / 4 1

# 5. Hadoop/Yarn/Historyserver 종료



The screenshot shows a terminal window titled "Master - VMware Workstation 12 Player (Non-commercial use only)". The terminal prompt is "hadoop@master:~/hadoop-2.9.2/etc/hadoop". The terminal output shows the execution of "stop-all.sh" and "mr-jobhistory-daemon.sh stop historyserver". The "stop-all.sh" command is highlighted with a red box, and its output is also highlighted with a red box. The "mr-jobhistory-daemon.sh stop historyserver" command is also highlighted with a red box.

```
hadoop@master:~/hadoop-2.9.2/etc/hadoop
[hadop@master hadoop]$
[hadop@master hadoop]$
[hadop@master hadoop]$ stop-all.sh
This script is Deprecated. Instead use stop-dfs.sh and stop-yarn.sh
Stopping namenodes on [master]
master: stopping namenode
slave2: stopping datanode
slave1: stopping datanode
Stopping secondary namenodes [slave1]
slave1: stopping secondarynamenode
stopping yarn daemons
stopping resourcemanager
slave2: ssh: connect to host slave2 port 22: Connection refused
slave1: stopping nodemanager
slave1: nodemanager did not stop gracefully after 5 seconds: killing with kill -9
no proxyserver to stop
[hadop@master hadoop]$
[hadop@master hadoop]$ mr-jobhistory-daemon.sh stop historyserver
stopping historyserver
[hadop@master hadoop]$
[hadop@master hadoop]$
```