

실무에 적용 가능한 Big Data 분석 개론

---

# 하둡 활용



한국기술교육대학교  
온라인평생교육원

## ■ 하둡 설치

### 1. 하둡 구현을 위한 환경



자유 소프트웨어와  
오픈 소스 개발의 가장 유명한  
표본으로 하둡을 구동시키는데  
안정적임



현재 전 세계 90%의  
개인용 컴퓨터에서 쓰고 있으며,  
서버용 운영 체제로도 영역을  
넓혀 나가고 있음



#6



가상머신 소프트웨어

- 하나의 컴퓨터에 설치하여 여러 대의 컴퓨터에 여러 개의 운영체제를 구동시키는 것과 유사한 효과를 제공



사용자 편의성에 초점을 맞춘  
리눅스 배포판

- 컴퓨터에서 프로그램과 주변기기를 사용할 수 있도록 해주는 운영체제 중 하나
- 리눅스에 기반한 운영체제로 모바일과 데스크톱 PC, 서버에 우분투 운영체제를 설치해 사용

## ■ 하둡 설치

### 1. 하둡 구현을 위한 환경

#### VMware와 우분투 설치

##### 1 VMware 설치 파일을 제공하는 사이트에서 다운받음

The screenshot shows the VMware website's download page for VMware Workstation 10.0.5 for Windows. The page has a header with the breadcrumb 'Home / VMware Workstation 10.0.5 for Windows'. The main heading is 'Download VMware Workstation 10.0.5 for Windows'. Below this, there's a 'Select Version' dropdown set to '10.0.5'. The 'Description' is 'VMware Workstation 10.0.5 for Windows'. The 'Documentation' link points to 'Release Notes'. The 'Release Date' is '2015-01-27'. The 'Type' is 'Product Binaries'. On the right, under 'Product Resources', there are links for 'View My Download History', 'Product Info', 'Documentation', 'Technical Papers', 'Knowledge Base', 'Community', 'System Requirements', and 'Receive Patch / Maintenance Alerts'. At the bottom right of this section is a 'Get Free Trial' button. Below the main content area, there are tabs for 'Product Downloads', 'Drivers & Tools', 'Open Source', and 'Custom ISOs'. The 'Product/Details' section shows 'Product installation including VMware Tools for all operating systems.', 'File size: 491 MB', 'File type: exe', and a 'Read More' link. A prominent blue 'Download Now' button is on the right.

##### 2 우분투 설치 파일을 제공하는 사이트에서 다운받음

The screenshot shows the Ubuntu website's download page for Ubuntu Desktop. The breadcrumb is 'Download > Overview > Cloud > Server > Desktop > Ubuntu Kylin > Alternative downloads'. The main heading is 'Download Ubuntu Desktop'. Below this, the version 'Ubuntu 14.04.2 LTS' is highlighted. A description states: 'The Long Term Support (LTS) version of the Ubuntu operating system for desktop PCs and laptops, Ubuntu 14.04.2 LTS comes with five years of security and maintenance updates, guaranteed.' It also says 'Recommended for most users.' and provides a link to 'Ubuntu 14.04.2 LTS release notes'. On the right, there's a 'Choose your flavour' dropdown set to '64-bit — recommended'. Below this is a large orange 'Download' button. At the bottom right, there's a link for 'Alternative downloads and torrents >'. The background is a dark purple gradient.

## ■ 하둡 설치

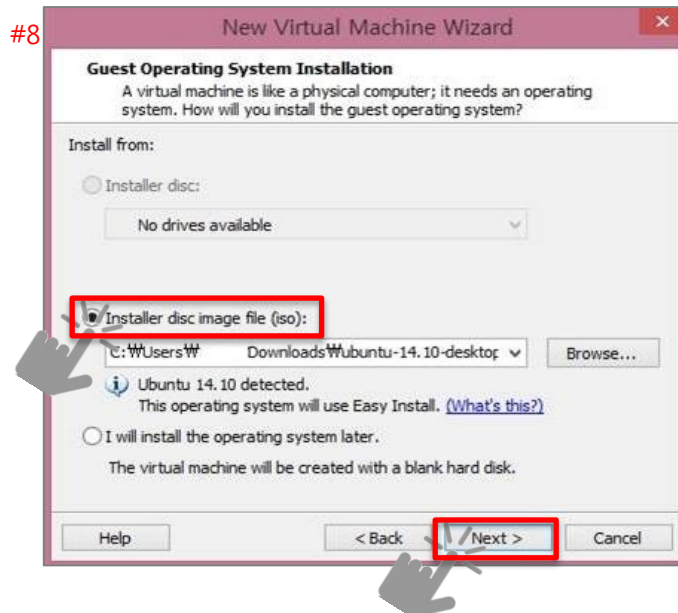
### 1. 하둡 구현을 위한 환경

#### VMware와 우분투 설치

- 3 VMware를 실행하여 Create A New Virtual Machine을 클릭함



- 4 미리 다운받은 우분투 .iso파일을 선택한 후 사용할 계정과 비밀번호를 설정하면 VMware 설치과정을 완료함

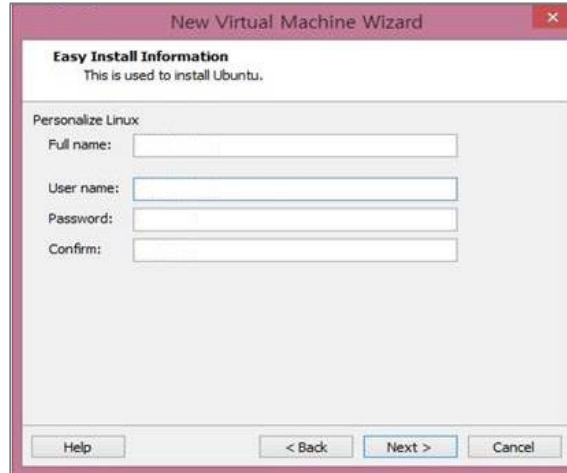


## ■ 하둡 설치

### 1. 하둡 구현을 위한 환경

#### VMware와 우분투 설치

- 4 미리 다운받은 우분투 .iso파일을 선택한 후 사용할 계정과 비밀번호를 설정하면 VMware 설치과정을 완료함



하둡을 설치하기 위한 가장 기본적인 환경 구축

## ■ 하둡 설치

### 2. 하둡 설치 모드별 설치방법

#### 하둡 설치 모드



#### 완전분산모드

- 모든 기능이 갖추어진 컴퓨터 클러스터를 구성할 수 있는 모드
- 설치를 통해 분산 저장과 분산 연산의 장점을 누릴 수 있음



단일컴퓨터로는 설치가 불가능함

#### 독립실행모드

- 다른 노드와 통신할 필요 없이 독립적으로 맵리듀스 프로그램의 로직을 개발하고 오류 수정 할 때의 모드



개발이나 오류 수정 목적으로 사용

#### 가상분산모드

- 컴퓨터 클러스터가 한 대로 구성
- 코드 오류 수정 시 독립실행모드에서의 기능을 보완
- 메모리 사용 정도, 하둡 분산파일시스템 입출력 관련 문제 등을 검사 가능함



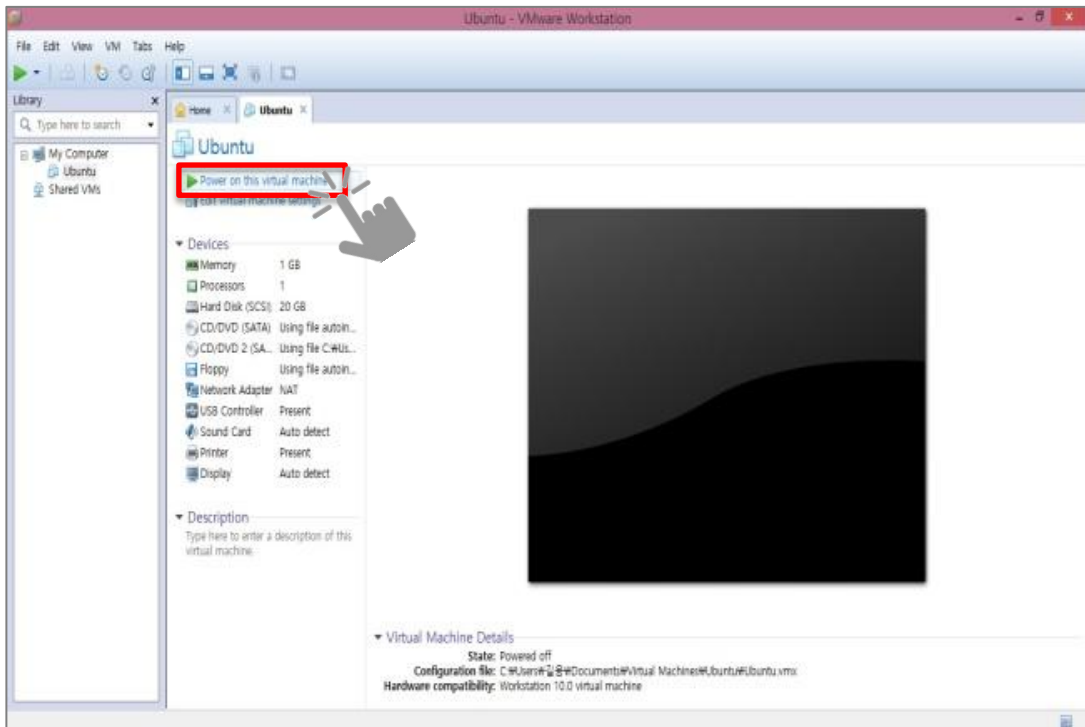
개발이나 오류 수정 목적으로 사용

## ■ 하둡 설치

### 2. 하둡 설치 모드별 설치방법

#### 독립실행모드 설치 과정

- 1 가상 머신인 Vmware을 활용하여 우분투를 실행시키고 접속함

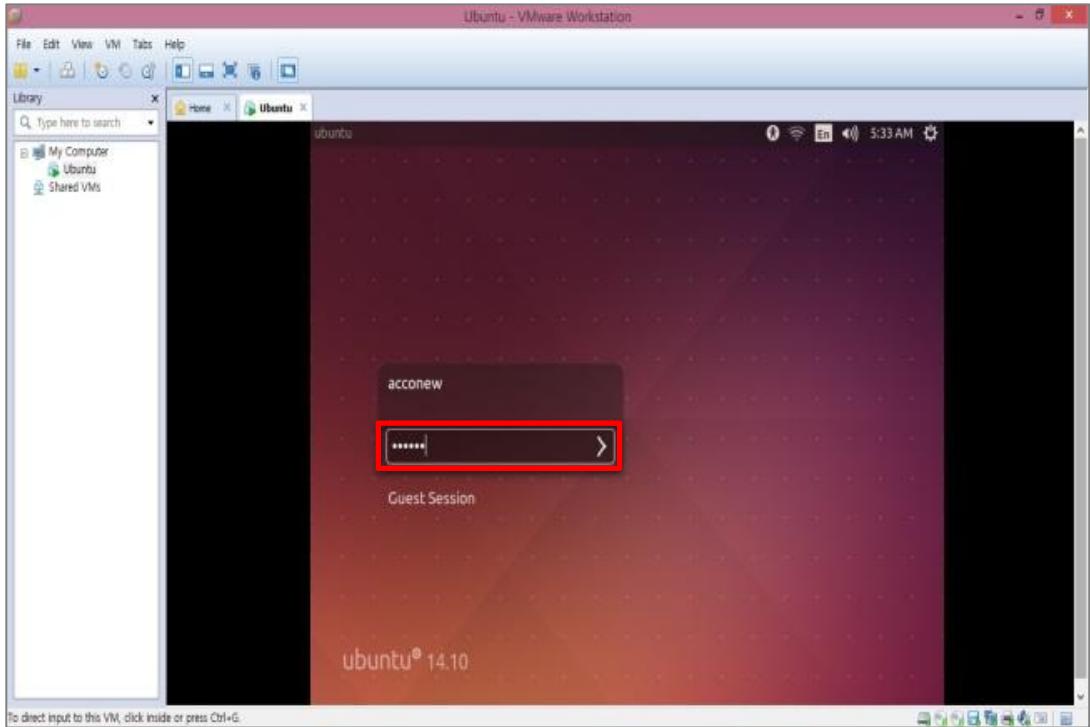


## ■ 하둡 설치

### 2. 하둡 설치 모드별 설치방법

#### 독립실행모드 설치 과정

2 가상머신인 Vmware을 활용하여 우분투를 실행시키고 접속함



#### 터미널(Terminal)

- 단말 장치(terminal unit)를 의미함
- 정보가 통신망에 들어가고 나가고 할 수 있는 지점
- 사용자가 데이터의 입출력을 행하는 가상의 장치
  - 키보드나 디스플레이 장치와 같은 기능을 갖고 통신망에 의해 정보를 송수신하는 장치

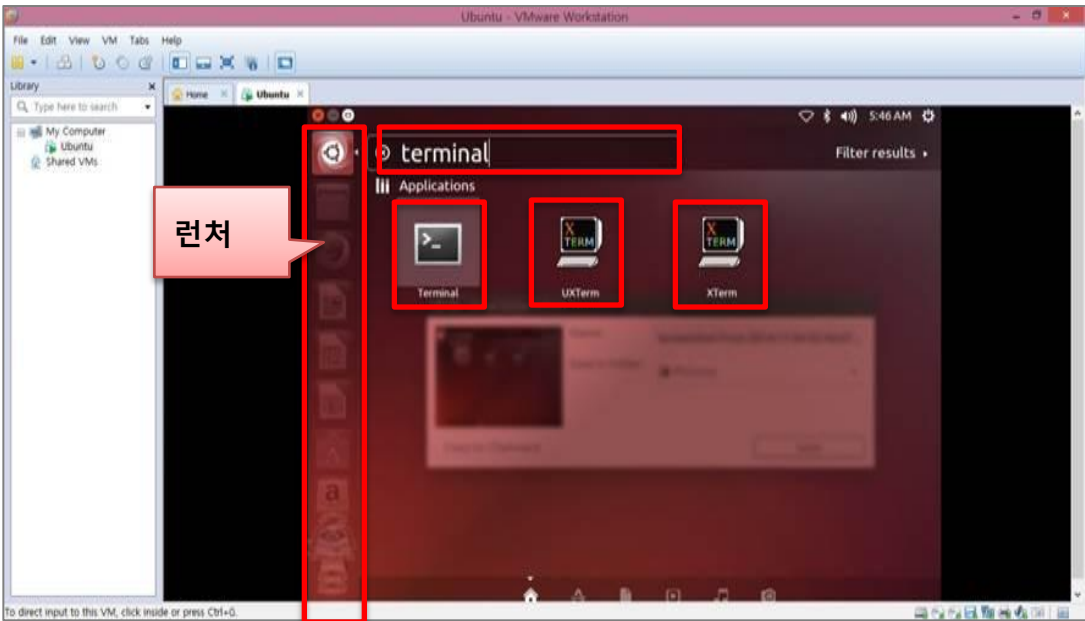


## ■ 하둡 설치

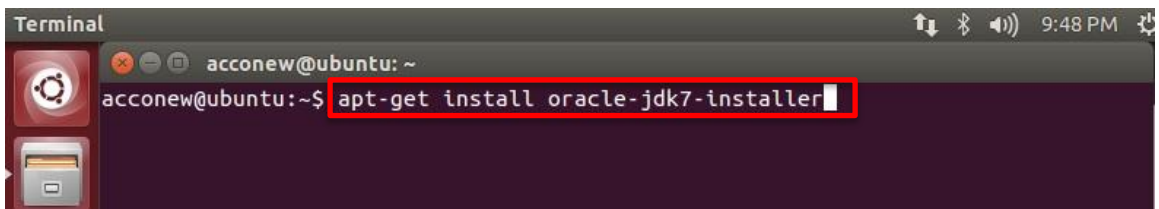
### 2. 하둡 설치 모드별 설치방법

#### 독립실행모드 설치 과정

#### 3 터미널(Terminal) 을 실행시킴



#### 4 실행된 터미널(Terminal)에서 다음 명령어를 입력해 우분투에 자바를 설치함

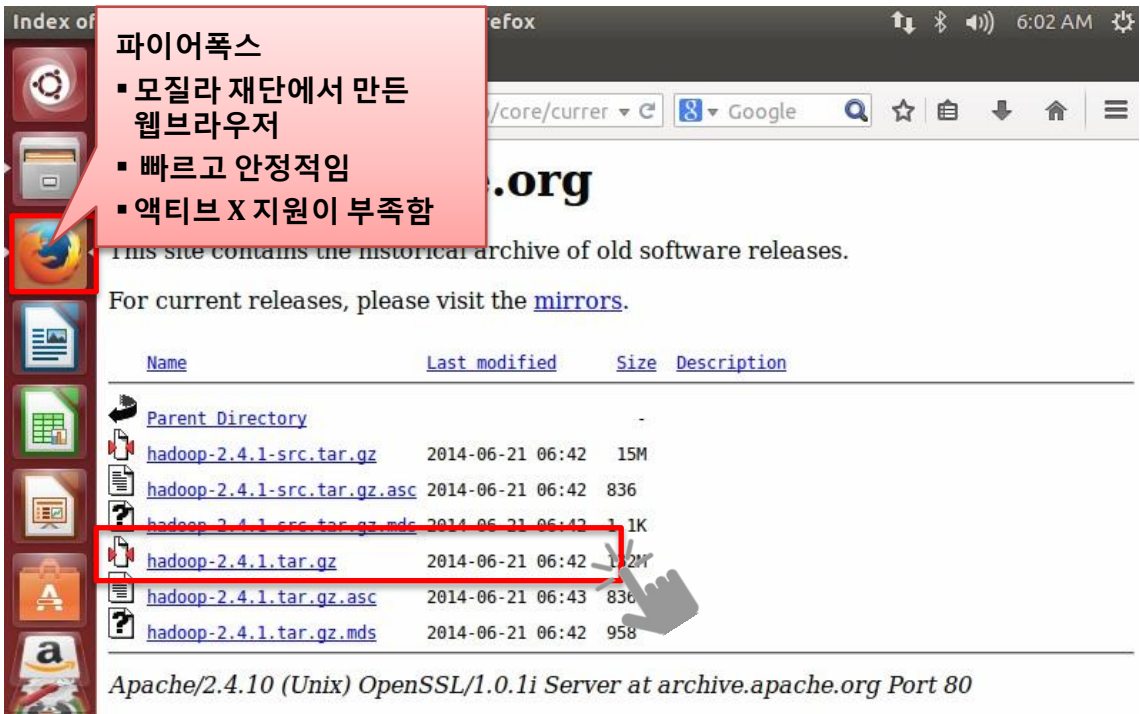


## ■ 하둡 설치

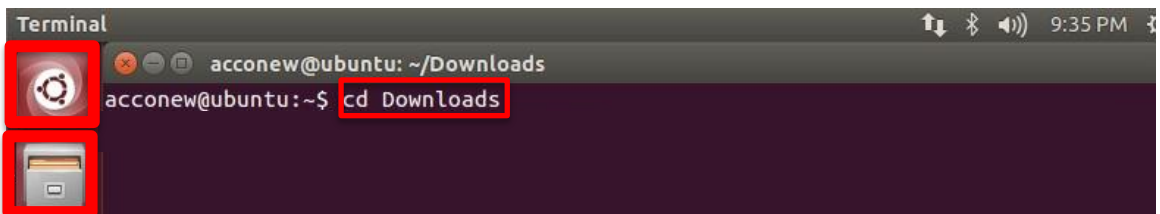
### 2. 하둡 설치 모드별 설치방법

#### 독립실행모드 설치 과정

5 우분투 내의 파이어폭스를 이용해 하둡을 설치함



6 터미널(Terminal)을 실행해주고 cd Downloads 명령어를 이용해 다운로드 된 폴더로 이동함

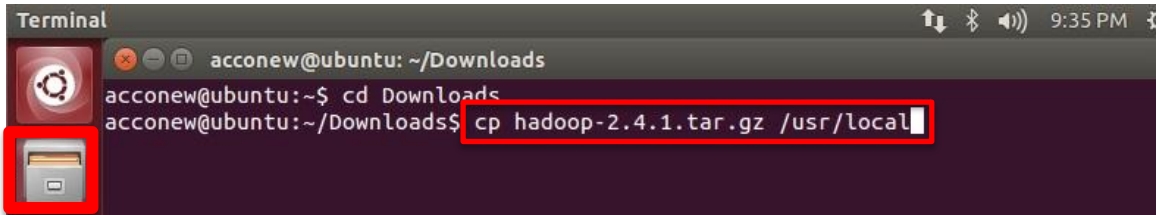


## ■ 하둡 설치

### 2. 하둡 설치 모드별 설치방법

#### 독립실행모드 설치 과정

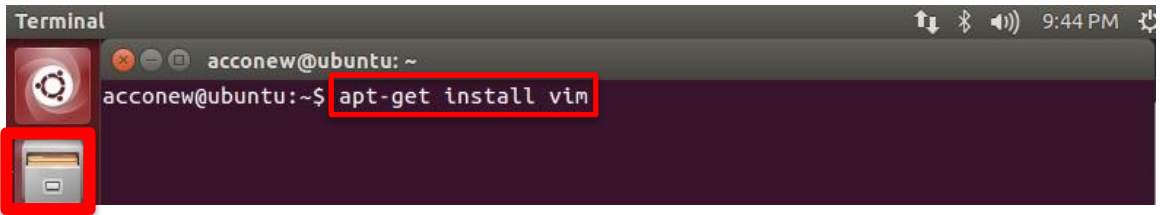
- 7 cp hadoop-2.4.1.tar.gz /usr/local 명령어를 이용해 하둡파일을 설정된 경로로 이동시킴



```
Terminal
acconew@ubuntu: ~/Downloads
acconew@ubuntu:~$ cd Downloads
acconew@ubuntu:~/Downloads$ cp hadoop-2.4.1.tar.gz /usr/local
```

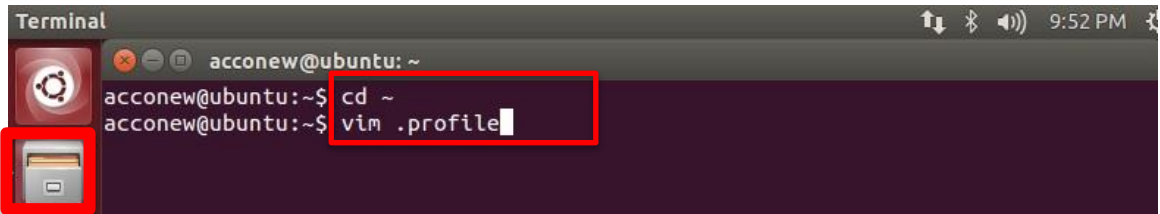
- 8 tar zxvf hadoop-2.4.1.tar.gz 명령어를 이용해 압축을 풀어 하둡을 설치함

- 9 경로를 설정해 주기 위해, 명령어를 입력해 vim을 설치함



```
Terminal
acconew@ubuntu: ~
acconew@ubuntu:~$ apt-get install vim
```

- 10 명령어를 이용해 profile을 엮



```
Terminal
acconew@ubuntu: ~
acconew@ubuntu:~$ cd ~
acconew@ubuntu:~$ vim .profile
```

## ■ 하둡 설치

### 2. 하둡 설치 모드별 설치방법

#### 독립실행모드 설치 과정

- 11 profile에 명령어를 입력한 후 저장함

```
export JAVA_HOME=/usr/lib/jvm/java-7-oracle
export HADOOP_INSTALL=/usr/local/hadoop-2.4.1
export PATH=$PATH:$HADOOP_INSTALL/bin
```

- 12 터미널(Terminal)을 이용해 source .profile 명령을 통해 profile을 등록함

- 13 터미널(Terminal)에서 hadoop을 입력하여 다음의 결과가 출력된다면 정상적으로 수행된 것임

```
Usage: hadoop [--config confdir] COMMAND
where COMMAND is one of:
  namenode -format      format the DFS filesystem
  secondarynamenode    run the DFS secondary namenode
  namenode              run the DFS namenode
  datanode              run a DFS datanode
  dfsadmin             run a DFS admin client
  mradmin              run a Map-Reduce admin client
  fsck                 run a DFS filesystem checking utility
  fs                   run a generic filesystem user client
  balancer              run a cluster balancing utility
  oiv                  apply the offline fsimage viewer to an fsimage
  fetchdt              fetch a delegation token from the NameNode
  jobtracker           run the MapReduce job Tracker node
  pipes                run a Pipes job
  tasktracker          run a MapReduce task Tracker node
  historyserver        run job history servers as a standalone daemon
  job                  manipulate MapReduce jobs
  queue                get information regarding JobQueues
  version              print the version
  jar <jar>            run a jar file
  distcp <srcurl> <desturl> copy file or directories recursively
  distcp2 <srcurl> <desturl> DistCp version 2
```

- 13 터미널(Terminal)에서 vim /usr/local/hadoop-2.4.1/conf/hadoop-env.sh 파일을 엮

- 14 제일 하단에 제시된 명령어를 추가 후 저장함

```
export JAVA_HOME=/usr/lib/jvm/java-7-oracle
export HADOOP_HOME=/usr/local/hadoop-2.4.1
```

## ■ 하둡 설치

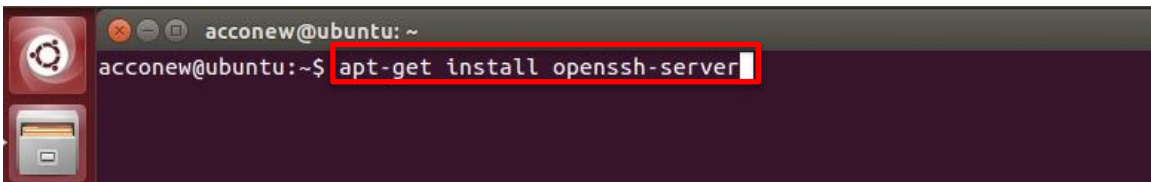
### 2. 하둡 설치 모드별 설치방법

#### 가상분산모드 설치 과정

기본적으로 독립실행모드 설치 과정이 완료되어 있어야 함  
독립실행모드 기능 보완이 가능함

#### 1 ssh를 설치함

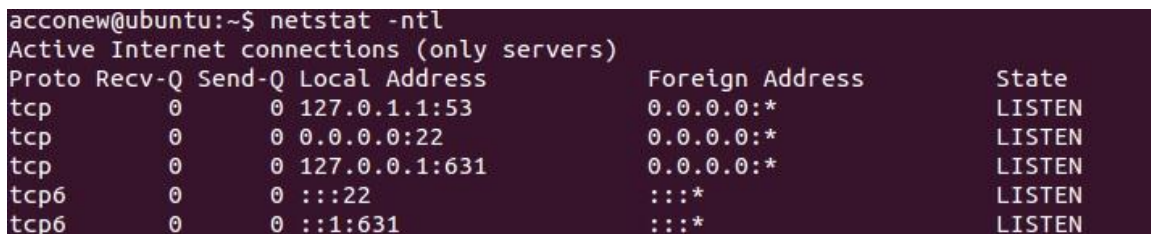
- ssh란 원격지 시스템에 접근하여 암호화된 메시지를 전송할 수 있는 프로그램임
- 하둡은 ssh프로토콜을 사용해 클러스터간에 내부통신을 하기 때문에 ssh서버를 설치해주어야 함
- 독립실행모드와 같이 터미널(Terminal)과 다음의 명령어를 이용해 ssh를 설치해줌



```
acconew@ubuntu: ~  
acconew@ubuntu:~$ apt-get install openssh-server
```

#### 2 /etc/init.d/ssh restart 명령어를 이용해 ssh를 실행시킴

#### 3 netstat -ntl명령어를 입력해서 아래와 같은 결과가 출력되면 성공적으로 설치된 것임



```
acconew@ubuntu:~$ netstat -ntl  
Active Internet connections (only servers)  
Proto Recv-Q Send-Q Local Address           Foreign Address         State  
tcp        0      0 127.0.1.1:53             0.0.0.0:*                LISTEN  
tcp        0      0 0.0.0.0:22               0.0.0.0:*                LISTEN  
tcp        0      0 127.0.0.1:631            0.0.0.0:*                LISTEN  
tcp6       0      0 :::22                   :::*                    LISTEN  
tcp6       0      0 :::1:631                 :::*                    LISTEN
```

## ■ 하둡 설치

### 2. 하둡 설치 모드별 설치방법

#### 가상분산모드 설치 과정

4 ssh는 키를 생성하고 생성키를 접속할 때 사용하도록 복사함

- rsa 공개키 암호를 사용하기 때문에 로그인할 때 별도의 키가 필요함
- 키를 생성하는 명령어

```
ssh-keygen -t rsa :키생성,
```

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys :  
생성키를 접속할 때 사용하도록 복사
```

5 ssh localhost를 입력하여 localhost로 접속함

```
acconew@ubuntu:~$ ssh localhost  
acconew@localhost's password:  
Welcome to Ubuntu 14.10 (GNU/Linux 3.16.0-23-generic i686)  
  
* Documentation:  https://help.ubuntu.com/  
  
Last login: Mon Nov  3 22:24:14 2014 from localhost  
acconew@ubuntu:~$
```

6 접속 후 xml설정 및 마스터 슬레이브 설정을 해 주면 설치가 끝남

실무에 적용 가능한 Big Data 분석 개론

---

# 하둡 활용



한국기술교육대학교  
온라인평생교육원

## ■ 하둡 분산파일시스템(HDFS)과 맵리듀스의 개념

### 1. 하둡 분산파일시스템(HDFS)의 개념

#### 하둡 분산파일시스템(HDFS)

빅데이터의 안정적인 저장 및 검색을 위해 설계된  
하둡의 분산파일시스템

#### HDFS의 클러스터 구성

하나의  
마스터

네임노드(NameNode)

여러 개의  
슬레이브

데이터노드  
(DataNode)

데이터노드  
(DataNode)

.....

데이터노드  
(DataNode)

- 대용량 파일이  
분산되어 여러  
데이터노드에  
저장



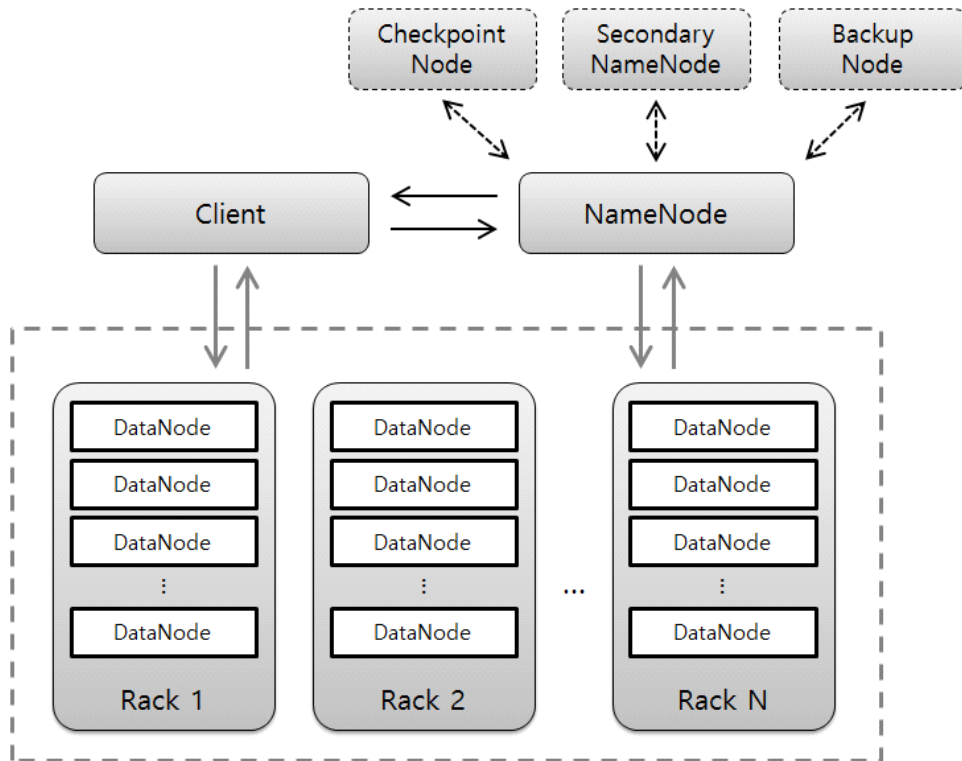
데이터의 안정성 보유  
고장에 대한 감내성 확보



## ■ 하둡 분산파일시스템(HDFS)과 맵리듀스의 개념

### 1. 하둡 분산파일시스템(HDFS)의 개념

#### 하둡 분산파일시스템(HDFS) 구조와 동작

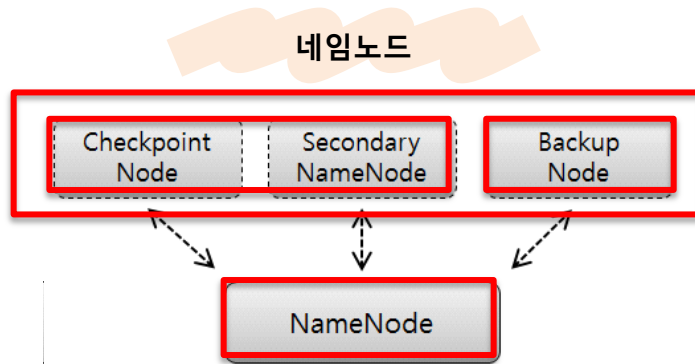


< 출처 : 김규식의 4명, 이기종 저장장치 기반의 HDFS 성능 분석,  
2014년 한국컴퓨터종합학술대회 논문집,2014 >

## ■ 하둡 분산파일시스템(HDFS)과 맵리듀스의 개념

### 1. 하둡 분산파일시스템(HDFS)의 개념

#### 하둡 분산파일시스템(HDFS) 구조와 동작



#### 네임노드(NameNode) 역할

- HDFS을 구성하는 파일, 디렉토리 등의 메타데이터를 메인 메모리에 유지함으로써 HDFS를 전반적으로 관리함

#### 체크포인트 노드와 세컨터리 노드, 백업 노드역할

- 네임노드에 문제가 생길 시를 대비하여 HDFS의 안정성을 높임
  - 체크포인트 노드와 세컨터리 노드 : 네임노드의 데이터를 자체 저장장치에 파일로 백업
  - 백업 노드 : 네임노드의 메타 데이터 정보를 자신의 메인 메모리에 백업
  - 세 노드는 하둡 버전마다 지원 여부가 상이하고, 역할이 일부 겹쳐서 동시에 사용할 수 없도록 제한되는 경우가 있으므로, 상황에 따라 탄력적으로 사용함

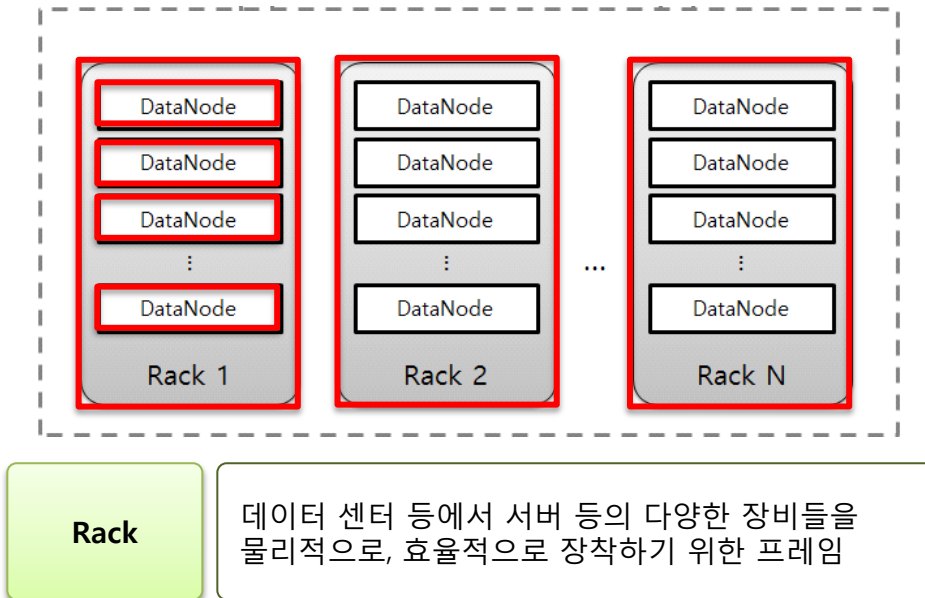
## ■ 하둡 분산파일시스템(HDFS)과 맵리듀스의 개념

### 1. 하둡 분산파일시스템(HDFS)의 개념

#### 하둡 분산파일시스템(HDFS) 구조와 동작

##### 데이터노드

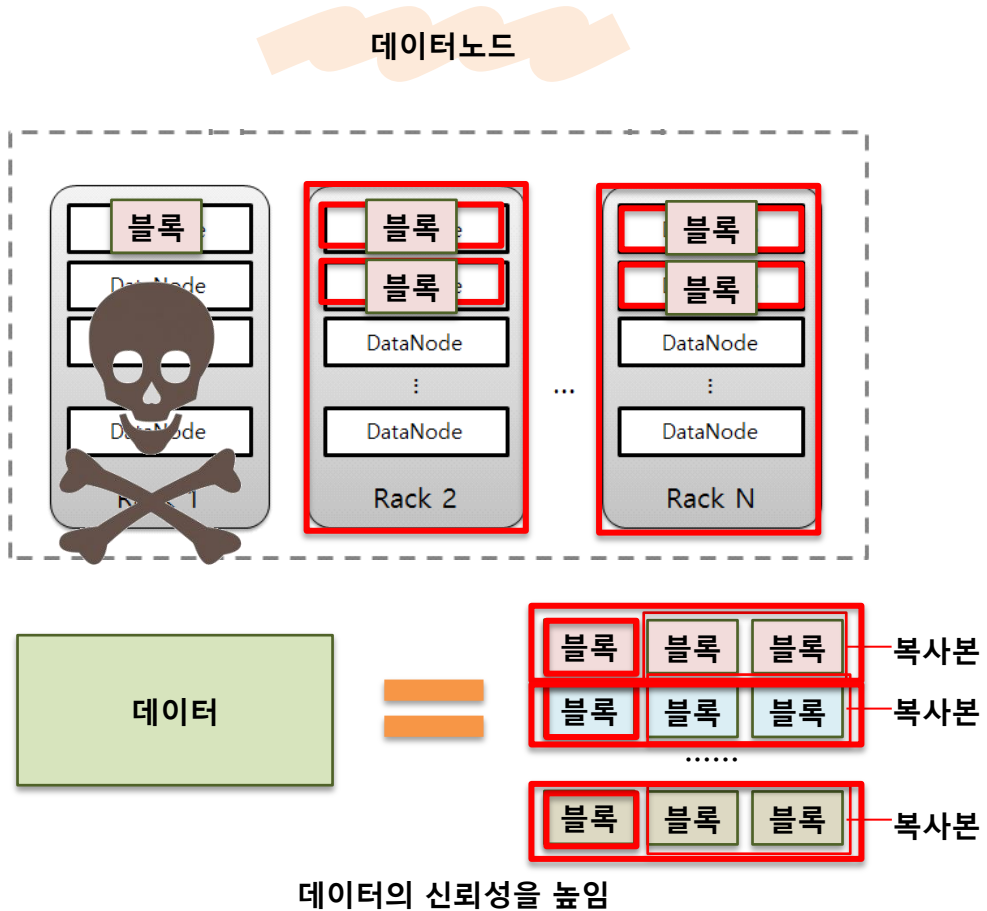
분산 처리할 정보가 실제로 저장되는 노드



## ■ 하둡 분산파일시스템(HDFS)과 맵리듀스의 개념

### 1. 하둡 분산파일시스템(HDFS)의 개념

#### 하둡 분산파일시스템(HDFS) 구조와 동작



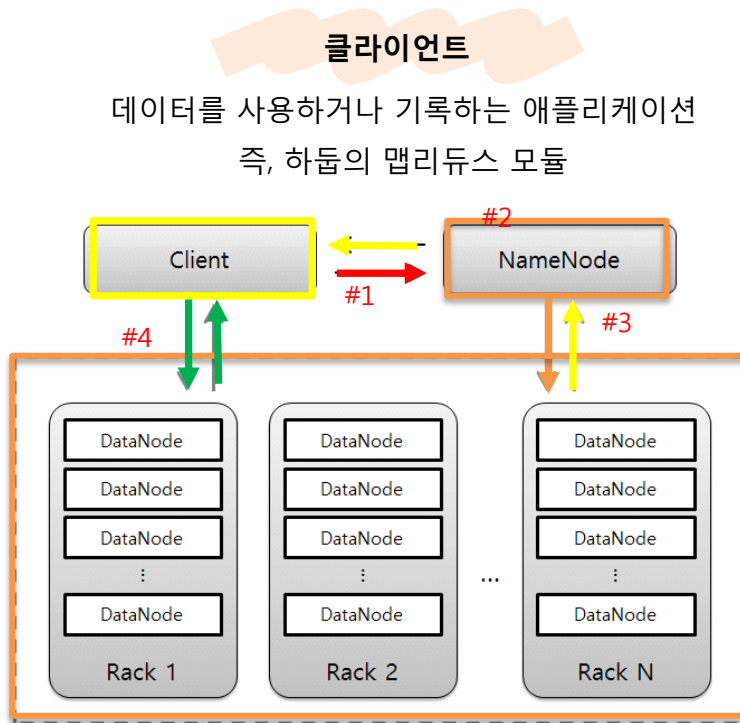
데이터의 신뢰성을 높임

- 하둡 분산파일시스템에서는 데이터를 저장할 때 기본적으로 블록 단위로 분할한 뒤 각 블록마다 두 개의 복사본을 만들어 총 세 개의 같은 블록을 유지함
- 복사본의 저장
  - 첫 번째 블록이 저장되지 않은 Rack을 찾아 그 Rack에 포함된 데이터노드 두 개를 선정하여 두 번째와 세 번째 블록을 저장함
  - 이를 통해 어떤 Rack 전체가 문제가 생기더라도 다른 Rack에서 해당 데이터를 구성하는 블록을 가져올 수 있게 되어 하둡 분산파일시스템에서의 데이터의 신뢰성을 높지게 됨

## ■ 하둡 분산파일시스템(HDFS)과 맵리듀스의 개념

### 1. 하둡 분산파일시스템(HDFS)의 개념

#### 하둡 분산파일시스템(HDFS) 구조와 동작



- 클라이언트는 하둡 분산파일시스템의 데이터를 사용하거나, 기록하는 애플리케이션으로 하둡의 맵리듀스 모듈이 이에 해당함
- 하둡 분산파일시스템에서 데이터를 읽거나 작성할 때 우선적으로 네임노드에 해당 요청을 보낸 뒤, 네임노드에서 반환해 준 데이터노드와 직접 통신하여 연산을 수행함

## ■ 하둡 분산파일시스템(HDFS)과 맵리듀스의 개념

### 1. 하둡 분산파일시스템(HDFS)의 개념

#### 하둡 분산파일시스템(HDFS) 설계

수십 PB에 이르는 대용량 데이터를 수천 대의 서버를 이용하여 빠르게 처리할 수 있도록 설계됨

##### 저가의 서버 이용

- 비용 문제로 **저가의 서버**를 이용하는 것을 전제로 하고 있음
- **고장 감내성**이 중요한 고려요소가 됨

##### 스토리지 용량 확장

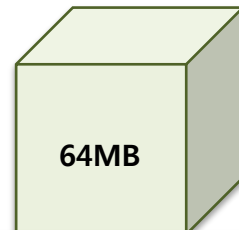
- 저가의 데이터노드를 **네임노드**에 추가 등록함
- 시스템의 중단 없이 용량을 확장시킴

데이터를 읽어오는 작업의 처리량을 높여 큰 데이터를 한 번에 빠르게 가져올 수 있도록 설계됨



수십 KB

일반적인 파일시스템 블록



64MB

하둡 분산파일시스템 블록

한번 쓰기 완료된 데이터는 **오직 덮어쓰기만 가능**하도록 설계됨

저장 방식이 간단해져 전체 시스템의 관리가 간편해짐

수천 대의 서버로 구성된 클러스터도 무리없이 운영 가능함

## ■ 하둡 분산파일시스템(HDFS)과 맵리듀스의 개념

### 2. 맵리듀스의 개념

#### 맵리듀스(MapReduce)

저렴한 머신들을 이용하여 빅 데이터를 병렬로  
분산 처리하기 위한 프로그래밍 모델



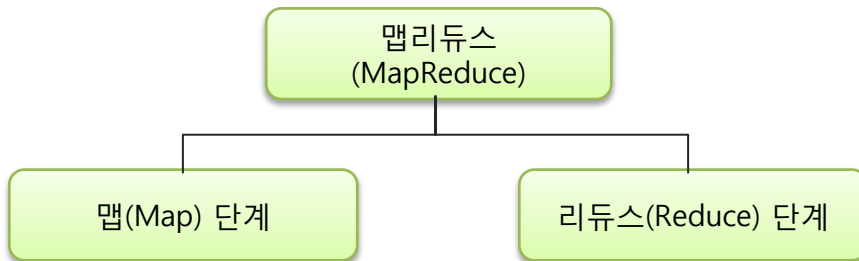
#### 개발동기

- 대용량 데이터를 신뢰할 수 없는 컴퓨터로 구성된 분산 클러스터 환경에서 대규모 데이터를 병렬로 처리하기 위함



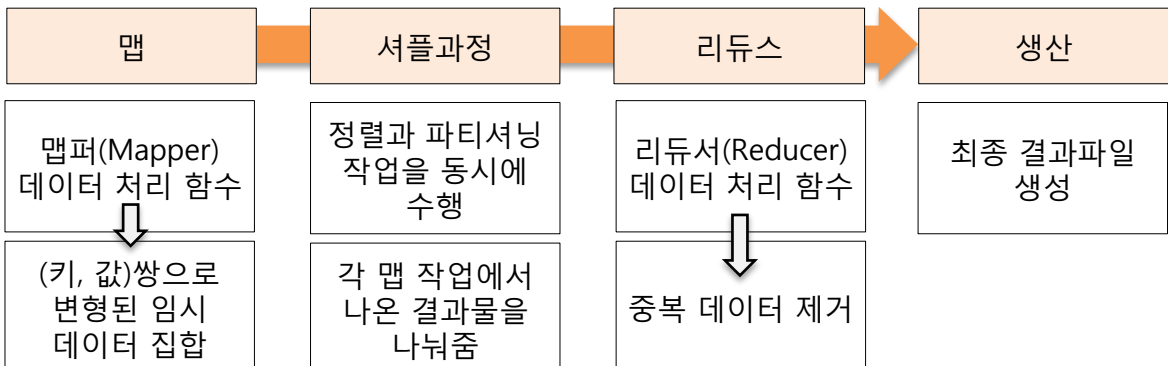
#### 혁신적인 부분

- 데이터 집합에 대한 쿼리를 입력 받아, 분할 한 후, 여러 개의 노드에서 병렬로 처리하는 것
- 단일 장비에서 처리하기에는 부적합한 대규모 데이터 처리 문제를 해결함



맵은 맵퍼라는 데이터 처리 함수를 통해 임시 데이터 집합으로 변형됨

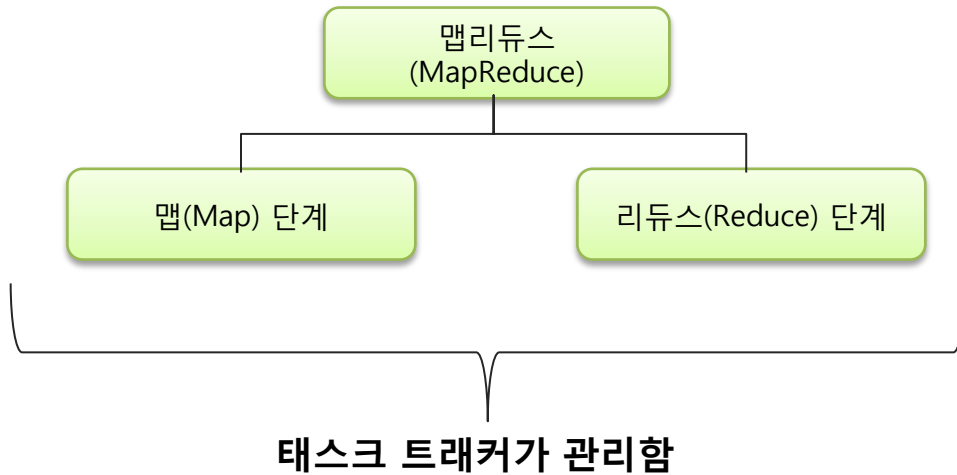
리듀스는 리듀서라는 데이터 처리 함수를 통해 원하는 데이터를 추출함



## ■ 하둡 분산파일시스템(HDFS)과 맵리듀스의 개념

### 2. 맵리듀스의 개념

#### 맵리듀스(MapReduce)



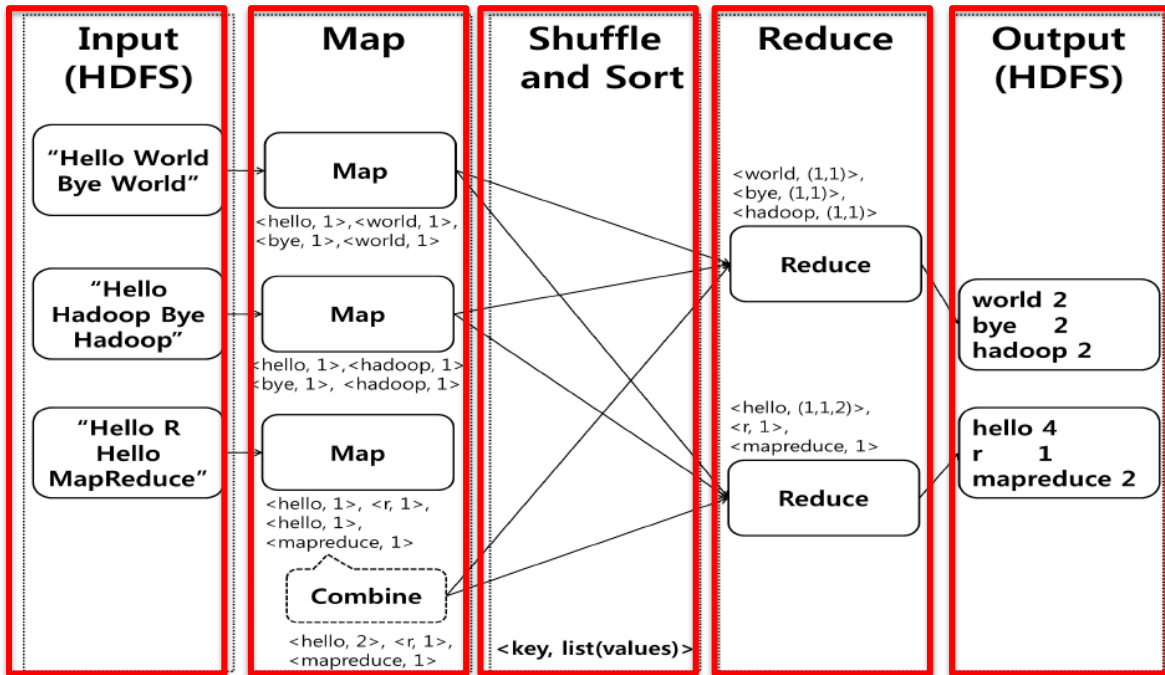
태스크 트래커	<ul style="list-style-type: none"><li>• 각각의 슬레이브 노드들에서 실행하는 태스크들을 관리함</li></ul>
잡 트래커	<ul style="list-style-type: none"><li>• 태스크들의 스케줄링을 담당함</li><li>• 하나의 마스터 노드 : 여러 슬레이브 노드들의 실행을 관리함</li></ul>



## ■ 하둡 분산파일시스템(HDFS)과 매프리듀스의 개념

### 2. 매프리듀스의 개념

#### 매프리듀스(MapReduce)



- 문서 집합에 등장하는 단어의 개수를 세는 작업
  - 3개의 문서가 하둡 분산파일시스템에서 입력으로 주어졌다고 가정하면, 3개의 맵 태스크 (task)가 발생함
  - 각 맵 태스크는 동일한 맵 함수를 실행하는데, 입력으로 문서의 ID가 키, 문서 텍스트가 값인 키나 값의 쌍이 주어지면, 이 맵 함수는 문서 텍스트를 단어 단위로 쪼개어 각 단어가 키가 되고 값은 1로 고정된 새로운 키나 값의 쌍인 집합을 출력함
  - 매프리듀스 시스템은 이 집합의 원소들에 대해 섞기 및 정렬과정을 통해 동일한 키를 가지는 값들을 하나로 묶음
  - 묶여진 <키, 값> 쌍들은 적절한 기준에 의해 2개의 리듀스 태스크로 분배되고, 각 리듀스 태스크는 역시 동일한 리듀스 함수를 실행하는데, 이 함수는 묶여진 값들을 모두 더하여 입력과 동일한 키의 새로운 값으로 만들
  - 그 결과가 하둡 분산파일시스템에 저장되며 이는 3개의 문서 집합에서 나타나는 단어의 빈도수가 되는데 여기서, 맵 태스크 및 리듀스 태스크의 갯수는 임의로 잡은 것이며, 실제로는 매프리듀스 시스템에서 자동적으로 생성함