

Hadoop 개요

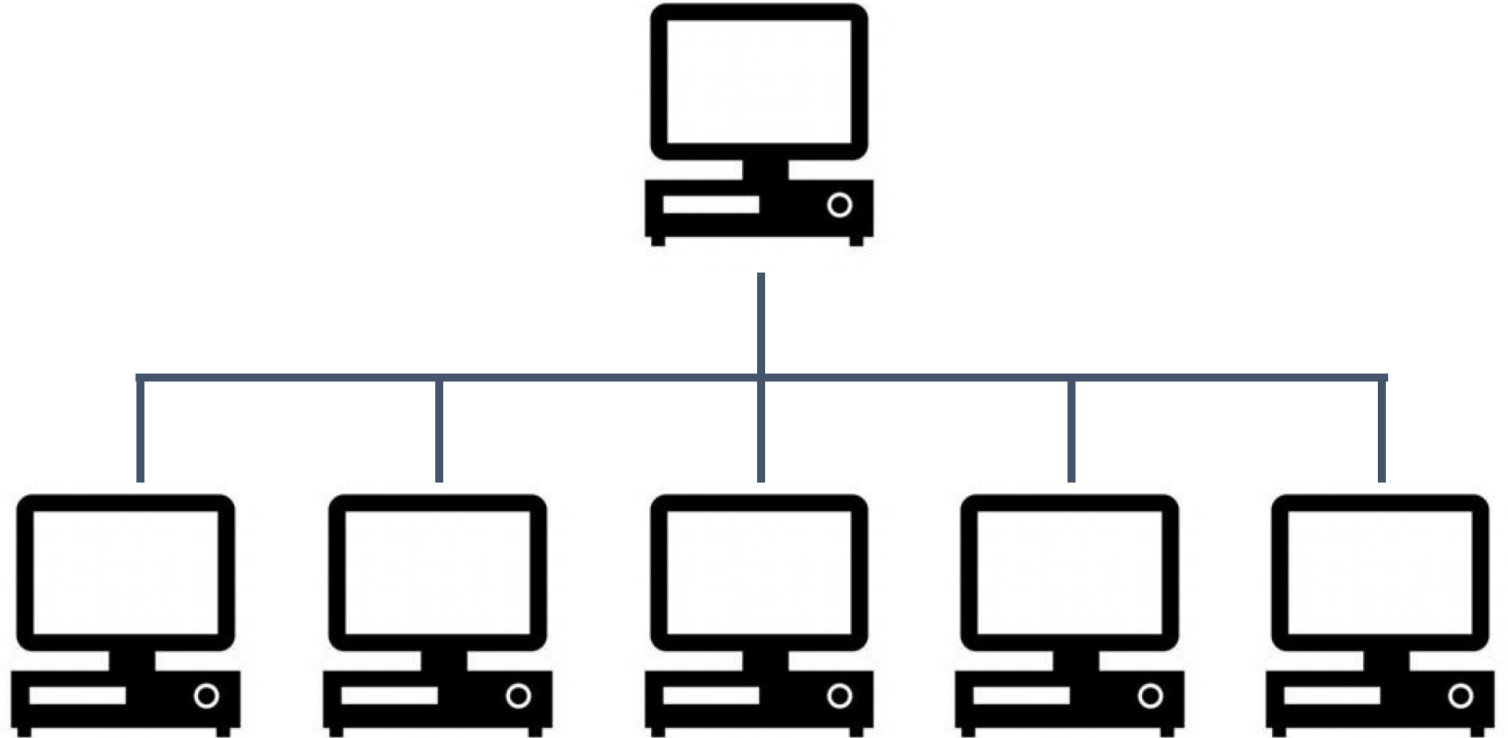
Local vs Distributed

- 로컬 컴퓨터에서는 RAM 용량에 따라 0-32GB 범위의 데이터를 다룰 수 있습니다.
- 더 큰 데이터셋은 SQL 데이터베이스를 사용하여 스토리지를 사용하거나, 여러대의 컴퓨터로 구성된 분산시스템을 사용해야 합니다.

Local



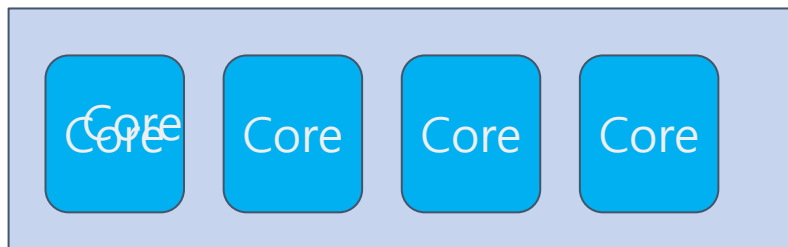
Distributed



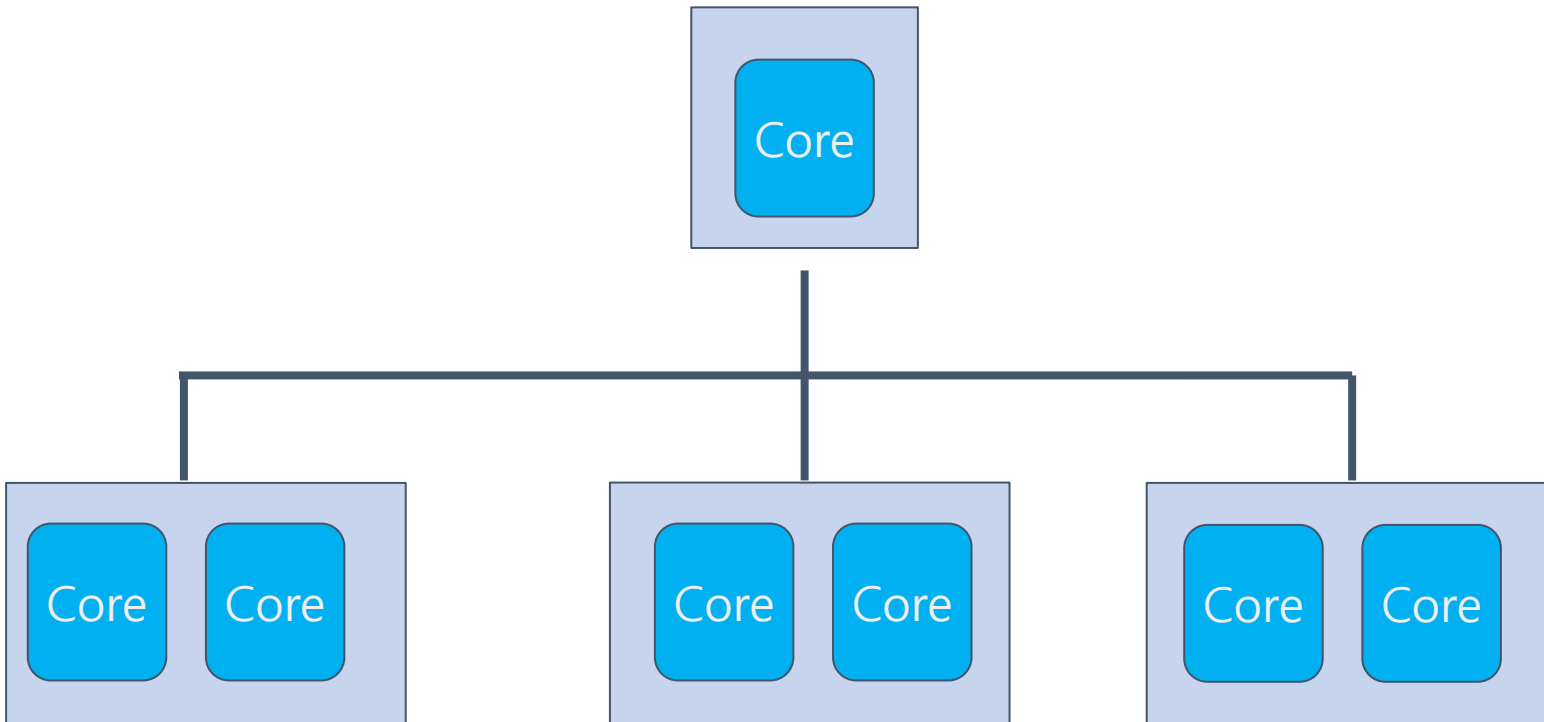
Local vs Distributed

- 로컬 프로세스는 단일 시스템의 컴퓨팅 리소스를 사용합니다.
- 분산 프로세스는 네트워크를 통해 연결된 여러 머신의 컴퓨팅 리소스를 액세스 할 수 있습니다.
- 단일 머신을 Scale Up 하는 것보다 여러대의 머신으로 확장(Scale Out)하는 것이 더 쉽습니다.
- 분산시스템에서는 한 대의 시스템에 장애가 발생해도 전체 네트워크가 계속 작동 할 수 있습니다.

Local



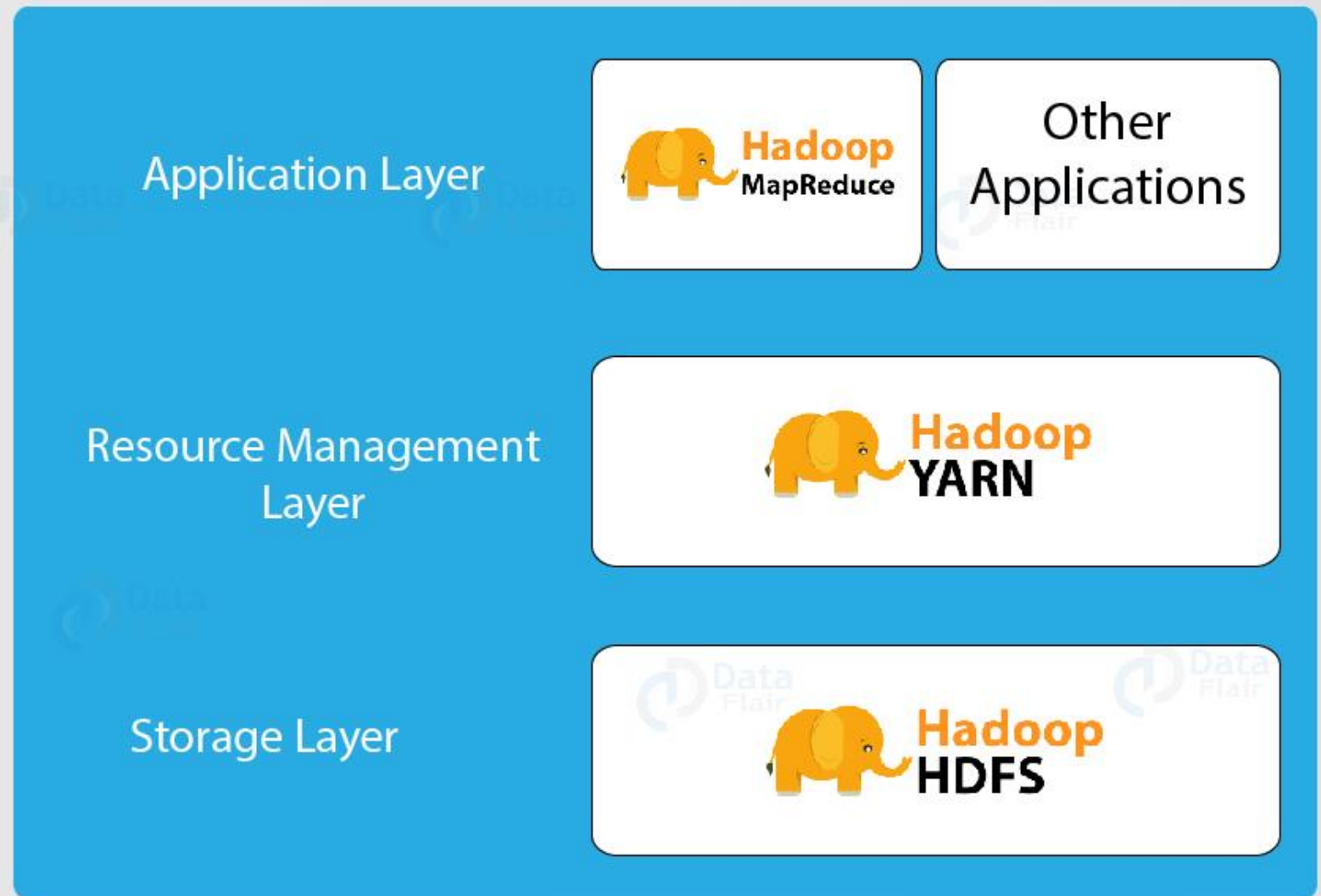
Distributed



Hadoop

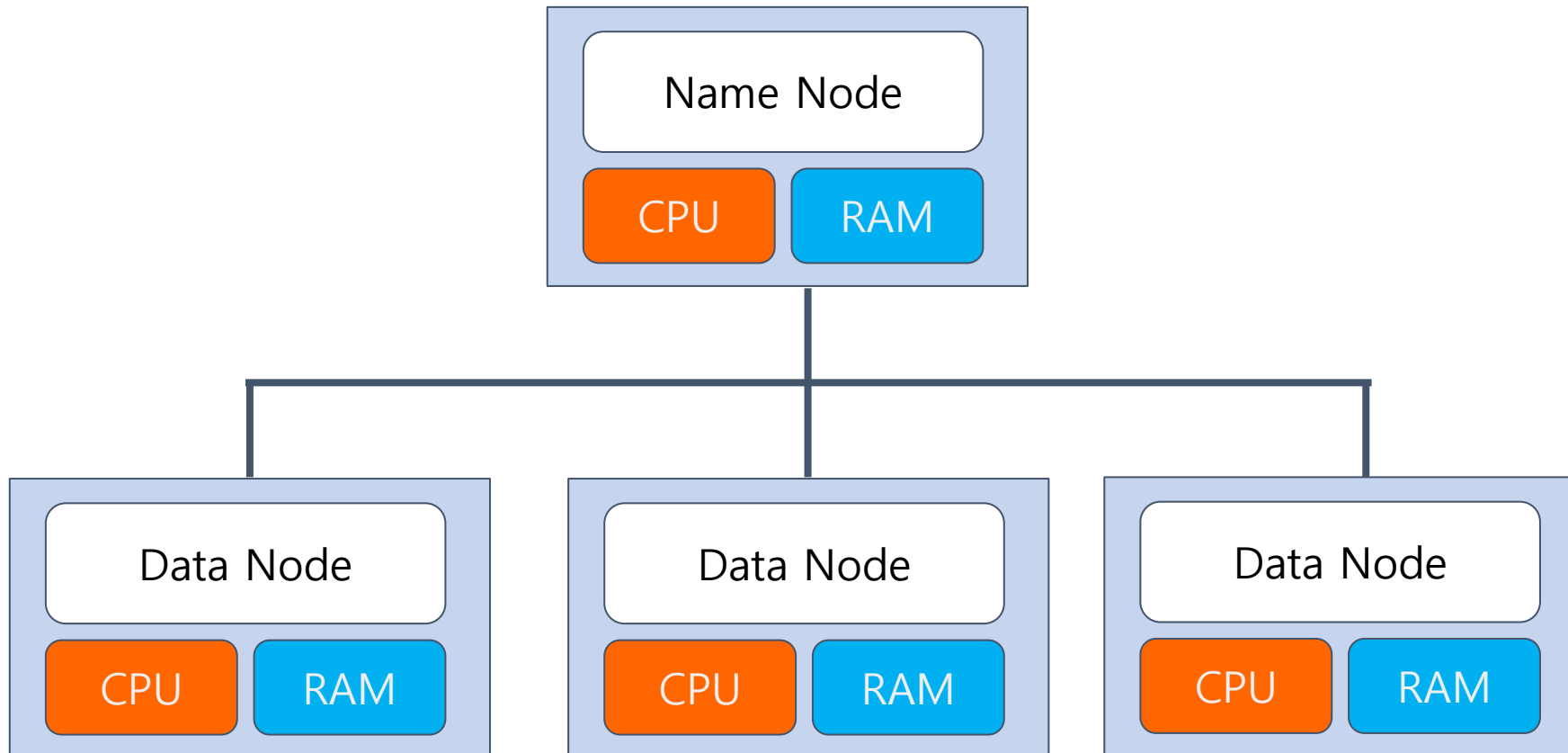


Hadoop Architecture



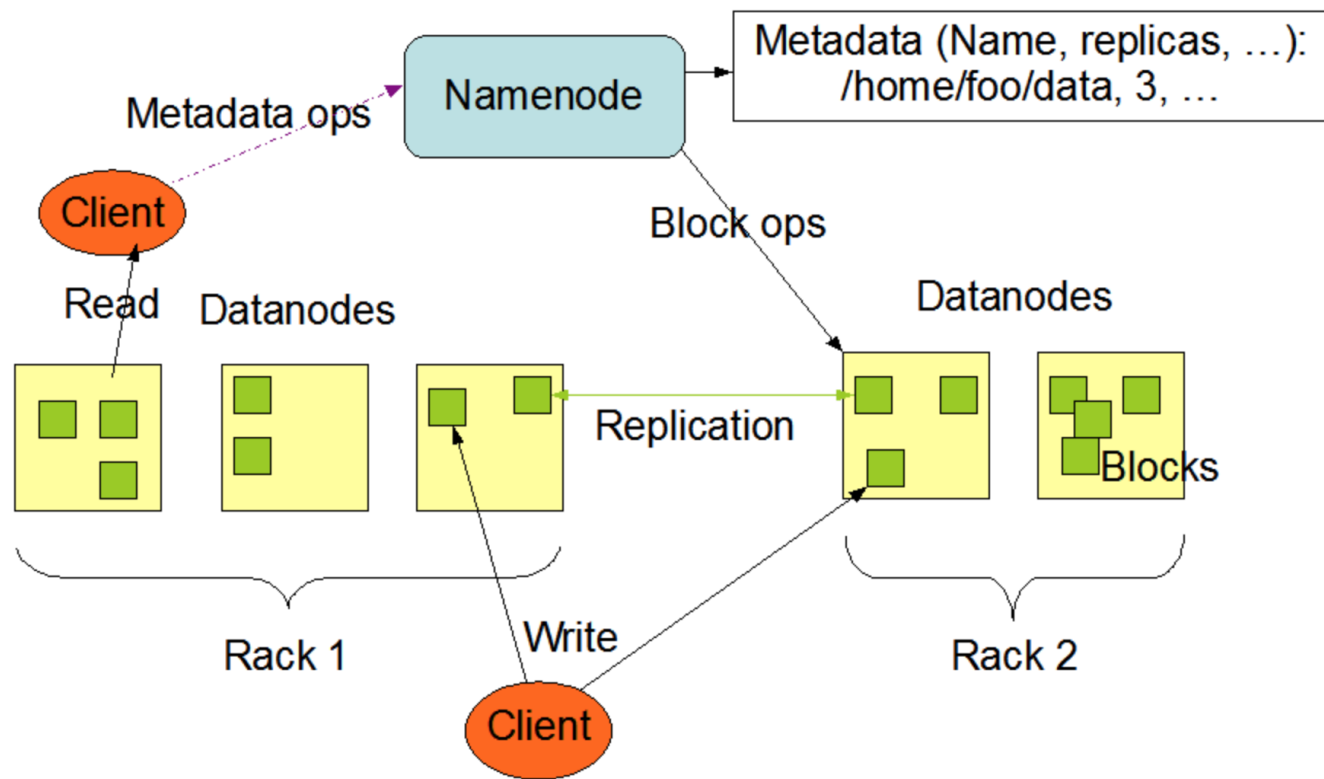
HDFS (Hadoop Distributed File System)

- HDFS는 기본적으로 128MB 크기의 데이터 블록(block)을 사용하며, 각 블록은 최소 3개로 복제됩니다.
- 블록은 Fault Tolerance을 지원하는 방식으로 분산(distributed) 됩니다.
- 더 작은 블록은 처리 중에 더 많은 병렬화를 제공합니다.
- 블록의 여러 복사본이 노드 장애로 인한 데이터 손실을 방지합니다.



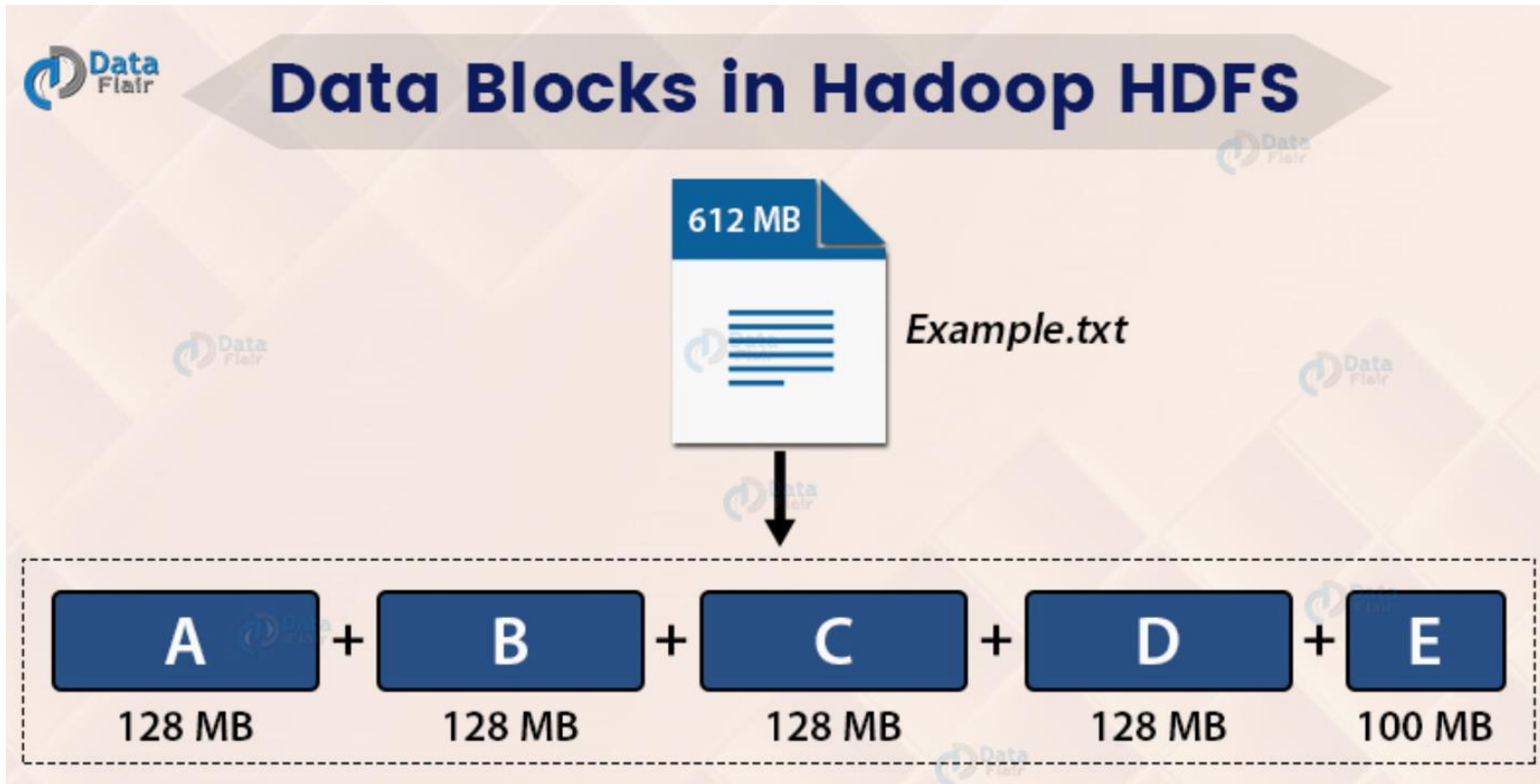
HDFS - 아키텍처

- HDFS는 클러스터에 파일을 분산저장/관리하는 역할을 하며, Namenode, Datanode, Client 모듈로 구성
- Namenode는 전체적인 HDFS가 어떻게 구성되어 있는지, 어디에 어떤 블록이 저장되어 있는지 등 HDFS의 메타 데이터를 관리하는 역할을 합니다.
- Datanode는 실제적으로 파일을 저장하는 역할을 합니다.
- Client는 사용자가 작성한 프로그램으로 이 HDFS에 파일을 쓰거나 읽는 작업을 요청합니다.



HDFS - 파일 저장 방식

- HDFS는 파일을 분산 저장하기 위해서 먼저 파일의 메타 데이터와 콘텐츠 데이터를 분리합니다.
- 메타 데이터 : 파일의 접근 권한, 생성일, 수정일, 네임 스페이스 등 파일에 대해 설명하는 정보
- 콘텐츠 데이터: 실제 파일에 저장된 데이터
- 파일의 메타데이터는 네임 노드에 저장되며 콘텐츠 데이터는 블록 단위로 쪼개져서 데이터 노드에
- 블록의 기본 크기는 128MB이며, 최소 3개의 복사본을 생성하여 분산 저장합니다.



HDFS - NameNode

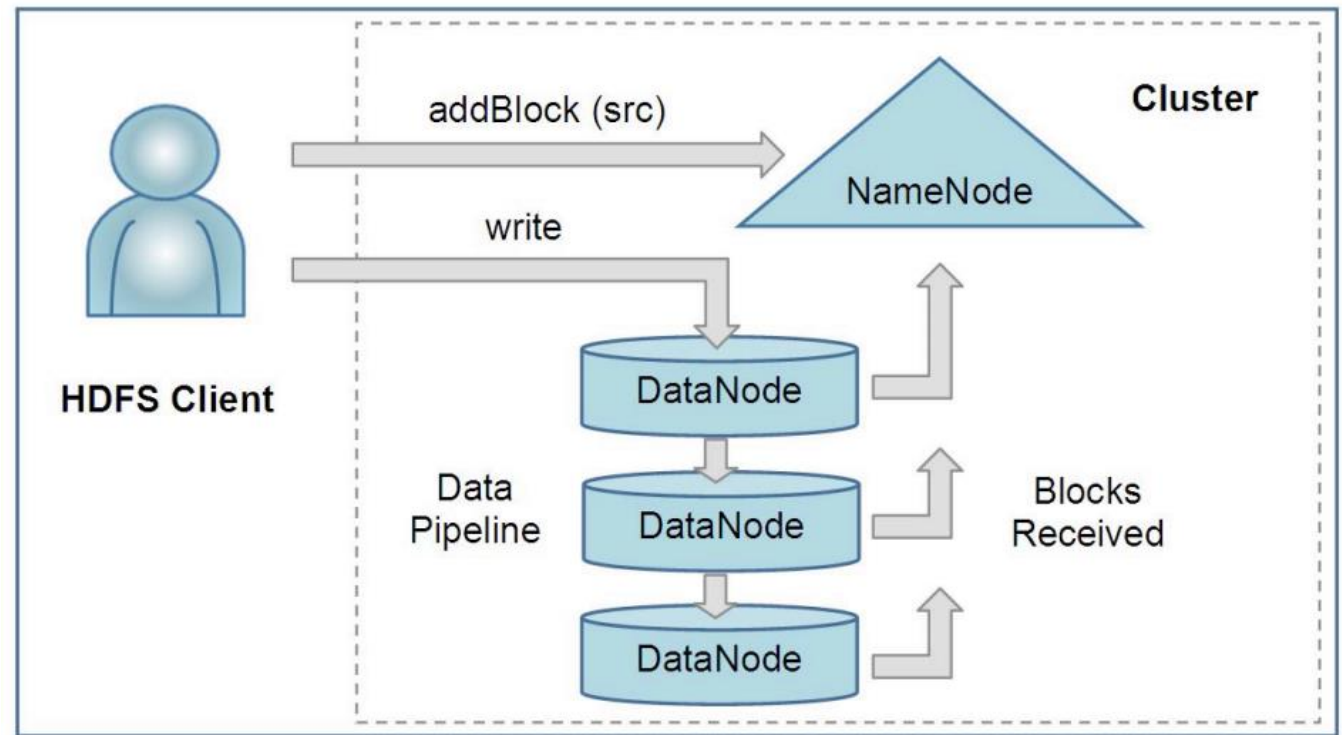
- 네임 노드는 파일의 메타 데이터를 inode에 저장합니다.
- inode란 unix 파일 시스템에서 사용되는 폴더 구조를 저장하는 구조체를 말합니다.
- 또한 네임 노드에는 파일 구성하는 블록들의 목록과 위치 정보가 저장되어 있습니다.
- 이러한 네임 노드는 HDFS에 파일을 읽거나 쓰는 작업의 시작점 역할을 수행합니다.

■ 파일 읽기 작업

1. 클라이언트는 네임 노드에 파일의 네임 스페이스에 해당하는 블록들의 목록과 주소를 요청
2. 클라이언트는 가장 가까이 위치한 데이터 노드에서 블록을 읽어옴

■ 파일 쓰기 작업

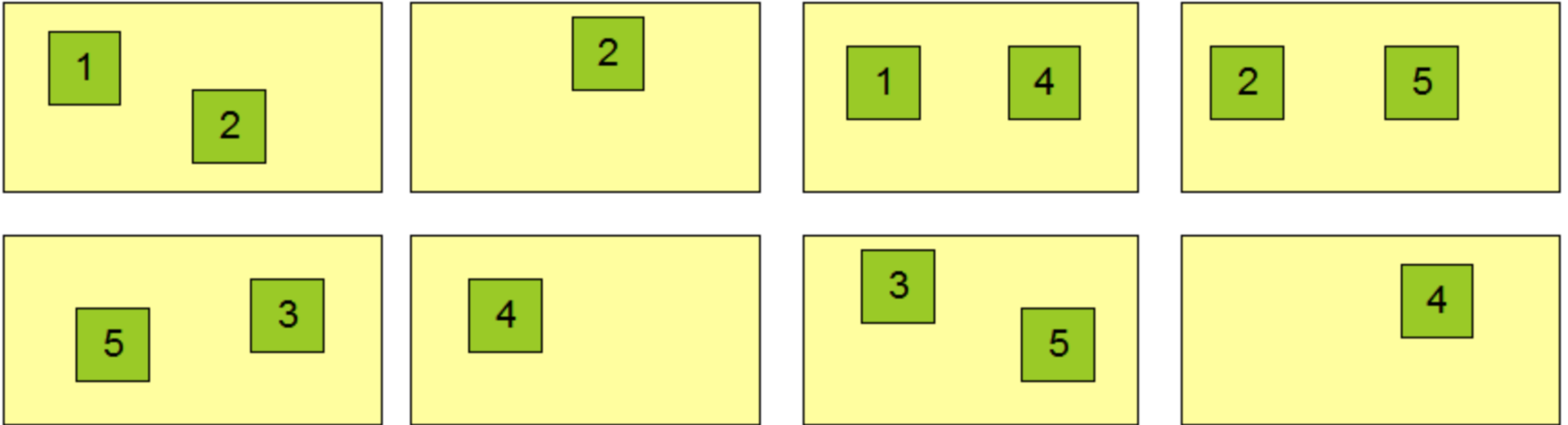
1. 클라이언트는 네임 노드에게 어느 데이터 노드에 블록을 쓰면 좋을지 요청
2. 네임 노드가 데이터 노드를 할당
3. 클라이언트는 블록을 쓰기 위한 데이터 파이프라인 생성
4. 데이터 파이프라인을 이용해서 블록 쓰기 작업 수행



HDFS - Datanode

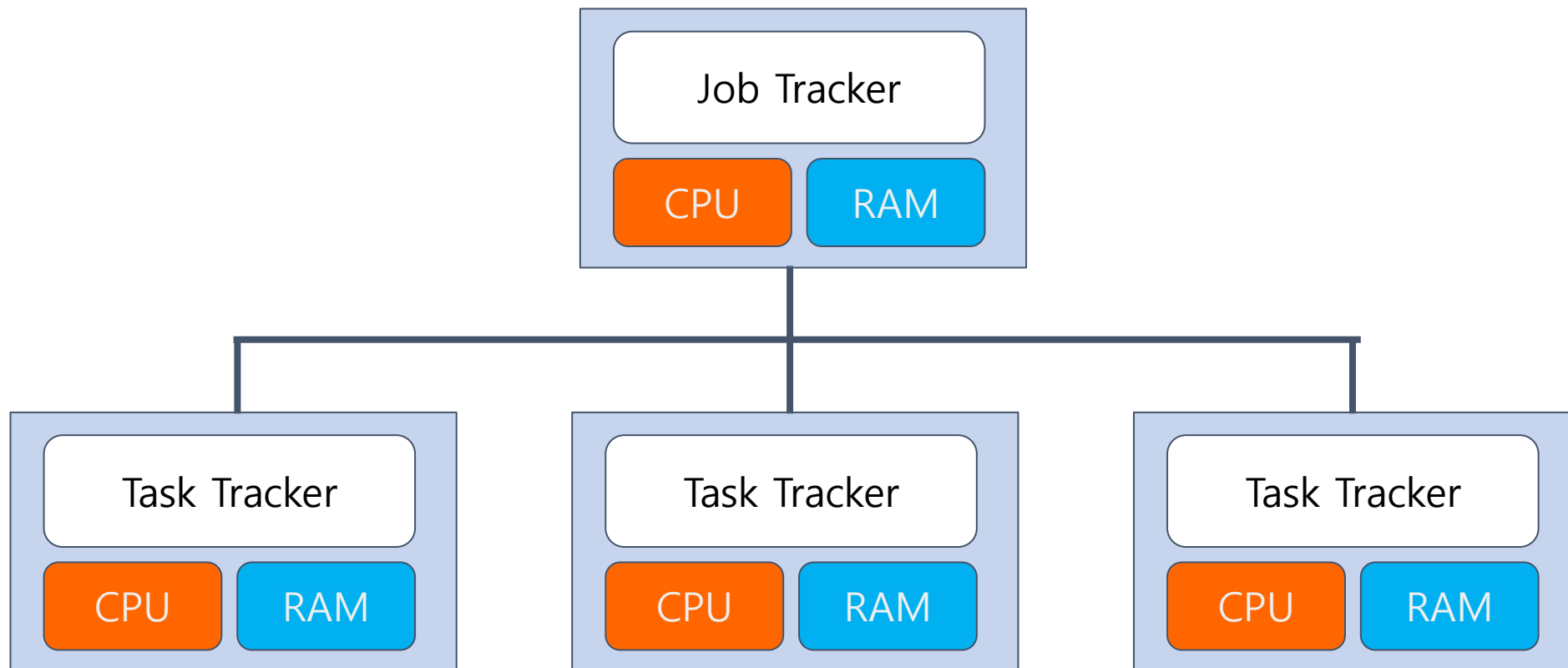
- 파일의 콘텐츠 데이터는 블록 단위로 나뉘며, 블록의 기본 크기는 128MB이며 최소 3개의 복사본을 생성하여 분산 저장합니다.
- 데이터 노드는 그 중 하나의 복사본을 저장하는 것입니다.

Datanodes



MapReduce

- MapReduce는 Computation Task를 분산 된 파일셋(예 : HDFS)으로 분할하는 방법입니다..
- Job Tracker와 여러 Task Tracker로 구성됩니다.
- Job Tracker는 Task Tracker에서 실행할 코드를 보냅니다.
- Job Tracker 는 Task에 CPU와 메모리를 할당하고 Worker 노드의 Task 들을 모니터링 합니다.



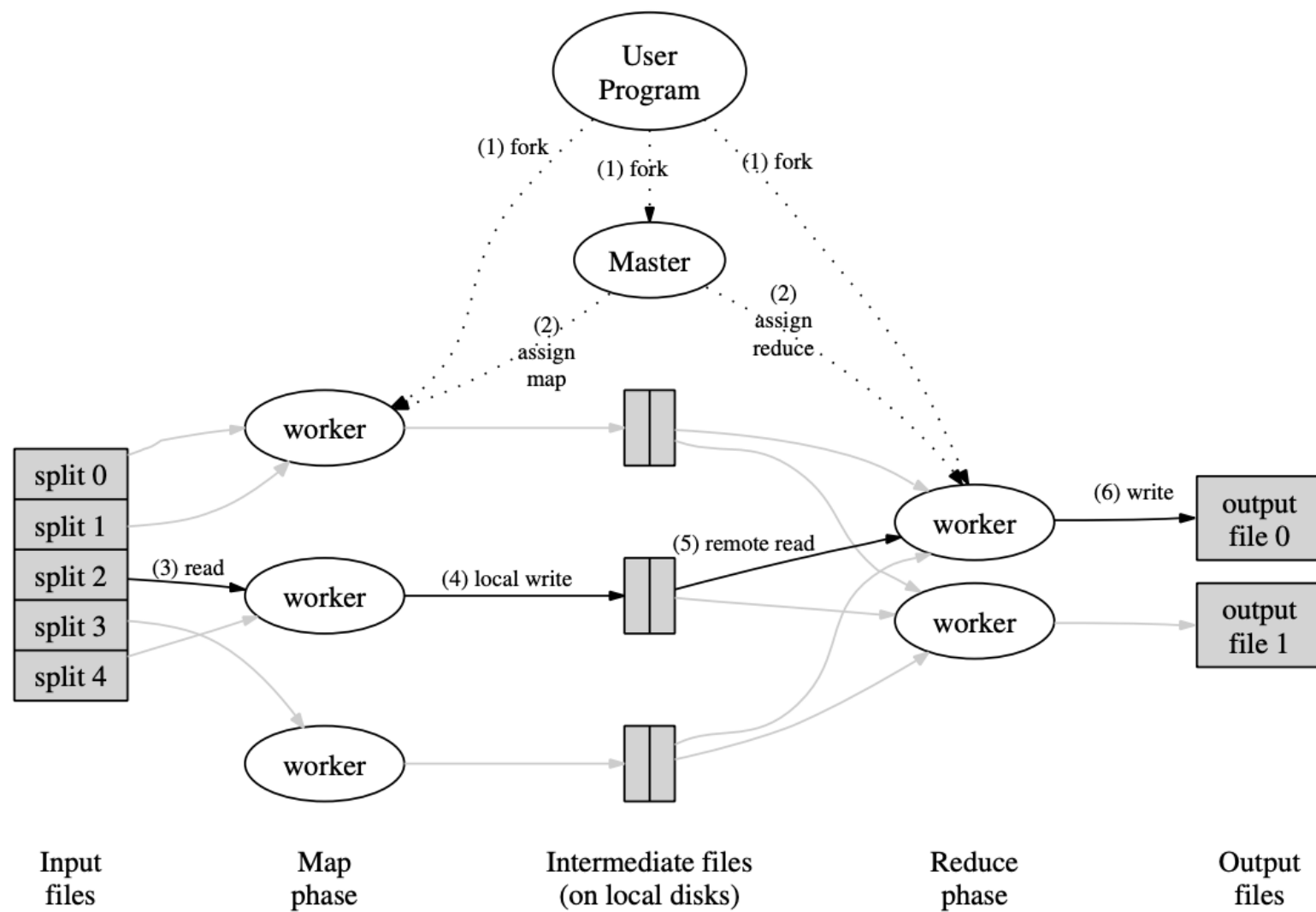
MapReduce

- 맵 리듀스는 구글 내부에서 크롤링 된 문서, 로그 등 방대한 양의 raw data를 분석하는 과정에서 느낀 불편함에서 출발했습니다.
- 프로그램 로직 자체는 단순한데 입력 데이터의 크기가 워낙 커서 연산을 하나의 물리 머신에서 수행할 수가 없었습니다.
- 이 거대한 인풋 데이터를 쪼개어 수많은 머신들에게 분산시켜서 로직을 수행한 다음 결과를 하나로 합치는 것이 핵심 아이디어 입니다.
- MapReduce 프레임워크에서 개발자가 코드를 작성하는 부분은 map과 reduce 두 가지 함수입니다.
- map은 전체 데이터를 쪼갠 청크에 대해서 실제로 수행할 로직입니다.
- reduce는 분산되어 처리된 결과 값들을 다시 하나로 합쳐주는 과정이며, 이 역시 분산된 머신들에서 병렬적으로 수행됩니다.

■ MapReduce 수행 절차

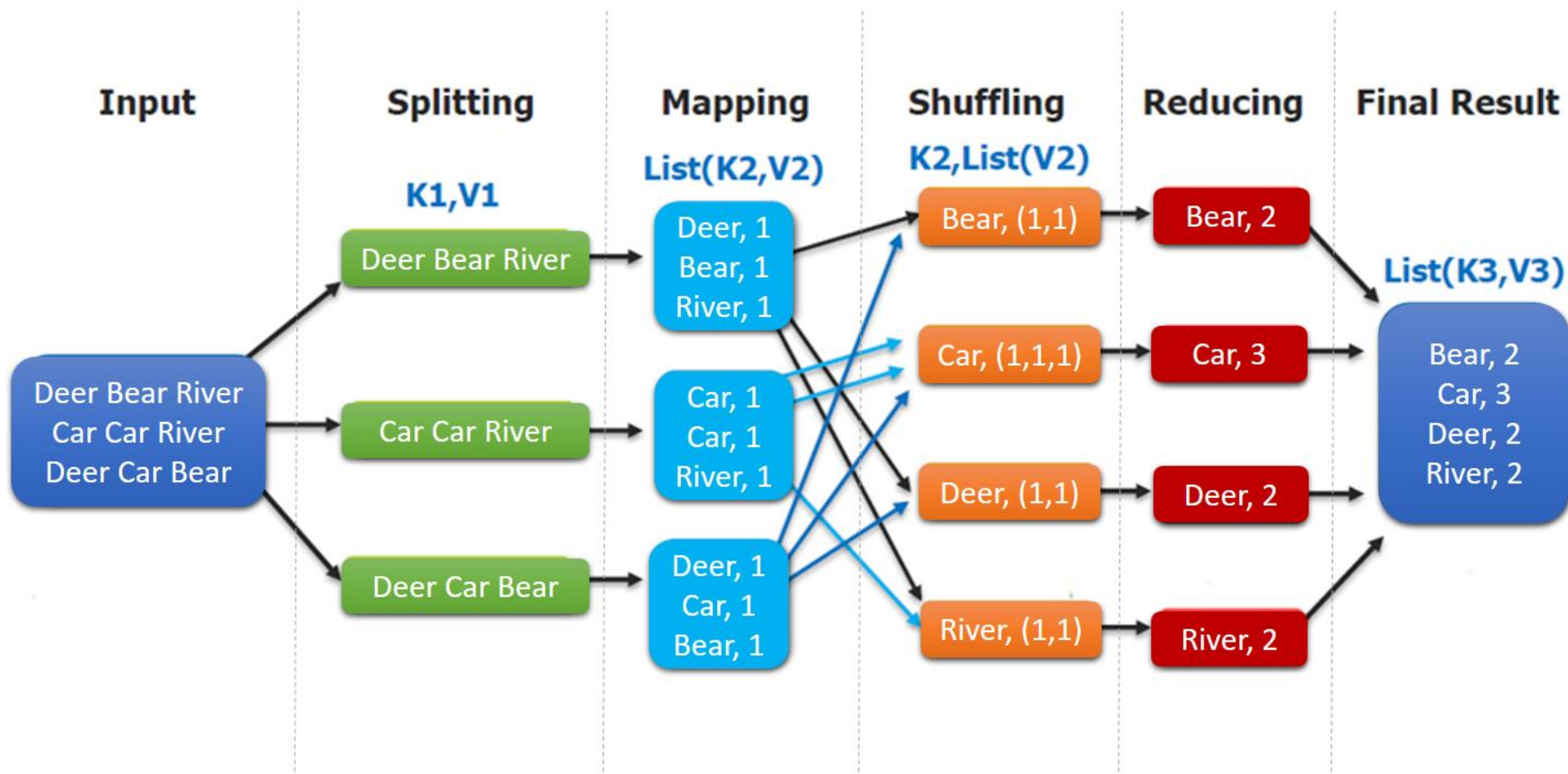
1. 쪼개기(Split): 크기가 큰 인풋 파일을 작은 단위의 청크들로 나누어 분산 파일 시스템(ex. HDFS)에 저장합니다.
2. 데이터 처리하기(Map): 잘게 쪼개어진 파일을 인풋으로 받아서 데이터를 분석하는 로직을 수행합니다.
3. 처리된 데이터 합치기(Reduce): 처리된 데이터를 다시 합칩니다.

MapReduce



MapReduce

3대의 Mapper와 4대의 Reducer 노드로 이루어진 클러스터에서 워드 카운팅을 수행하는 예시



MapReduce Word Count Process

Hadoop 설치

VirtualBox 설치

<https://www.virtualbox.org/>



■ 설치방법 참고 : <https://bit.ly/2T2rR7L>

VirtualBox 6.1.20 platform packages

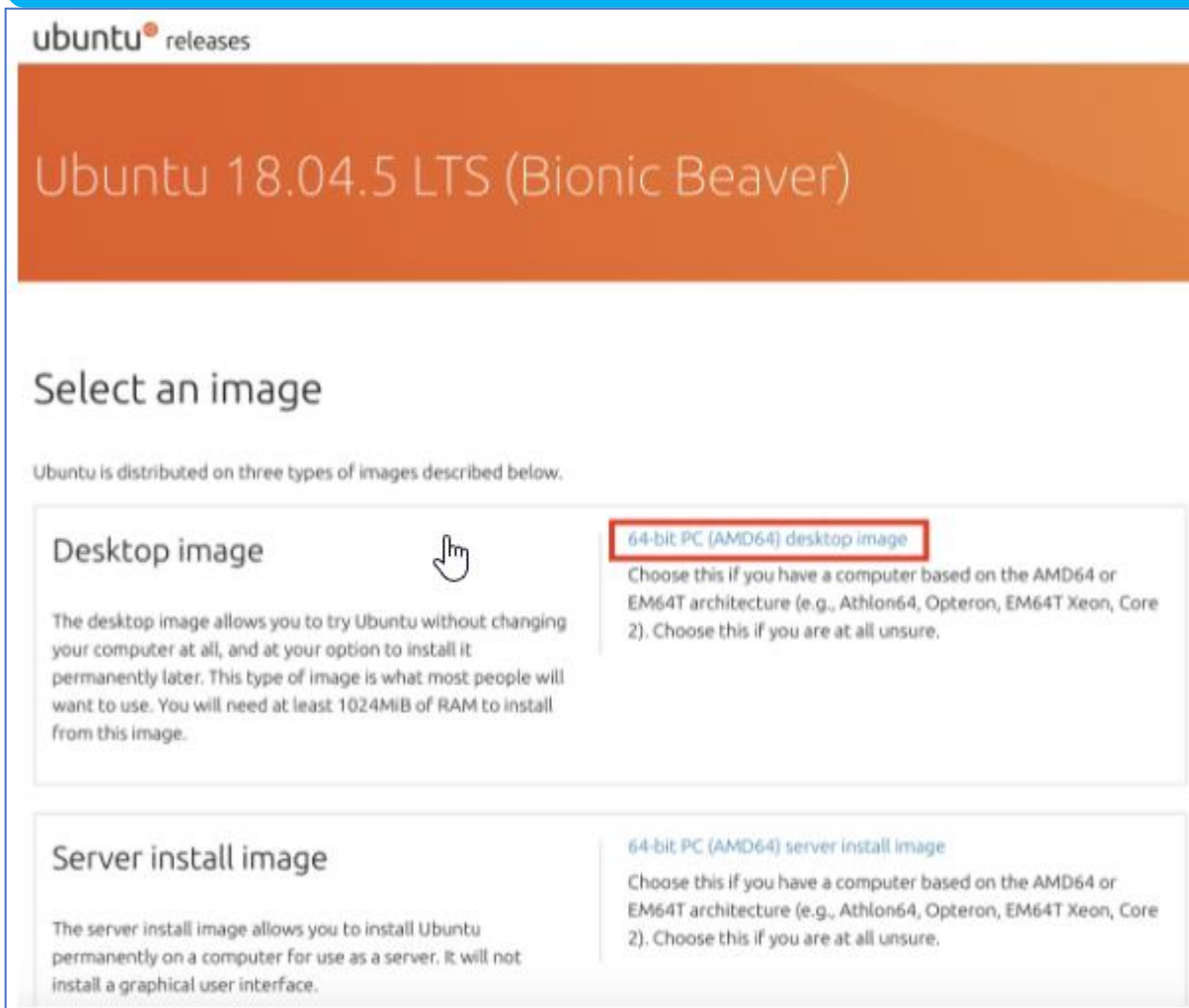
- [Windows hosts](#) ←
- [OS X hosts](#)
- [Linux distributions](#)
- [Solaris hosts](#)
- [Solaris 11 IPS hosts](#)

VirtualBox 6.1.20 platform packages

- [Windows hosts](#)
- [OS X hosts](#) ←
- [Linux distributions](#)
- [Solaris hosts](#)
- [Solaris 11 IPS hosts](#)

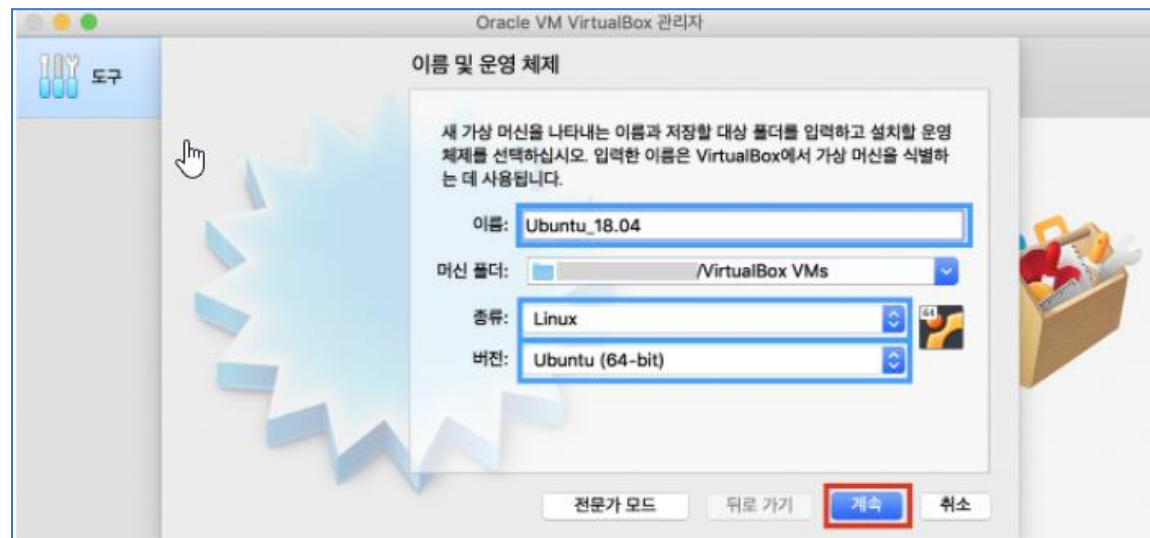
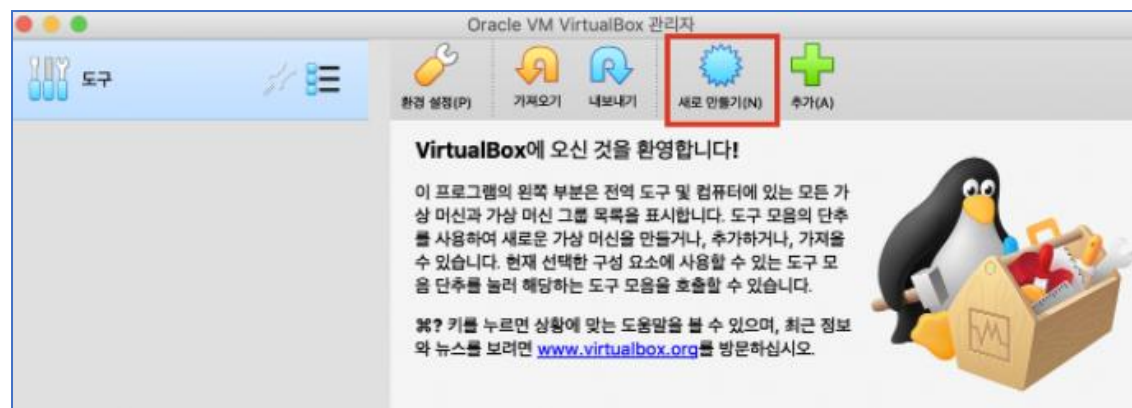
Ubuntu 설치(참고)

<http://mirror.kakao.com/ubuntu-releases/bionic/>



■ 설치방법 참고

<https://soobarkbar.tistory.com/215>



Spark 설치 (참고)

■ 시스템 패키지 업데이트

```
sudo apt update  
sudo apt -y upgrade
```

■ Java 설치

```
sudo apt-get install openjdk-8-jdk  
java -version
```

■ Apache Hadoop 다운로드 및 설치

```
wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz --no-check-certificate  
tar zxvf hadoop-3.3.0.tar.gz
```

설정방법 <https://eyeballs.tistory.com/420> 참고

■ Apache Spark 다운로드 및 설치

```
wget https://downloads.apache.org/spark/spark-3.1.1/spark-3.1.1-bin-hadoop3.2.tgz --no-check-certificate  
tar zxvf spark-3.1.1-bin-hadoop3.2.tgz
```

설정방법 <https://eyeballs.tistory.com/422> 참고

Spark 설치 (참고)

■ Standalone Master Server 시작

```
start-master.sh  
sudo ss -tunelp | grep 8080
```

■ Spark Worker Process 시작

```
start-worker.sh spark://127.0.0.1:7077
```

Spark Master at spark://dev:7077

URL: spark://dev:7077
Alive Workers: 0
Cores in use: 0 Total, 0 Used
Memory in use: 0.0 B Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

▼ Workers (0)

Worker Id	Address	State	Cores	Memory
-----------	---------	-------	-------	--------

▼ Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time
----------------	------	-------	---------------------	------------------------	----------------

▼ Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time
----------------	------	-------	---------------------	------------------------	----------------

Spark 설치 (참고)

■ Spark shell 사용

SPARK_HOME/bin/spark-shell

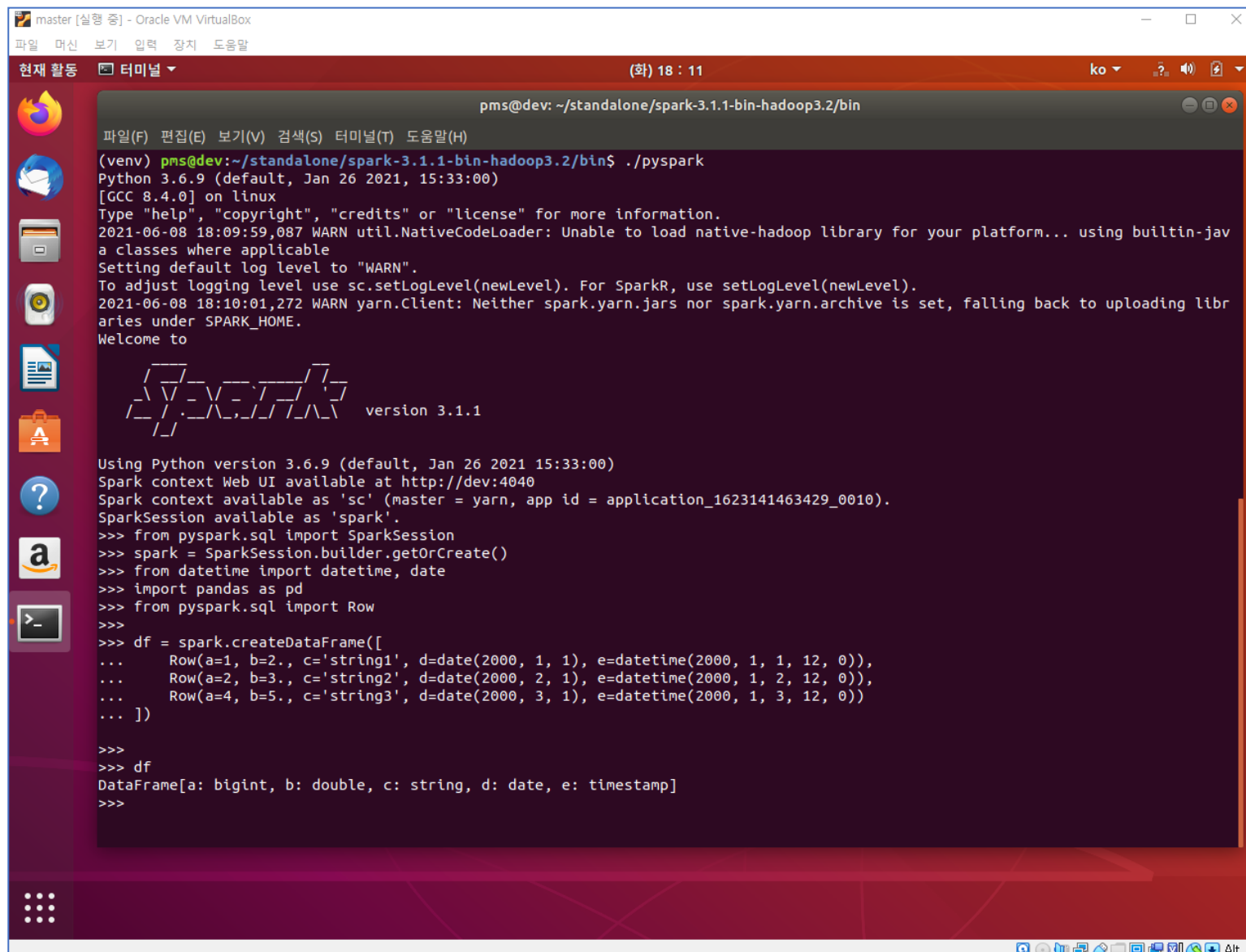
■ PySpark shell 사용(파이썬)

SPARK_HOME/bin/pyspark

■ Spark 정지

SPARK_HOME/sbin/stop-slave.sh

SPARK_HOME/sbin/stop-master.sh



```
master [실행 중] - Oracle VM VirtualBox
파일  머신  보기  입력  장치  도움말
현재 활동  터미널
(화) 18 : 11  ko

pms@dev: ~/standalone/spark-3.1.1-bin-hadoop3.2/bin

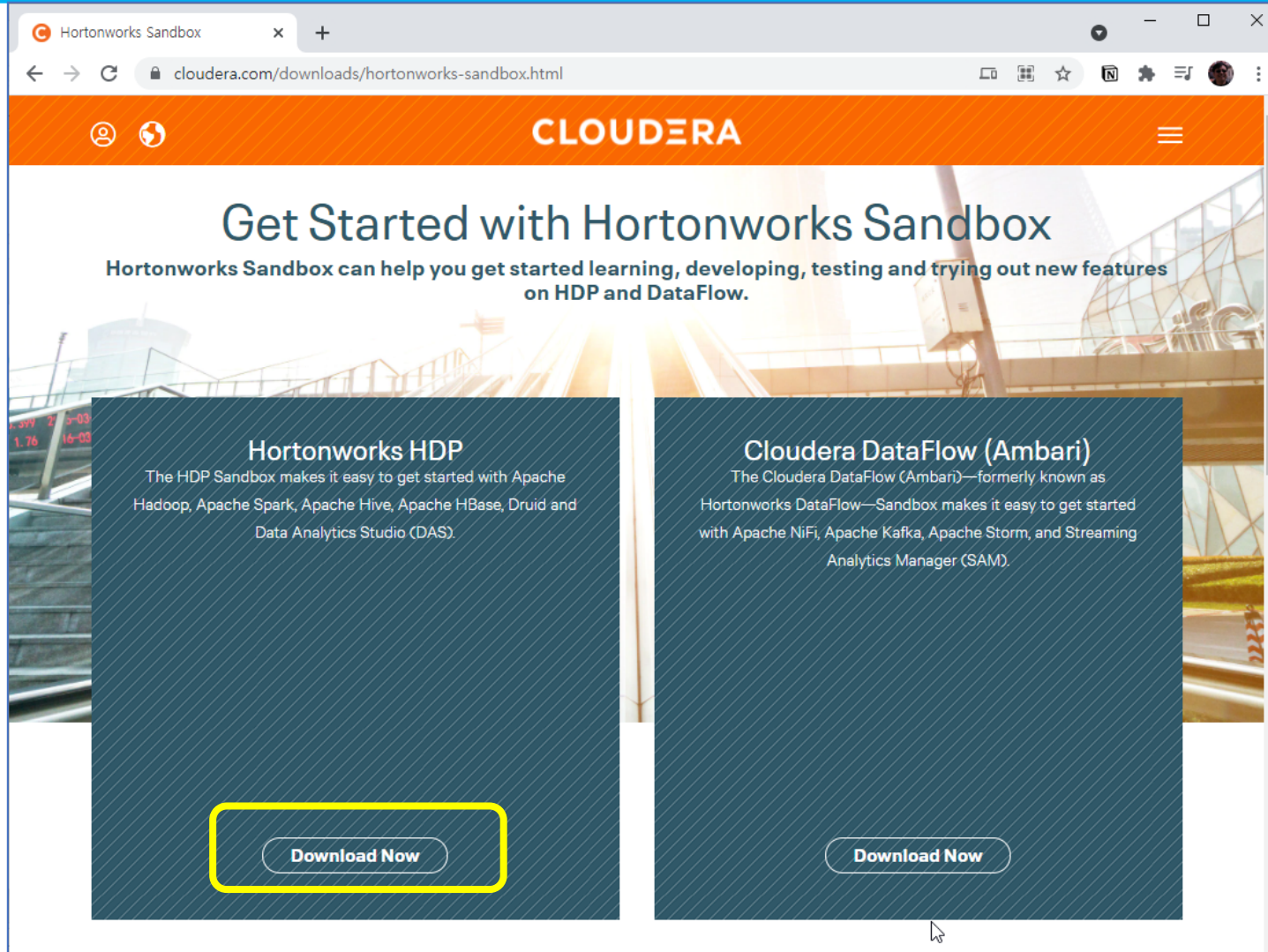
(venv) pms@dev:~/standalone/spark-3.1.1-bin-hadoop3.2/bin$ ./pyspark
Python 3.6.9 (default, Jan 26 2021, 15:33:00)
[GCC 8.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
2021-06-08 18:09:59,087 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
a classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
2021-06-08 18:10:01,272 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libr
aries under SPARK_HOME.
Welcome to

      _/ _ \| | | | _ \| | | |
     / ___ \| |_| | |_) | |_| |
    /_/   \_\_  \___|___|___|___|
    version 3.1.1

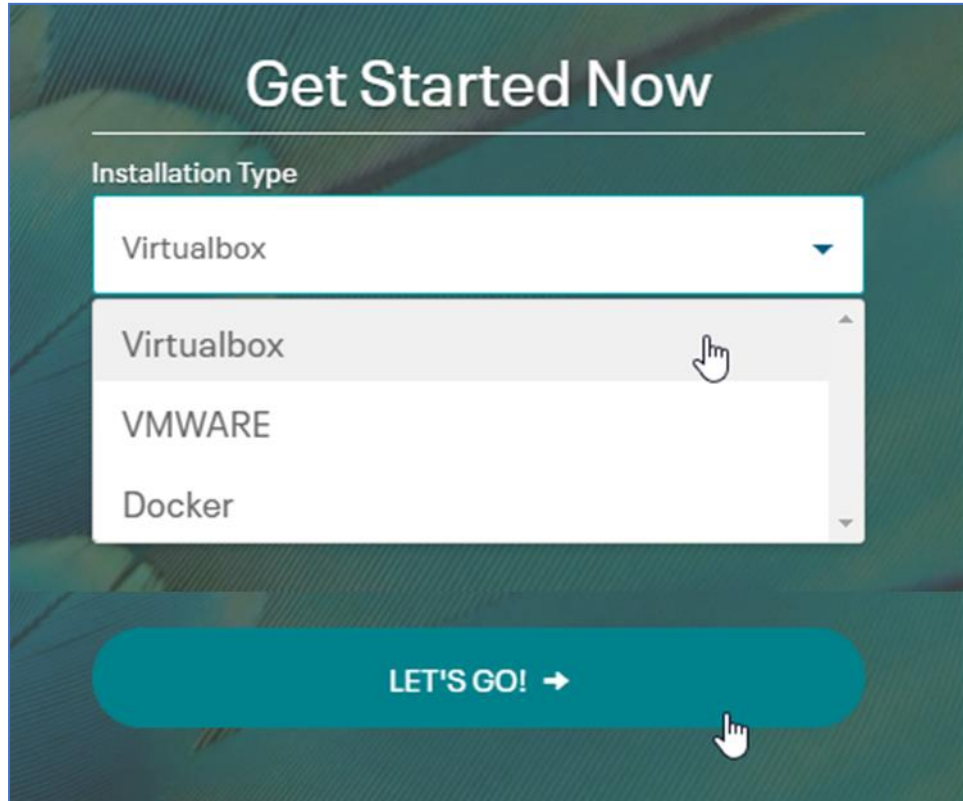
Using Python version 3.6.9 (default, Jan 26 2021 15:33:00)
Spark context Web UI available at http://dev:4040
Spark context available as 'sc' (master = yarn, app id = application_1623141463429_0010).
SparkSession available as 'spark'.
>>> from pyspark.sql import SparkSession
>>> spark = SparkSession.builder.getOrCreate()
>>> from datetime import datetime, date
>>> import pandas as pd
>>> from pyspark.sql import Row
>>>
>>> df = spark.createDataFrame([
...   Row(a=1, b=2., c='string1', d=date(2000, 1, 1), e=datetime(2000, 1, 1, 12, 0)),
...   Row(a=2, b=3., c='string2', d=date(2000, 2, 1), e=datetime(2000, 1, 2, 12, 0)),
...   Row(a=4, b=5., c='string3', d=date(2000, 3, 1), e=datetime(2000, 1, 3, 12, 0))
... ])
>>>
>>> df
DataFrame[a: bigint, b: double, c: string, d: date, e: timestamp]
>>>
```

HDP(Hortonworks Data Platform) 다운로드

<https://www.cloudera.com/downloads/hortonworks-sandbox.html>



HDP 다운로드



Sign in or complete our product interest form to continue.

Sign In

For self-learning


First Name
Danny

Last Name
Park

Business Email

Company

Job Title

 Phone

HDP 다운로드

Thank you for choosing Hortonworks Data Platform
(HDP) on Sandbox

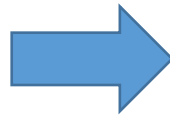
Sandbox HDP Virtualbox Downloads

HDP Sandbox 3.0.1 (Latest)

[Install Guide on VirtualBox](#)

Older Versions

- [2.6.5](#)
- [2.5.0](#)



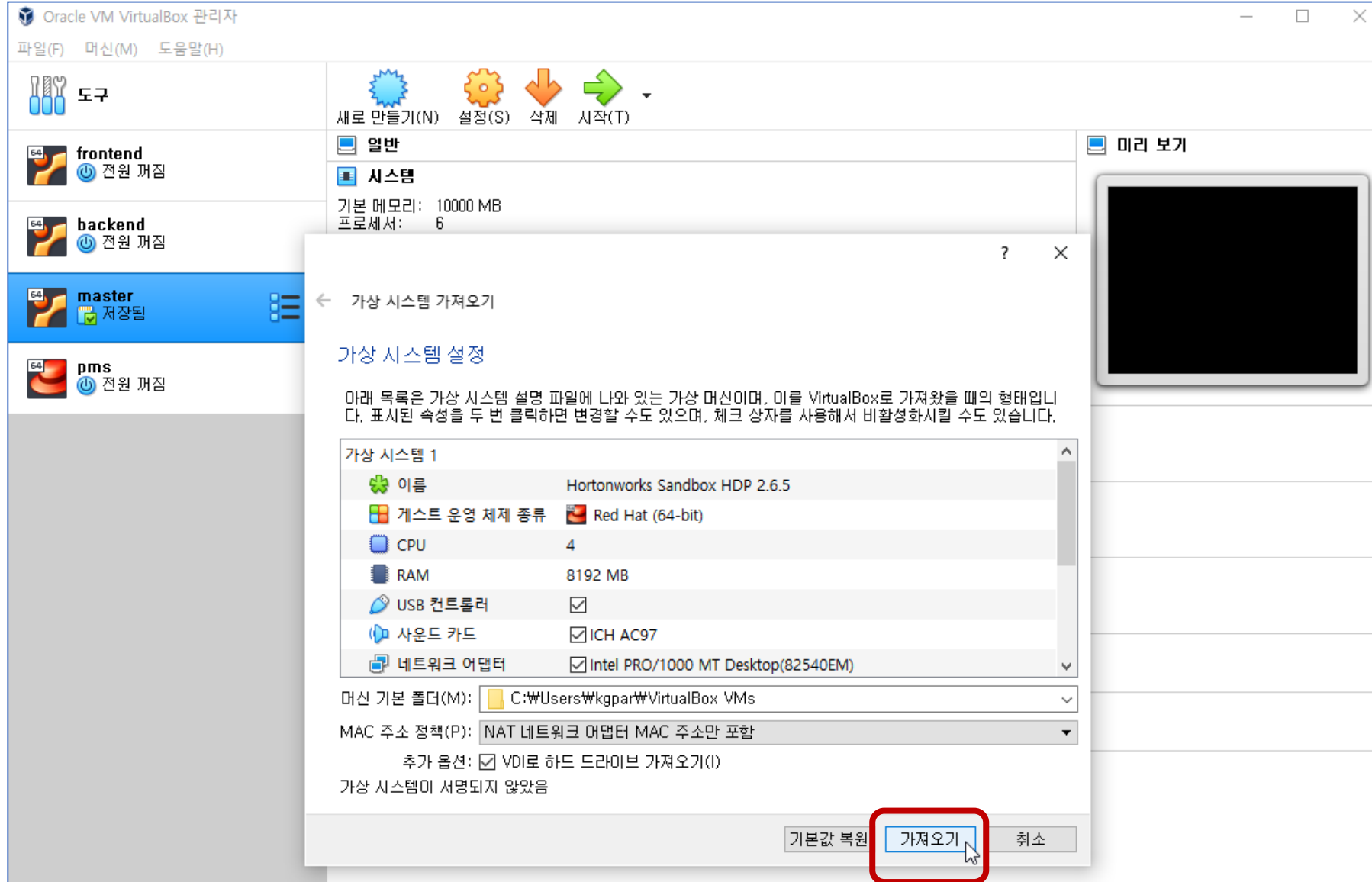
HDP_2.6.5_virtualbox_180626.ova 15.0GB

HDP 설치

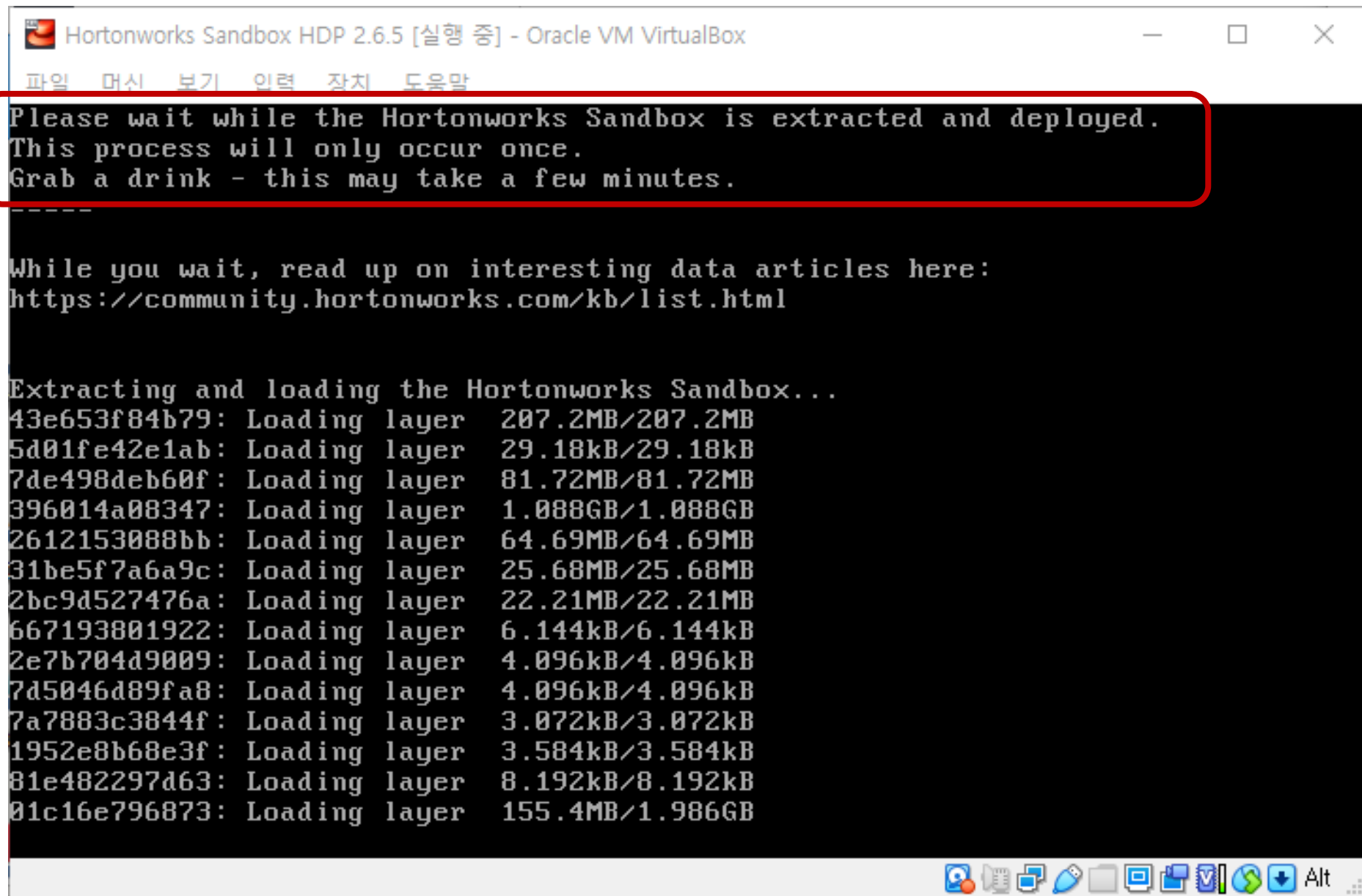
VirtualBox 실행



HDP 설치



HDP 설치



```
Hortonworks Sandbox HDP 2.6.5 [실행 중] - Oracle VM VirtualBox
파일  머신  보기  입력  장치  도움말

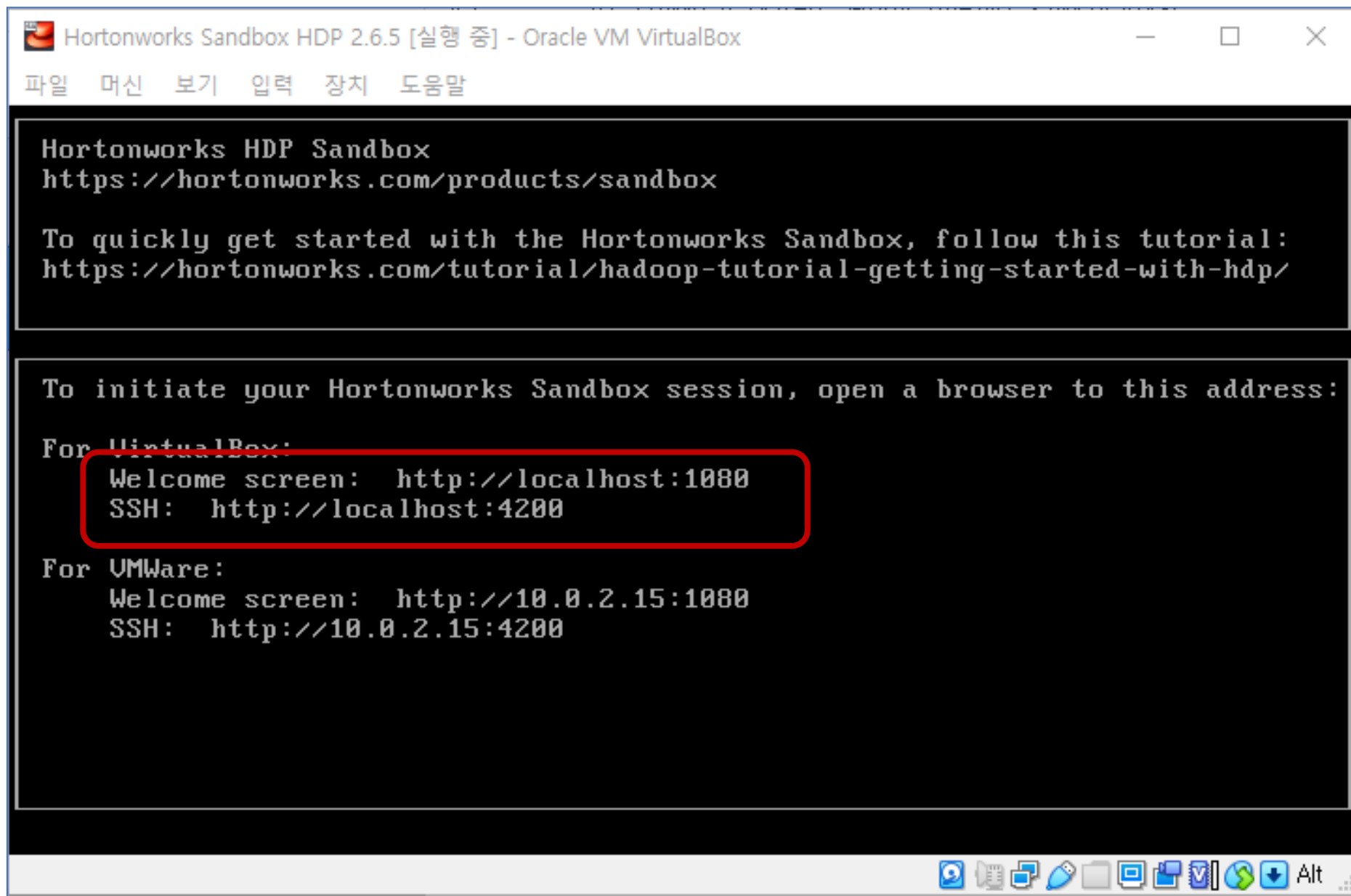
Please wait while the Hortonworks Sandbox is extracted and deployed.
This process will only occur once.
Grab a drink - this may take a few minutes.

-----

While you wait, read up on interesting data articles here:
https://community.hortonworks.com/kb/list.html


Extracting and loading the Hortonworks Sandbox...
43e653f84b79: Loading layer 207.2MB/207.2MB
5d01fe42e1ab: Loading layer 29.18kB/29.18kB
7de498deb60f: Loading layer 81.72MB/81.72MB
396014a08347: Loading layer 1.088GB/1.088GB
2612153088bb: Loading layer 64.69MB/64.69MB
31be5f7a6a9c: Loading layer 25.68MB/25.68MB
2bc9d527476a: Loading layer 22.21MB/22.21MB
667193801922: Loading layer 6.144kB/6.144kB
2e7b704d9009: Loading layer 4.096kB/4.096kB
7d5046d89fa8: Loading layer 4.096kB/4.096kB
7a7883c3844f: Loading layer 3.072kB/3.072kB
1952e8b68e3f: Loading layer 3.584kB/3.584kB
81e482297d63: Loading layer 8.192kB/8.192kB
01c16e796873: Loading layer 155.4MB/1.986GB
```

HDP 사용



HDP 사용

<http://127.0.0.1:1080/>



The screenshot shows a web browser window with the title "Hortonworks Sandbox with HDP". The address bar displays "127.0.0.1:1080/splash.html". The page features the Hortonworks logo in the top left and a "GET HELP" button in the top right. The main heading is "SAND BOX HDP2.6.5", where "BOX" is inside an orange hexagon. Below this, there are two columns. The left column is for "NEW TO HDP", featuring a green hexagon with the HDP logo and the text "Explore the Hortonworks Data Platform (HDP)" and "Walk through a typical use case with the tutorial". A red rectangle highlights a green "LAUNCH DASHBOARD" button. The right column is for "ADVANCED HDP", featuring a blue hexagon with gear icons and the text "Expand your Hortonworks Data Platform (HDP) experience" and "Access components in Sandbox". A green "QUICK LINKS" button is at the bottom of this column.

Hortonworks Sandbox with HDP

127.0.0.1:1080/splash.html

GET HELP

SAND BOX HDP2.6.5

NEW TO HDP

HDP
HORTONWORKS
DATA PLATFORM
powered by Apache Hadoop®

Explore the Hortonworks Data Platform (HDP)
Walk through a typical use case with the tutorial

LAUNCH DASHBOARD

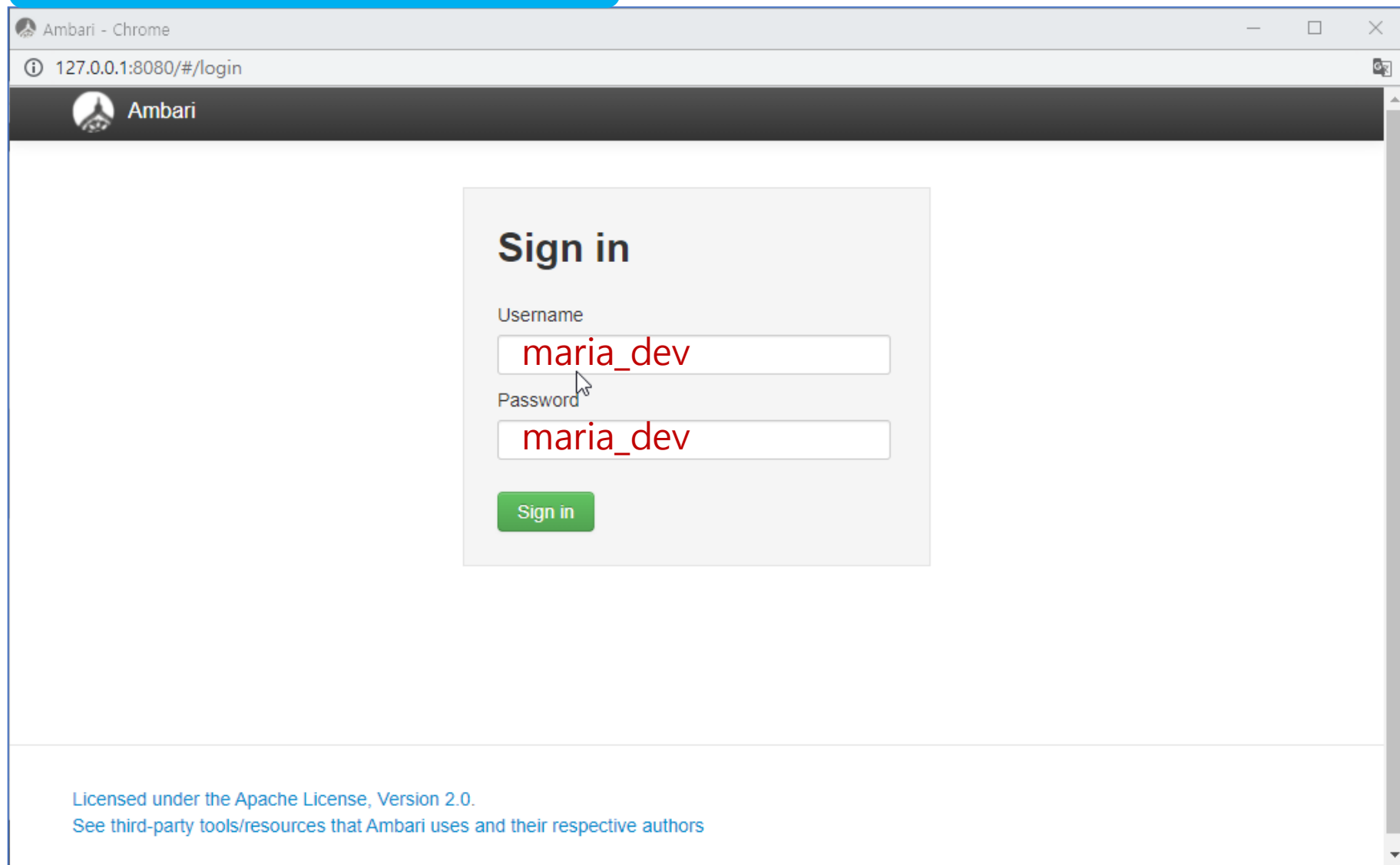
ADVANCED HDP

Expand your Hortonworks Data Platform (HDP) experience
Access components in Sandbox

QUICK LINKS

HDP 사용

<http://127.0.0.1:8080/>

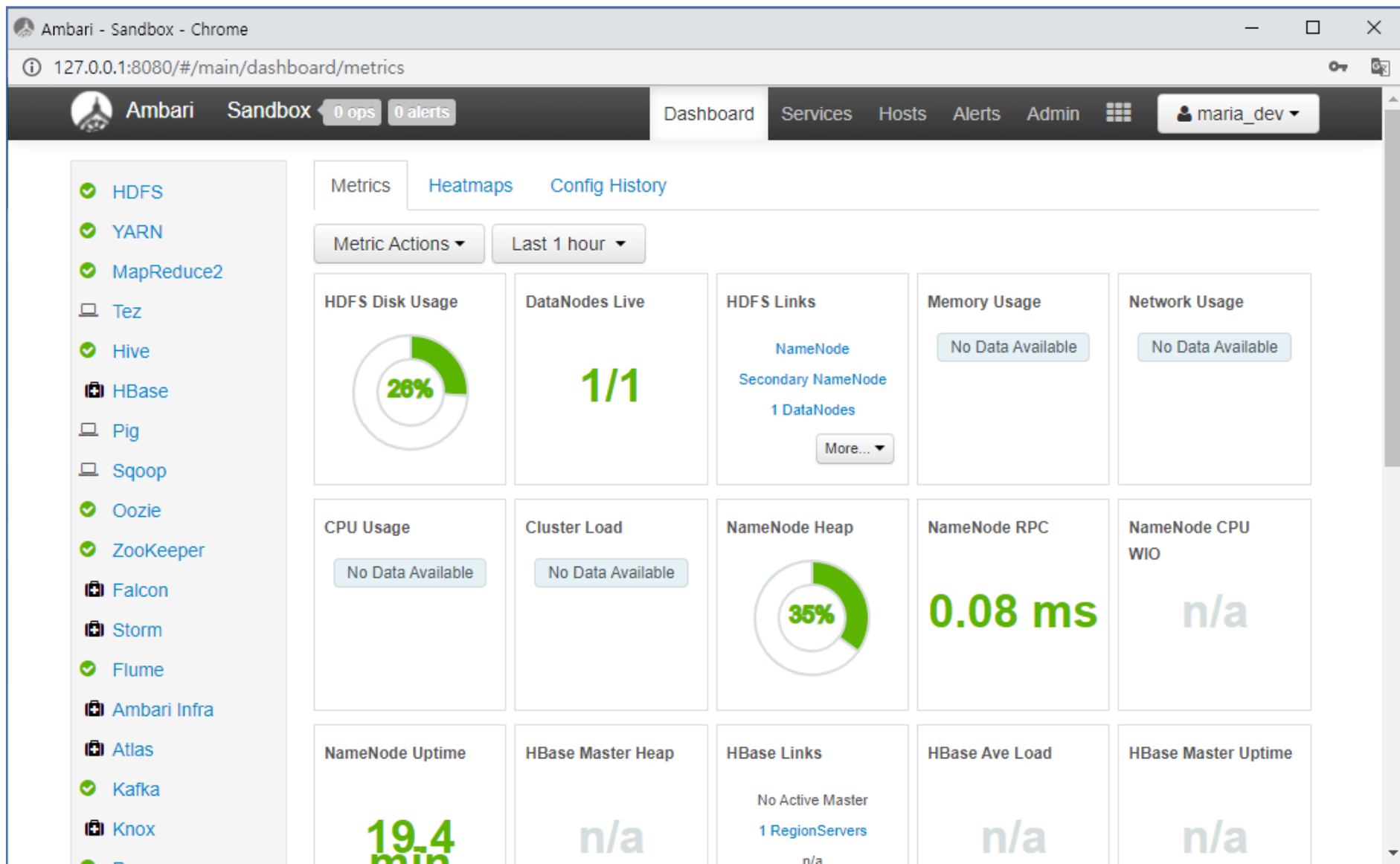


The screenshot shows a web browser window titled "Ambari - Chrome" with the address bar displaying "127.0.0.1:8080/#/login". The page features a dark header with the Ambari logo and name. The main content area contains a "Sign in" form with the following elements:

- Sign in** (Section Header)
- Username** label above a text input field containing "maria_dev".
- Password** label above a text input field containing "maria_dev".
- A green **Sign in** button.

At the bottom of the page, there is a footer with the text: "Licensed under the Apache License, Version 2.0. See third-party tools/resources that Ambari uses and their respective authors".

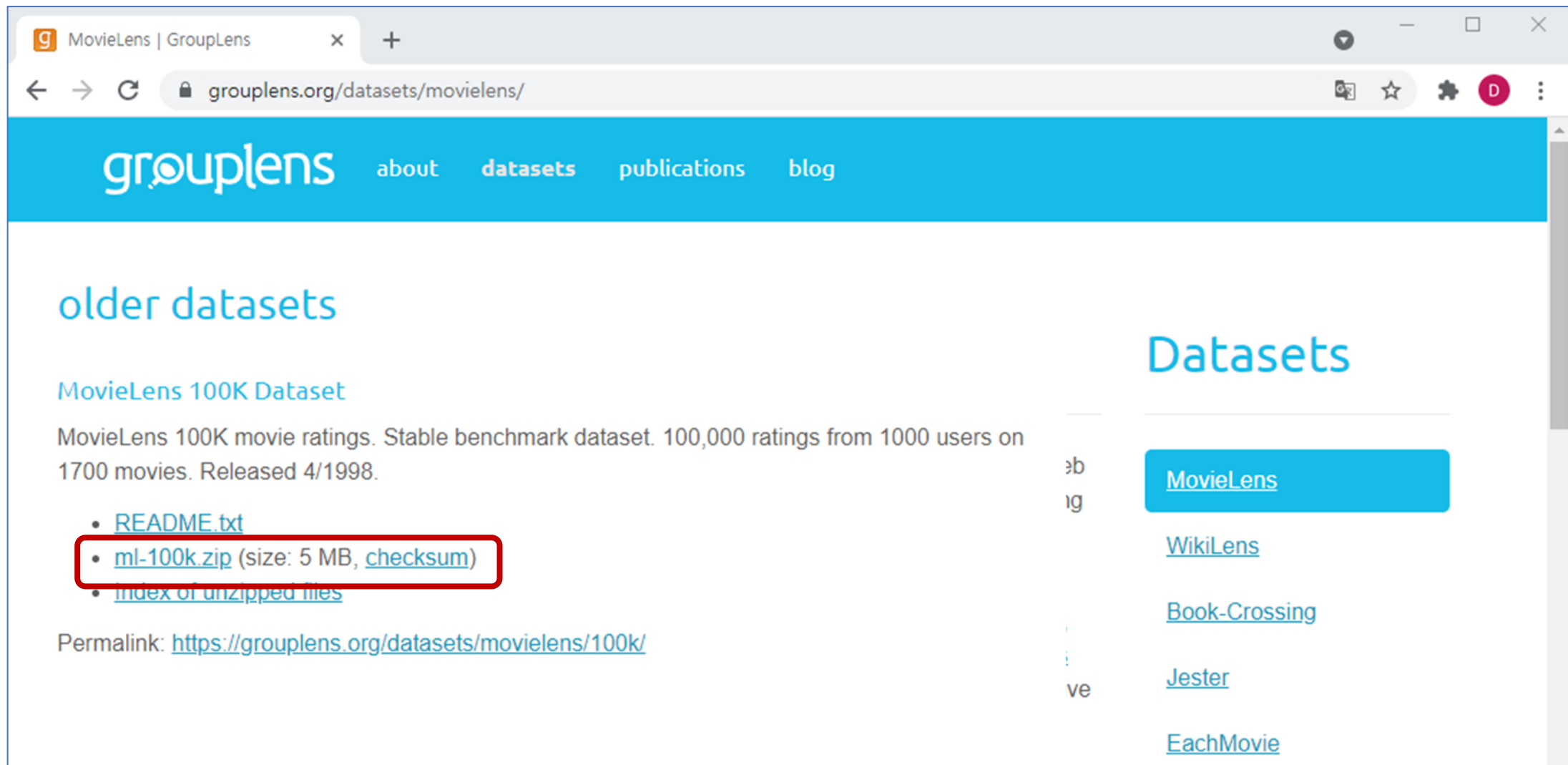
HDP 사용



Hadoop 실습

MovieLens 데이터 실습

<https://grouplens.org/datasets/movielens/>



The screenshot shows the MovieLens website interface. The browser's address bar displays the URL <https://grouplens.org/datasets/movielens/>. The website has a blue header with the 'grouplens' logo and navigation links for 'about', 'datasets', 'publications', and 'blog'. The main content area is titled 'older datasets' and features the 'MovieLens 100K Dataset' section. This section describes the dataset as 'MovieLens 100K movie ratings. Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies. Released 4/1998.' Below the description is a list of links: 'README.txt', 'ml-100k.zip (size: 5 MB, checksum)', and 'index of unzipped files'. The 'ml-100k.zip' link is highlighted with a red rectangular box. At the bottom of the section, a 'Permalink' is provided: <https://grouplens.org/datasets/movielens/100k/>. On the right side of the page, there is a 'Datasets' sidebar with a list of dataset names: 'MovieLens', 'WikiLens', 'Book-Crossing', 'Jester', and 'EachMovie'. The 'MovieLens' link in this sidebar is highlighted with a blue button.

grouplens | GroupLens

about datasets publications blog

older datasets

MovieLens 100K Dataset

MovieLens 100K movie ratings. Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies. Released 4/1998.

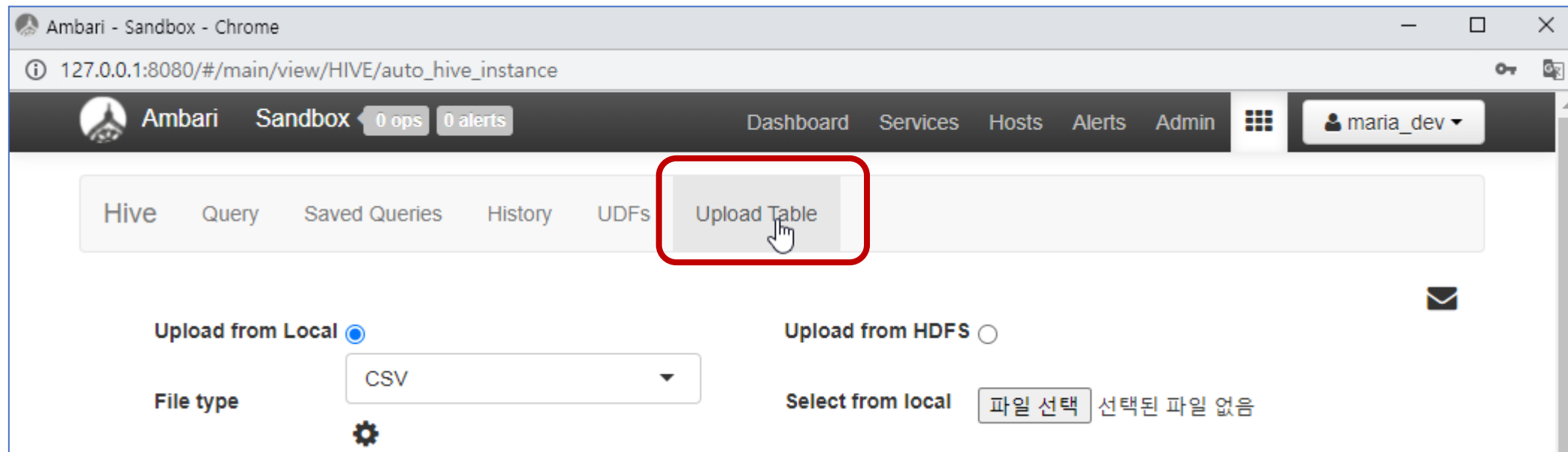
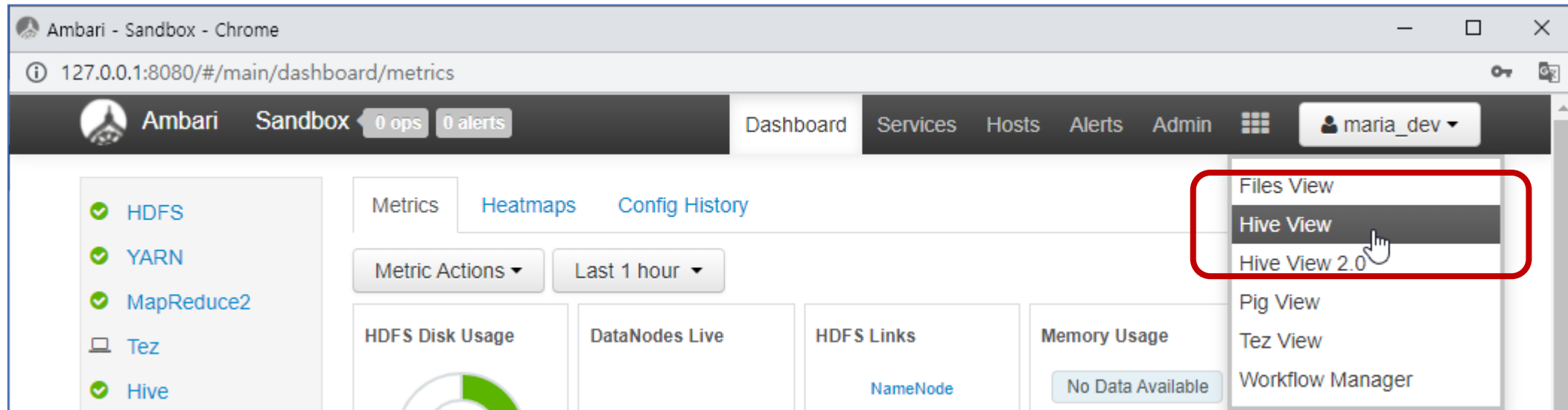
- [README.txt](#)
- [ml-100k.zip](#) (size: 5 MB, [checksum](#))
- [index of unzipped files](#)

Permalink: <https://grouplens.org/datasets/movielens/100k/>

Datasets

- [MovieLens](#)
- [WikiLens](#)
- [Book-Crossing](#)
- [Jester](#)
- [EachMovie](#)

데이터 업로드



u.data 데이터 업로드

Ambari - Sandbox - Chrome
127.0.0.1:8080/#/main/view/HIVE/auto_hive_instance

Hive Query Saved Queries History UDFs Upload Table

Upload from Local ☒ Upload from HDFS ☐

File type CSV Database default Stored as ORC

Field Delimiter: | Escape Character: Quote Character: Is first row header?

9 TAB(horizontal tab) 11 VT(vertical tab) 12 FF(NP form feed - new page) 14 SO(shift out) 15 SI(shift in) 16 DLE(data link escape) 17 DC1(device control 1)

1 2 3 4 5 6

파일 선택 u.data ratings ratings

Upload Table


user_id	movie_id	rating	rating_time
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	246	1	880606923

u.item 데이터 업로드


Ambari - Sandbox - Chrome
127.0.0.1:8080/#/main/view/HIVE/auto_hive_instance

Hive Query Saved Queries History UDFs Upload Table

Upload from Local ☒ Upload from HDFS ☐

File type CSV **1** 

Database default

Stored as ORC 

Select from **3** 파일 선택 u.item **u_item**

Table name **4** movie_names **movie_names**

Contains endlines? ☐

6 Upload Table

5

movie_id	name	column3	column4
1	Toy Story (1995)	01-Jan-1995	
2	GoldenEye (1995)	01-Jan-1995	
3	Four Rooms (1995)	01-Jan-1995	

Field Delim **2** ||

Escape Character:

Quote Character:

Is first row header ?

119 w
120 x
121 y
122 z
123 {
124 |
125 }

Hive

The screenshot shows the Ambari web interface for a Hive instance. The browser address bar indicates the URL `127.0.0.1:8080/#/main/view/HIVE/auto_hive_instance`. The top navigation bar includes links for Dashboard, Services, Hosts, Alerts, and Admin, along with a user profile dropdown for `maria_dev`. Below this, a secondary navigation bar contains tabs for Hive, Query, Saved Queries, History, UDFs, and Upload Table. The `Hive` tab is currently selected and highlighted with a red box. On the right side of this bar, a dropdown menu is open, listing several options: Files View, Hive View (highlighted with a red box and a mouse cursor), Hive view 2.0, Pig View, Tez View, and Workflow Manager. The main interface is divided into two panels. The left panel, titled 'Database Explorer', shows a tree view of the database structure under the 'default' schema, listing tables like 'movie_names' and their columns (movie_id, name, column3, column4, column5). The right panel, titled 'Query Editor', displays a 'Worksheet' with a single line of text '1 |'. On the far right, a vertical sidebar contains icons for SQL, settings, a chart, a link, and a TEZ icon with a red notification badge.

Ambari - Sandbox

127.0.0.1:8080/#/main/view/HIVE/auto_hive_instance

Ambari Sandbox 0 ops 0 alerts

Dashboard Services Hosts Alerts Admin

maria_dev

Hive Query Saved Queries History UDFs Upload Table

Files View

Hive View

Hive view 2.0

Pig View

Tez View

Workflow Manager

Database Explorer

default

Search tables...

Databases

default

movie_names

movie_id INT

name STRING

column3 STRING

column4 STRING

column5 STRING

Query Editor

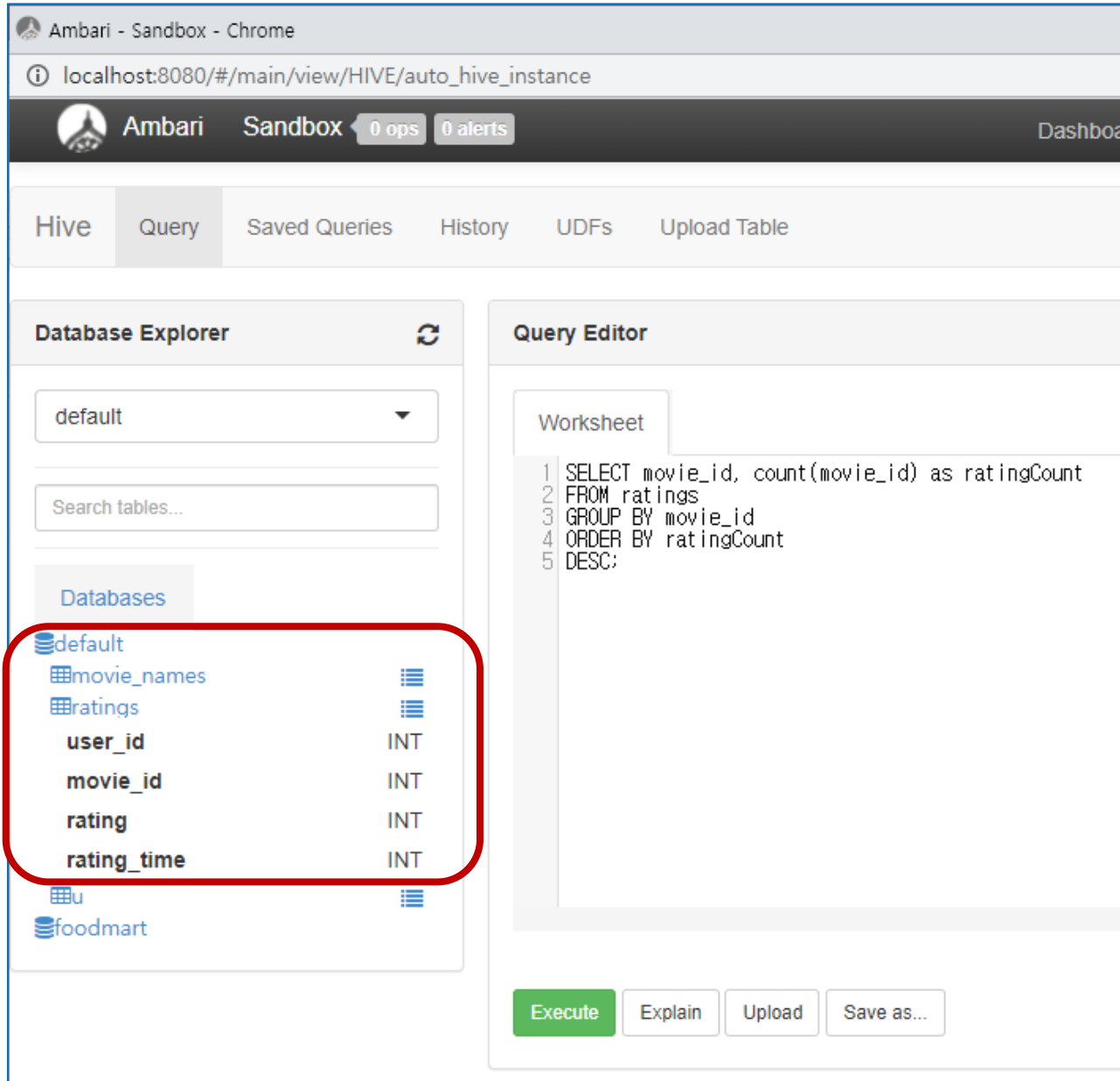
Worksheet

1 |

SQL

TEZ

Hive



Ambari - Sandbox - Chrome

localhost:8080/#/main/view/HIVE/auto_hive_instance

Ambari Sandbox 0 ops 0 alerts Dashboard

Hive Query Saved Queries History UDFs Upload Table

Database Explorer

default

Search tables...

Databases

- default
 - movie_names
 - ratings
 - user_id INT
 - movie_id INT
 - rating INT
 - rating_time INT
- u
- foodmart

Query Editor

Worksheet

```
1 SELECT movie_id, count(movie_id) as ratingCount
2 FROM ratings
3 GROUP BY movie_id
4 ORDER BY ratingCount
5 DESC;
```

Execute Explain Upload Save as...

```
SELECT movie_id, count(movie_id) as ratingCount
FROM ratings
GROUP BY movie_id
ORDER BY ratingCount
DESC;
```



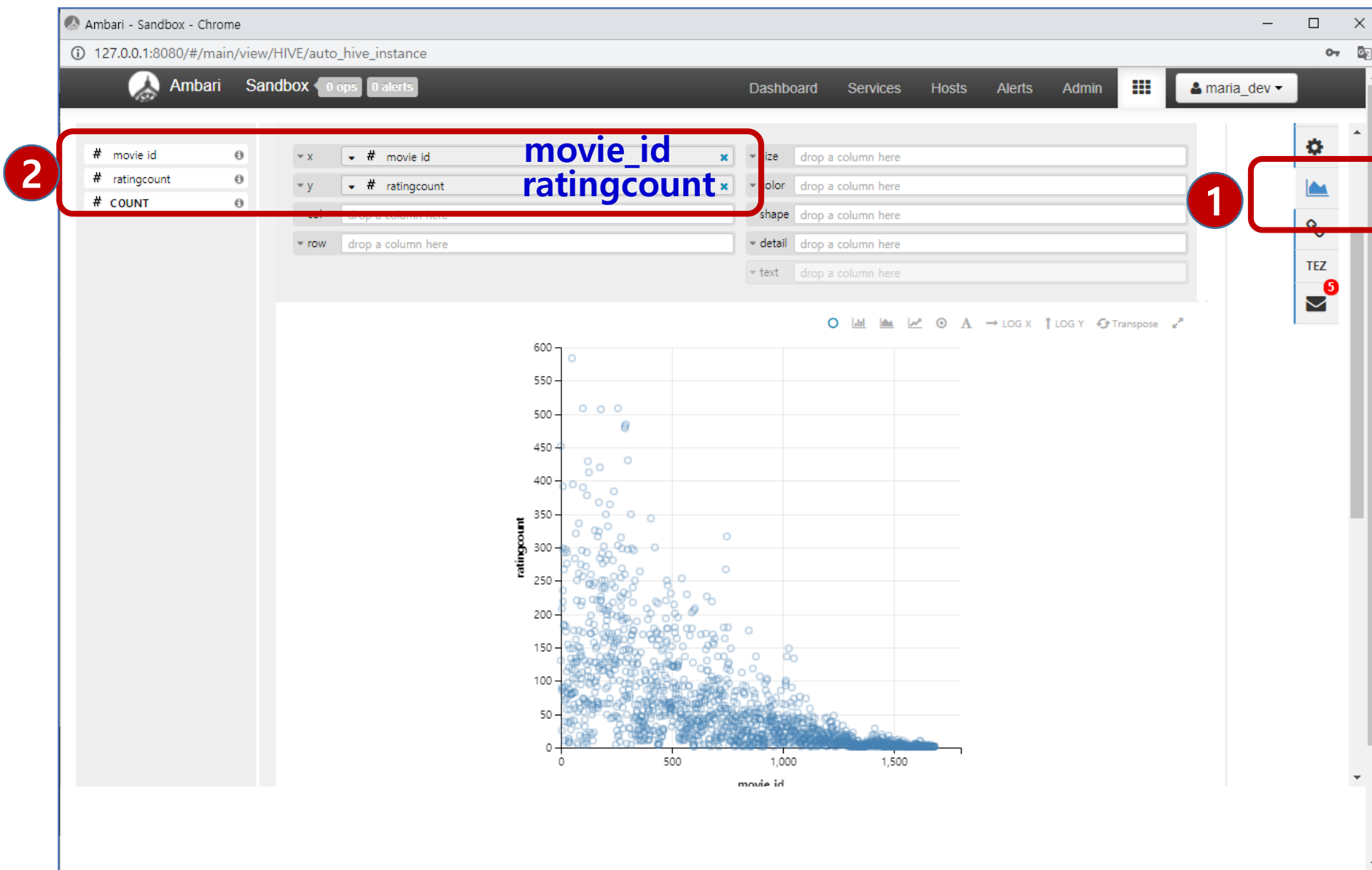
Query Process Results (Status: SUCCEEDED)

Logs Results

Filter columns...

movie_id	ratingcount
50	583
258	509
100	508
181	507

Hive



Hive

Ambari - Sandbox - Chrome

127.0.0.1:8080/#/main/view/HIVE/auto_hive_instance

Search tables...

Databases

- default
 - movie_names
 - movie_id INT
 - name STRING
 - column3 STRING
 - column4 STRING
 - column5 STRING
 - column6 INT
 - column7 INT
 - column8 INT
 - column9 INT
 - column10 INT
 - ratings
 - foodmart

Load more...

```
1 SELECT name
2 FROM movie_names
3 WHERE movie_id = 50;
```

SELECT name
FROM movie_names
WHERE movie_id = 50;

Execute Explain Upload Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

previous next

name

Star Wars (1977)

HDFS

The screenshot shows the Ambari web interface for the HDFS service. A red box labeled '1' highlights the 'HDFS' service in the left-hand navigation menu. Another red box labeled '2' highlights the user profile dropdown menu in the top right corner, which is open and showing a list of view options: 'Files View', 'Hive View', 'Hive View 2.0', 'Pig View', 'Tez View', and 'Workflow Manager'. A mouse cursor is pointing at 'Files View'.

Ambari - Sandbox
127.0.0.1:8080/#/main/services/HDFS/summary

Ambari Sandbox 0 ops 0 alerts Dashboard Services Hosts Alerts Admin

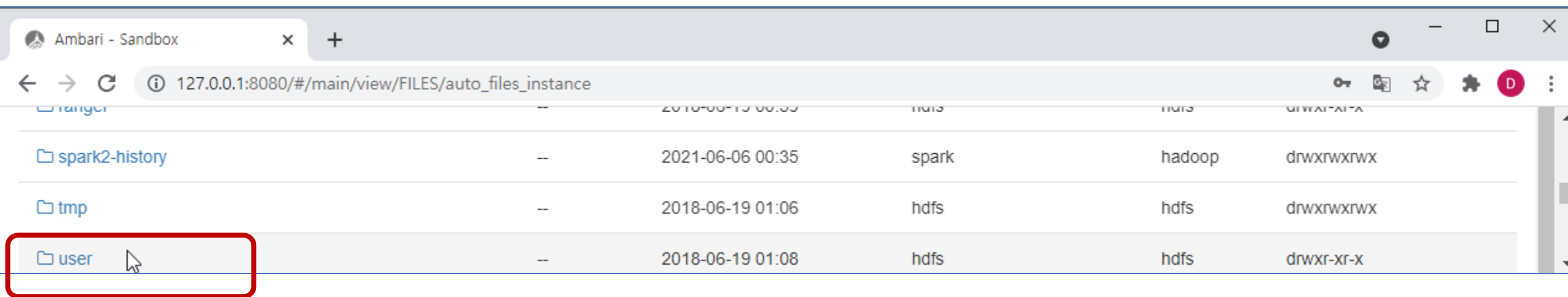
1 HDFS

Summary Heatmaps Configs Quick Links

Summary

NameNode	Started	No alerts	Disk Remaining	7
SNameNode	Started	No alerts		
DataNodes	1/1	Started	Blocks (total)	1
DataNodes Status	1 live / 0 dead / 0 decommissioning		Block Errors	0 corrupt replica / 0 missing / 0 under replicated
JournalNodes	0/0	JournalNodes Live	Total Files + Directories	1379
NFSGateways	0/0	Started	Upgrade Status	No pending upgrade
NameNode Uptime	4.37 hours		Safe Mode Status	Not in safe mode
NameNode Heap	55.3 MB / 240.0 MB (23.0% used)			
Disk Usage (DFS Used)	1.9 GB / 106.0 GB (1.77%)			
Disk Usage (Non DFS Used)	25.9 GB / 106.0 GB (24.41%)			

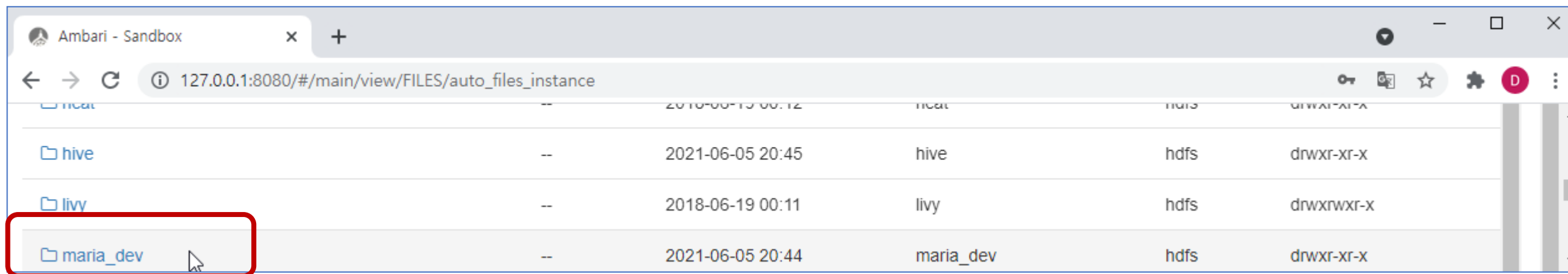
HDFS



Ambari - Sandbox

127.0.0.1:8080/#/main/view/FILES/auto_files_instance

spark2-history	--	2021-06-06 00:35	spark	hadoop	drwxrwxrwx
tmp	--	2018-06-19 01:06	hdfs	hdfs	drwxrwxrwx
user	--	2018-06-19 01:08	hdfs	hdfs	drwxr-xr-x



Ambari - Sandbox

127.0.0.1:8080/#/main/view/FILES/auto_files_instance

hive	--	2021-06-05 20:45	hive	hdfs	drwxr-xr-x
livy	--	2018-06-19 00:11	livy	hdfs	drwxrwxr-x
maria_dev	--	2021-06-05 20:44	maria_dev	hdfs	drwxr-xr-x

HDFS

Ambari - Sandbox x +

127.0.0.1:8080/#/main/view/FILES/auto_files_instance

Ambari Sandbox 0 ops 0 alerts Dashboard Services Hosts Alerts Admin maria_dev

/ > user > maria_dev Total: 1 files or folders

+ All New Folder Upload

Search in current directory...

Group > Permission

hdfs drwxr-xr-x

Add new folder

Name

ml-100k ml-100k

Cancel + Add

Ambari - Sandbox x +

127.0.0.1:8080/#/main/view/FILES/auto_files_instance

hdfs drwxr-xr-x

hdfs drwxr-xr-x

hive

ml-100k

--	2021-06-05 20:44	maria_dev	hdfs	drwxr-xr-x
--	2021-06-06 00:38	maria_dev	hdfs	drwxr-xr-x

HDFS

Ambari - Sandbox

127.0.0.1:8080/#/main/view/FILES/auto_files_instance

Ambari Sandbox 0 ops 0 alerts

Dashboard Services Hosts Alerts Admin

maria_dev

/ > user > maria_dev > ml-100k

Total: 0 files or folders

+ Select All

1

Upload

Search in current directory...

Group > Permission

2

u.data

u.item

Drag file to upload or click to browse

Currently supports single file upload

Cancel

HDFS

Ambari - Sandbox

127.0.0.1:8080/#/main/view/FILES/auto_files_instance

Ambari Sandbox

0 ops0 alerts

DashboardServicesHostsAlertsAdmin

maria_dev

Home

Files

Refresh

/ > user > maria_dev > ml-100k

Total: 2 files or folders

+ Select All




New Folder

Upload

1

Search in current directory...

Q

Name >	Size >	Last Modified >	Owner >	Group >	Permission
					
 u.data	1.9 MB	2021-06-06 00:41	maria_dev	hdfs	-rw-r--r--
 u.item	230.8 kB	2021-06-06 00:42	maria_dev	hdfs	-rw-r--r--

HDFS

The screenshot shows the Ambari web interface in a browser window. The address bar displays the URL `127.0.0.1:8080/#/main/view/FILES/auto_files_instance`. The top navigation bar includes links for Dashboard, Services, Hosts, Alerts, and Admin, along with a user profile dropdown for `maria_dev`. The main content area shows a file browser view for the path `/user/maria_dev/ml-100k`, indicating `1 Files, 0 Folders selected`. A red box highlights the `Open` button in the file actions menu. A `File Preview` modal window is open, displaying the content of the file `/user/maria_dev/ml-100k/u.data`. The modal contains a table with four columns of data and a `Download` button at the bottom right.

Name
u.data
u.item

File Preview			
/user/maria_dev/ml-100k/u.data			
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596
298	474	4	884182806
115	265	2	881171488
253	465	5	891628467
305	451	3	886324817
6	86	3	883603013
62	257	2	879372434
286	1014	5	879781125
200	222	5	876042340
210	40	3	891035994
224	29	3	888104457
303	785	3	879485318
122	387	5	879270459
194	274	2	879539794
291	1042	4	874834944

HDFS

Ambari - Sandbox

127.0.0.1:8080/#/main/view/FILES/auto_files_instance

Ambari Sandbox 0 ops 0 alerts Dashboard Services Hosts Alerts Admin maria_dev

/ > user > maria_dev > ml-100k 2 Files, 0 Folders selected

Deselect All New Folder Upload

Open Rename Permissions Delete Copy Move Download concatenate Search in current directory...

Name	Size	Last Modified	Owner	Group	Permission
u.data	1.9 MB	2021-06-06 00:41	maria_dev	hdfs	-rw-r--r--
u.item	230.8 kB	2021-06-06 00:42	maria_dev	hdfs	-rw-r--r--

127.0.0.1:8080/views/FILES/1.0.0/AUTO_FILES_INSTANCE/#

HDFS

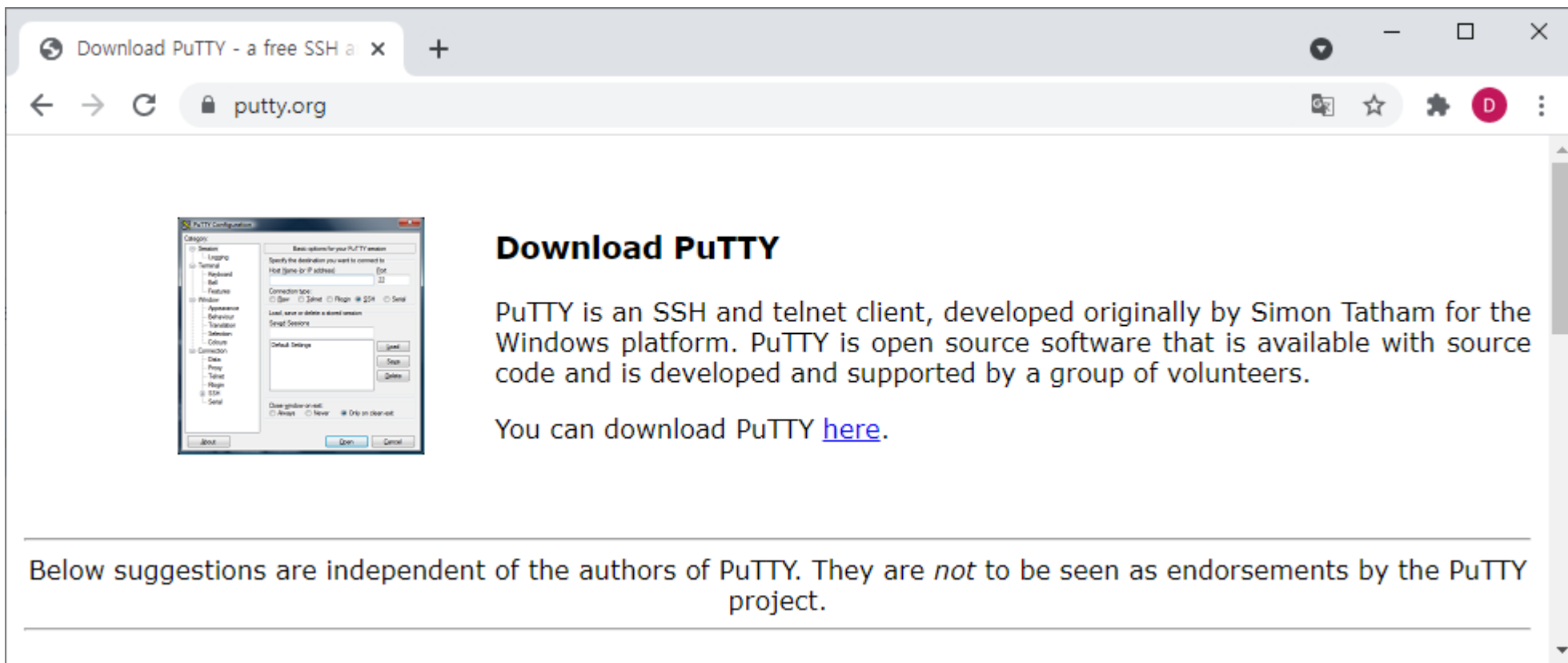
The screenshot shows the Ambari Sandbox web interface for managing HDFS files. The browser address bar shows the URL `127.0.0.1:8080/#/main/view/FILES/auto_files_instance`. The Ambari header includes navigation links for Dashboard, Services, Hosts, Alerts, and Admin, along with a user profile for `maria_dev`. The breadcrumb path is `/ > user > maria_dev`. A yellow status bar indicates `0 Files, 1 Folders selected`. The action bar contains icons for Open, Rename, Permissions, Delete (highlighted), Copy, Move, Download, and concatenate. A search bar is present on the right. The file list table has columns for Name, Size, Last Modified, Owner, Group, and Permission.

Name	Size	Last Modified	Owner	Group	Permission
↶					
📁 .Trash	--	2021-06-06 00:48	maria_dev	hdfs	drwxr-xr-x
📁 hive	--	2021-06-05 20:44	maria_dev	hdfs	drwxr-xr-x
📁 ml-100k	--	2021-06-06 00:48	maria_dev	hdfs	drwxr-xr-x

127.0.0.1:8080/views/FILES/1.0.0/AUTO_FILES_INSTANCE/#

HDFS (터미널 환경)

<https://www.putty.org/>



Download PuTTY

PuTTY is an SSH and telnet client, developed originally by Simon Tatham for the Windows platform. PuTTY is open source software that is available with source code and is developed and supported by a group of volunteers.

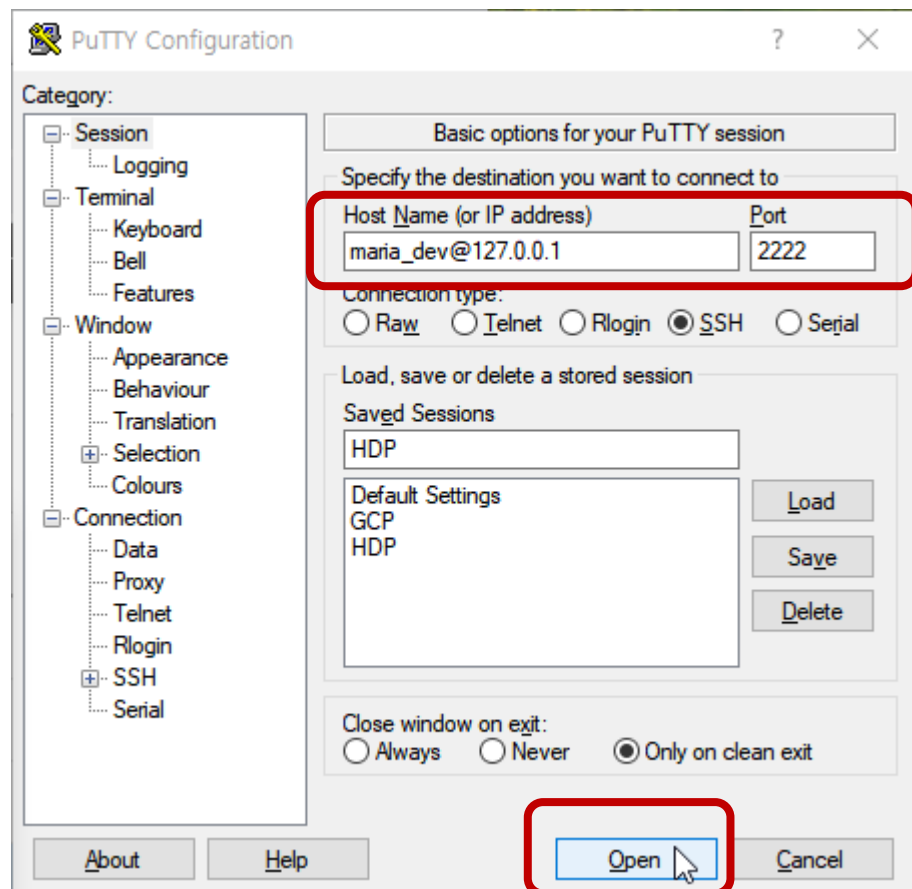
You can download PuTTY [here](#).

Below suggestions are independent of the authors of PuTTY. They are *not* to be seen as endorsements by the PuTTY project.

HDFS (터미널 환경)

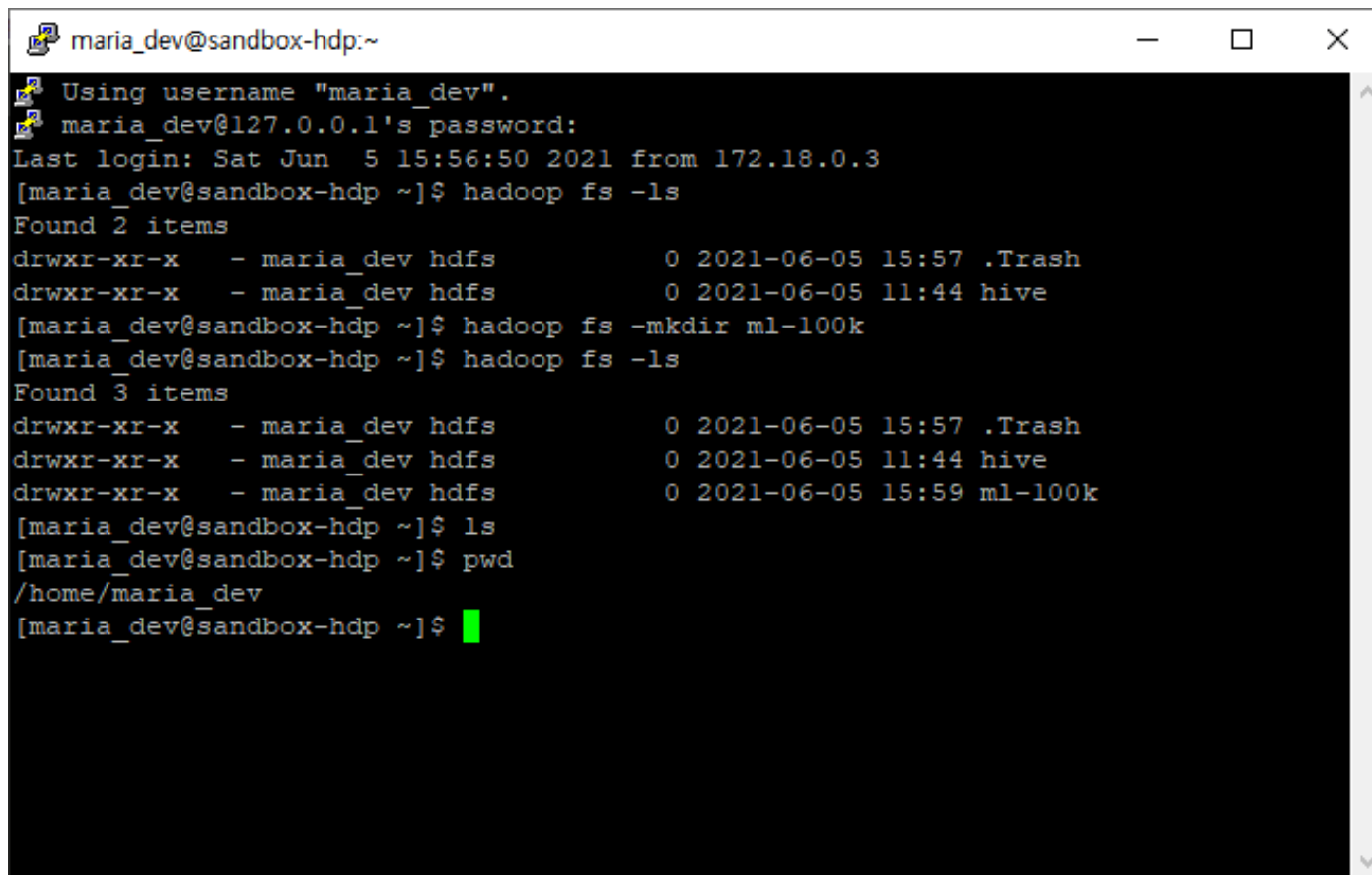


putty 실행



hadoop fs -ls

hadoop fs -mkdir ml-100k



HDFS(터미널 환경)

wget https://github.com/kgpark88/bigdata/raw/main/ml-100k/u.data

ls

```
maria_dev@sandbox-hdp:~  
[maria_dev@sandbox-hdp ~]$ wget https://github.com/kgpark88/bigdata/raw/main/ml-100k/u.data  
--2021-06-05 16:12:19-- https://github.com/kgpark88/bigdata/raw/main/ml-100k/u.data  
Resolving github.com (github.com)... 15.164.81.167  
Connecting to github.com (github.com)|15.164.81.167|:443... connected.  
HTTP request sent, awaiting response... 302 Found  
Location: https://raw.githubusercontent.com/kgpark88/bigdata/main/ml-100k/u.data [following]  
--2021-06-05 16:12:19-- https://raw.githubusercontent.com/kgpark88/bigdata/main/ml-100k/u.data  
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.111.133, ...  
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 1979173 (1.9M) [text/plain]  
Saving to: 'u.data'  
  
100%[=====>] 1,979,173 8.33MB/s in 0.2s  
  
2021-06-05 16:12:19 (8.33 MB/s) - 'u.data' saved [1979173/1979173]  
  
[maria_dev@sandbox-hdp ~]$ ls  
u.data  
[maria_dev@sandbox-hdp ~]$
```

HDFS(터미널 환경)

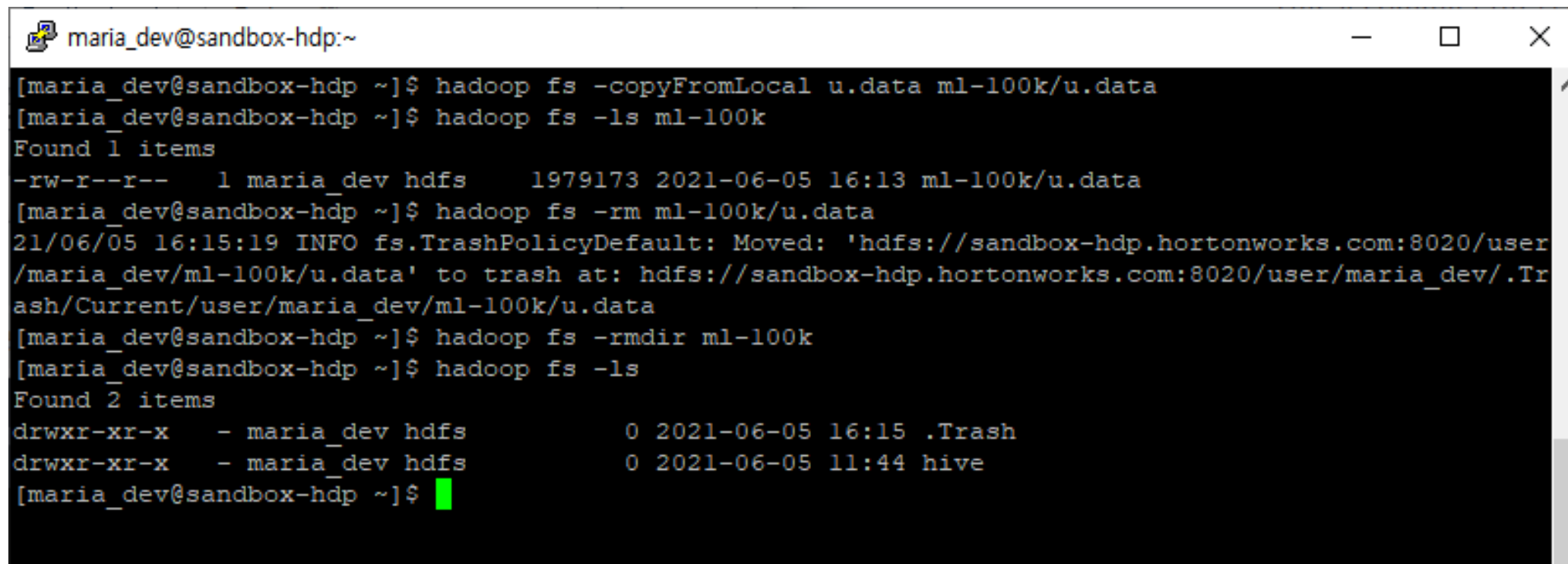
hadoop fs -copyFromLocal u.data ml-100k/u.data

hadoop fs -ls ml-100k

hadoop fs -rm ml-100k/u.data

hadoop fs -rmdir ml-100k

hadoop fs -ls



```

maria_dev@sandbox-hdp:~
[maria_dev@sandbox-hdp ~]$ hadoop fs -copyFromLocal u.data ml-100k/u.data
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls ml-100k
Found 1 items
-rw-r--r--  1 maria_dev hdfs      1979173 2021-06-05 16:13 ml-100k/u.data
[maria_dev@sandbox-hdp ~]$ hadoop fs -rm ml-100k/u.data
21/06/05 16:15:19 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonworks.com:8020/user/maria_dev/ml-100k/u.data' to trash at: hdfs://sandbox-hdp.hortonworks.com:8020/user/maria_dev/.Trash/Current/user/maria_dev/ml-100k/u.data
[maria_dev@sandbox-hdp ~]$ hadoop fs -rmdir ml-100k
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x  - maria_dev hdfs          0 2021-06-05 16:15 .Trash
drwxr-xr-x  - maria_dev hdfs          0 2021-06-05 11:44 hive
[maria_dev@sandbox-hdp ~]$
```

HDFS(터미널 환경)

hadoop fs

```

maria_dev@sandbox-hdp:~
[maria_dev@sandbox-hdp ~]$ hadoop fs
Usage: hadoop fs [generic options]
    [-appendToFile <localsrc> ... <dst>]
    [-cat [-ignoreCrc] <src> ...]
    [-checksum <src> ...]
    [-chgrp [-R] GROUP PATH...]
    [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
    [-chown [-R] [OWNER][:[GROUP]] PATH...]
    [-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
    [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
    [-count [-q] [-h] [-v] [-t [<storage type>]] [-u] <path> ...]
    [-cp [-f] [-p | -p[topax]] <src> ... <dst>]
    [-createSnapshot <snapshotDir> [<snapshotName>]]
    [-deleteSnapshot <snapshotDir> <snapshotName>]
    [-df [-h] [<path> ...]]
    [-du [-s] [-h] <path> ...]
    [-expunge]
    [-find <path> ... <expression> ...]
    [-get [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
    [-getfacl [-R] <path>]
    [-getfattr [-R] {-n name | -d} [-e en] <path>]
    [-getmerge [-nl] <src> <localdst>]
    [-help [cmd ...]]
    [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]
    [-mkdir [-p] <path> ...]
    [-moveFromLocal <localsrc> ... <dst>]
    [-moveToLocal <src> <localdst>]

```

MapReduce

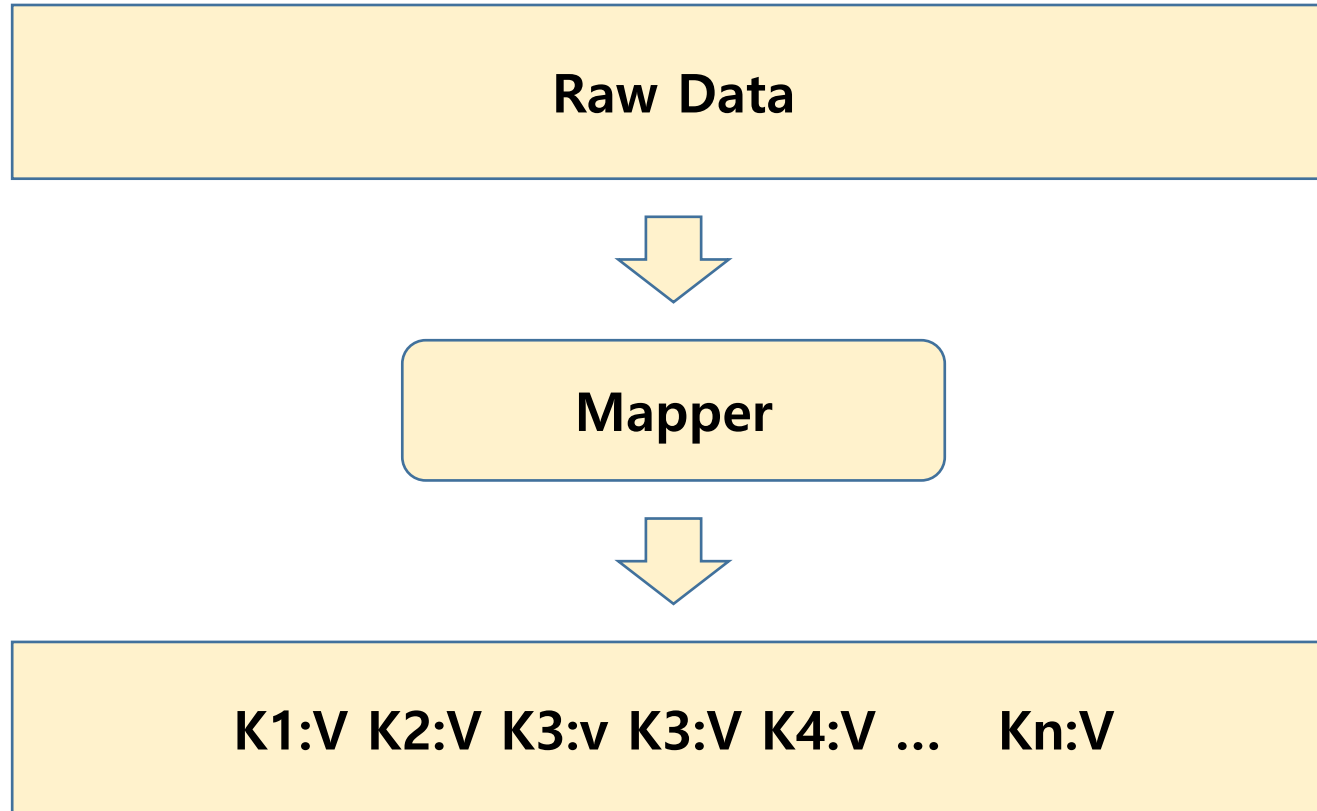
Mapper는 데이터를 변환(transform)하고 Reducer는 데이터를 집계(aggregate) 하는 것입니다.

MovieLens 데이터셋에서 각 사용자는 몇 개의 영화에 대해 평점을 매겼을까요?



Mapper Function

Mapper는 raw 소스 데이터를 key/value pair로 변환합니다.



Mapper on Movie Data

Extract and Organize What we care about

USER ID	MOVIE ID	RATING	TIMESTAMP
196	242	3	881250949
186	302	3	891717742
196	377	1	878887116
244	51	2	880606923
166	346	1	886397596
186	474	4	884182806
186	265	2	881171488



Mapper



196:242 186:302 196:377 244:51 166:346 186:274 186:265

Sort and Shuffle Mapper Data

MapReduce Sorts and Groups the Mapped Data

196:242 186:302 196:377 244:51 166:346 186:274 186:265



Shuffle & Sort



166:346 186:302,274,265 196:242,377 244:51

Reducer Process Each Key Value

The REDUCER Processes Each Key's Values

166:346 186:302,274,265 196:242,377 244:51



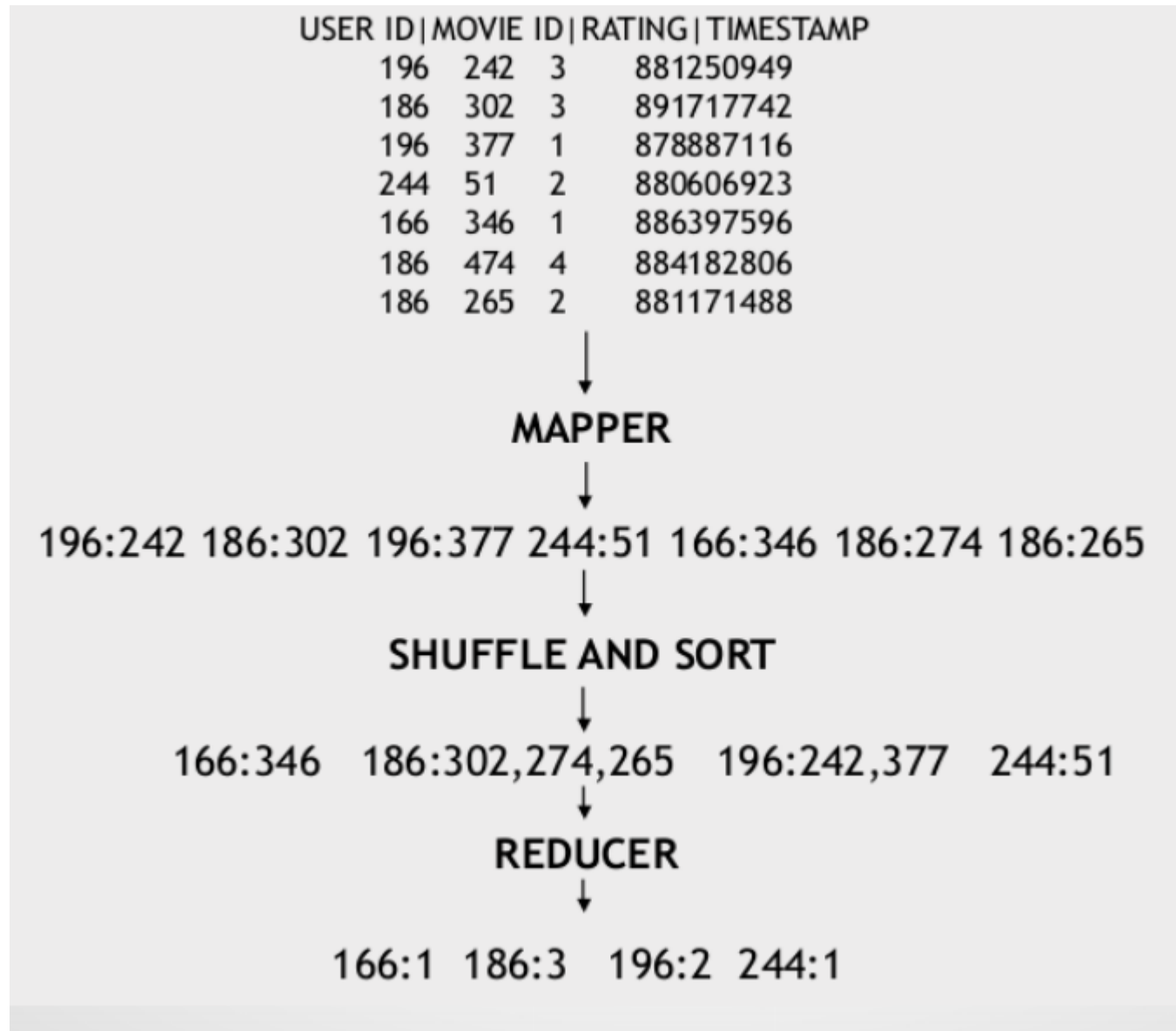
REDUCER

len(movies)

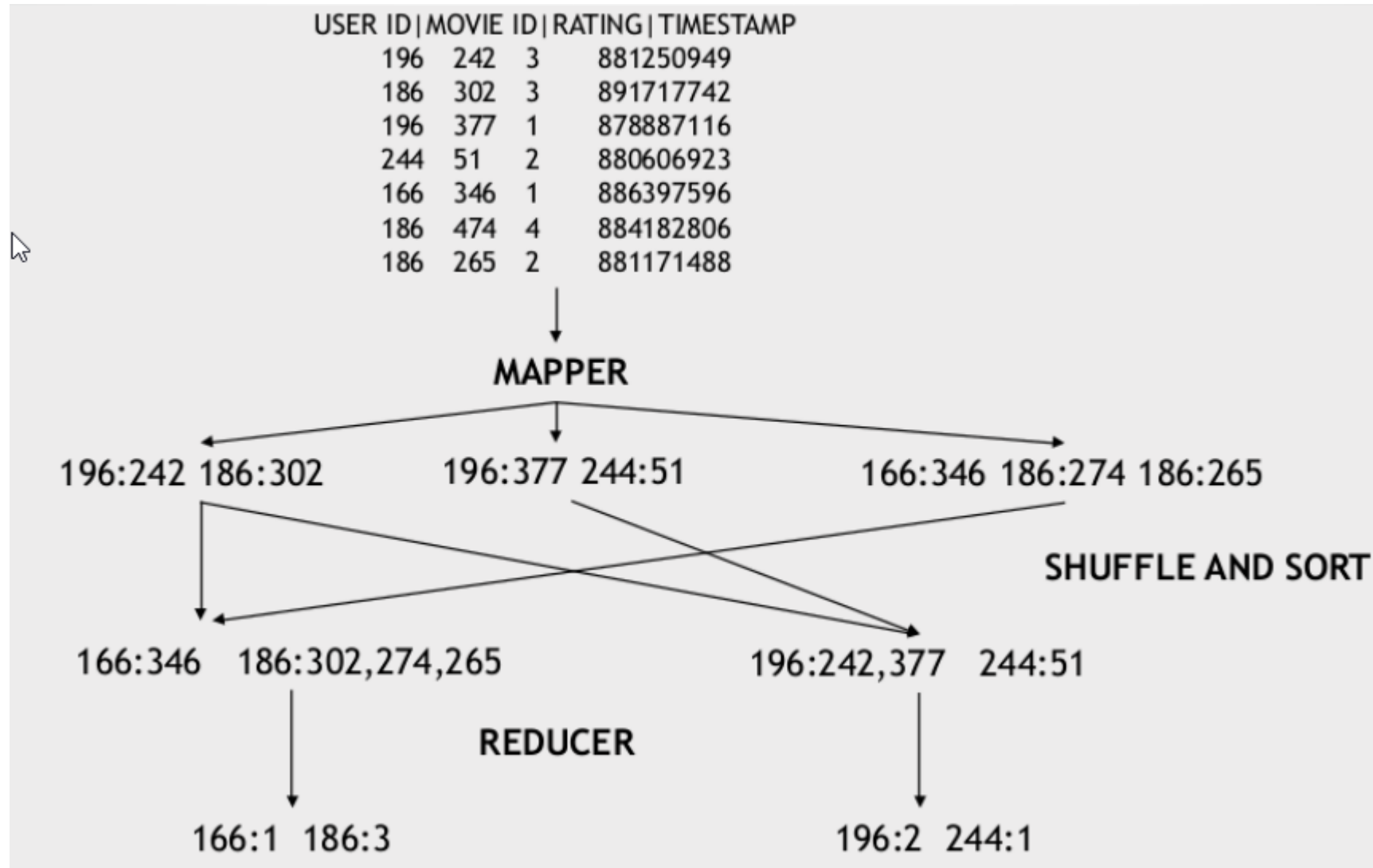


166:1 186:3 196:2 244:1

Mapper and Reduce



Mapper & Reducer in Cluster



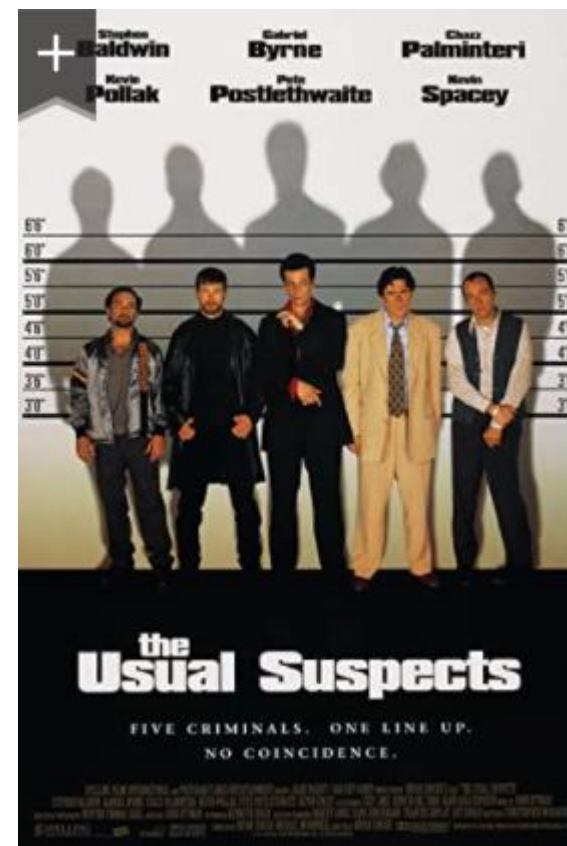
MapReduce Program (참고)

```
RatingsBreakdown.py
1  from mrjob.job import MRJob
2  from mrjob.step import MRStep
3
4  class RatingsBreakdown(MRJob):
5      def steps(self):
6          return [
7              MRStep(mapper=self.mapper_get_ratings,
8                    reducer=self.reducer_count_ratings)
9          ]
10
11     def mapper_get_ratings(self, _, line):
12         (userID, movieID, rating, timestamp) = line.split('\t')
13         yield rating, 1
14
15     def reducer_count_ratings(self, key, values):
16         yield key, sum(values)
17
18 if __name__ == '__main__':
19     RatingsBreakdown.run()
```

■ 실행방법 참고

<http://www.nitesh-research.com/wp-content/uploads/nitesh-hadoop-12-09-2020.pdf>

Find the oldest 5-star movies



The screenshot shows the Ambari Sandbox web interface in a Chrome browser. The address bar displays `localhost:8080/#/main/views/PIG/1.0.0/PIG_INSTANCE`. The top navigation bar includes the Ambari logo, 'Sandbox' tab, '0 ops' and '0 alerts' indicators, and links to 'Dashboard', 'Services', 'Hosts', 'Alerts', and 'Admin'. A user profile dropdown for 'maria_dev' is visible on the right.

On the left sidebar, there are three menu items: 'Scripts' (with a code icon), 'UDFs' (with a plug icon), and 'History' (with a clock icon). The 'Scripts' menu item is selected.

The main content area is titled 'Scripts'. It features a table with the following headers: 'Name', 'Last Executed', 'Last Results', and 'Actions'. Below the table, a light blue message box states: 'No pig scripts have been created. To get started, click New Script.'

A dropdown menu is open from the top right, showing options: 'Files View', 'Hive View', 'Hive View 2.0', 'Pig View' (highlighted with a red rectangle and a mouse cursor), 'tez view', and 'Workflow Manager'.

The screenshot shows the Ambari Sandbox web interface in a Chrome browser. The address bar shows the URL `localhost:8080/#/main/views/PIG/1.0.0/PIG_INSTANCE`. The top navigation bar includes the Ambari logo, 'Sandbox' tab, '0 ops' and '0 alerts' indicators, and links to 'Dashboard', 'Services', 'Hosts', 'Alerts', and 'Admin'. A user profile dropdown for 'maria_dev' is visible on the right.

On the left sidebar, there are three menu items: 'Scripts' (with a code icon), 'UDFs' (with a plug icon), and 'History' (with a clock icon). The 'Scripts' menu is currently selected.

The main content area is titled 'Scripts'. It features a table with the following headers: 'Name', 'Last Executed', 'Last Results', and 'Actions'. Below the table, a light blue message box states: 'No pig scripts have been created. To get started, click New Script.'

A dropdown menu is open from the 'Actions' column of the table. The menu items are: 'Files View', 'Hive View', 'Hive View 2.0', 'Pig View' (highlighted with a red box and a mouse cursor), 'tez view', and 'Workflow Manager'.

Ambari - Sandbox - Chrome

localhost:8080/#/main/views/PIG/1.0.0/PIG_INSTANCE

Ambari Sandbox 0 ops 0 alerts

Dashboard Services Hosts Alerts Admin

maria_dev

Scripts

UDFs

History

New Script

Name

Oldest five-star movie **Oldest five-star movie**

Script HDFS Location (optional)

/hdfs/path/to/pig/script

Leave empty to create file automatically.

Cancel Create

+ New Script

Actions

History Copy Delete

Show: 10 1 - 1 of 1

Pig

Ambari - Sandbox - Chrome

localhost:8080/#/main/views/PIG/1.0.0/PIG_INSTANCE

Ambari Sandbox 0 ops 0 alerts Dashboard Services Hosts Alerts Admin maria_dev

Script History

Oldest five-star movie

Save Copy Delete

Execute Execute

PIG helper UDF helper /user/maria_dev/pig/scripts/oldest_fivestar_movie-2021-06-06_04-27.pig

```
1 ratings = LOAD '/user/maria_dev/ml-100k/u.data' AS (userID:int, movieID:int, rating:int, ratingTime:int);
2 metadata = LOAD '/user/maria_dev/ml-100k/u.item' USING PigStorage('|')
3           AS (movieID:int, movieTitle:chararray, releaseDate:chararray, videoRealese:chararray, imdblink:chararray);
4
5 nameLookup = FOREACH metadata GENERATE movieID, movieTitle,
6             ToUnixTime(ToDate(releaseDate, 'dd-MMM-yyyy')) AS releaseTime;
7
8 ratingsByMovie = GROUP ratings BY movieID;
9 avgRatings = FOREACH ratingsByMovie GENERATE group as movieID, AVG(ratings.rating) as avgRating;
10 fiveStarMovies = FILTER avgRatings BY avgRating > 4.0;
11 fiveStarsWithData = JOIN fiveStarMovies BY movieID, nameLookup BY movieID;
12 oldestFiveStarMovies = ORDER fiveStarsWithData BY nameLookup::releaseTime;
13 DUMP oldestFiveStarMovies;
14
```

1

pig.txt 내용 복사

2

Ambari - Sandbox - Chrome

localhost:8080/#/main/views/PIG/1.0.0/PIG_INSTANCE

Oldest five-star movie

Save
Copy
Delete

Oldest five-star movie - COMPLETED

Job ID job_1622947887856_0007

Started 2021-06-06 13:29

▼ Results [Download](#)

```
(493,4.15,493,Thin Man, The (1934),-1136073600)
(604,4.012345679012346,604,It Happened One Night (1934),-1136073600)
(615,4.0508474576271185,615,39 Steps, The (1935),-1104537600)
(1203,4.0476190476190474,1203,Top Hat (1935),-1104537600)
(613,4.037037037037037,613,My Man Godfrey (1936),-1073001600)
(633,4.057971014492754,633,Christmas Carol, A (1938),-1009843200)
(132,4.0772357723577235,132,Wizard of Oz, The (1939),-978307200)
(1122,5.0,1122,They Made Me a Criminal (1939),-978307200)
(136,4.123809523809523,136,Mr. Smith Goes to Washington (1939),-978307200)
(478,4.115384615384615,478,Philadelphia Story, The (1940),-946771200)
(524,4.021739130434782,524,Great Dictator, The (1940),-946771200)
(484,4.2101449275362315,484,Maltese Falcon, The (1941),-915148800)
(134,4.292929292929293,134,Citizen Kane (1941),-915148800)
(483,4.45679012345679,483,Casablanca (1942),-883612800)
(659,4.078260869565217,659,Arsenic and Old Lace (1944),-820540800)
```

Thank you