

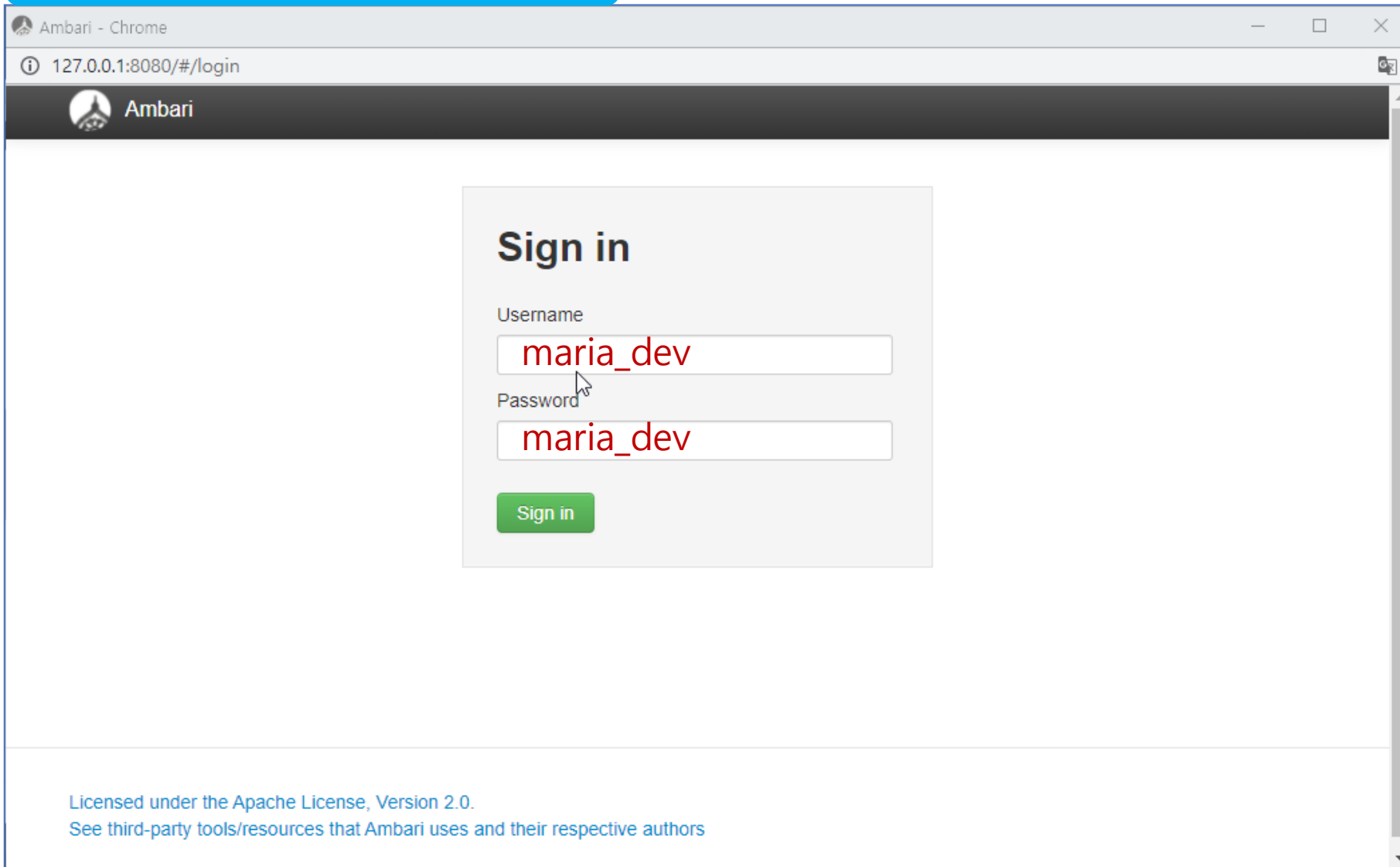
Hadoop 에코시스템



Ambari

Ambari 접속

<http://127.0.0.1:8080/>

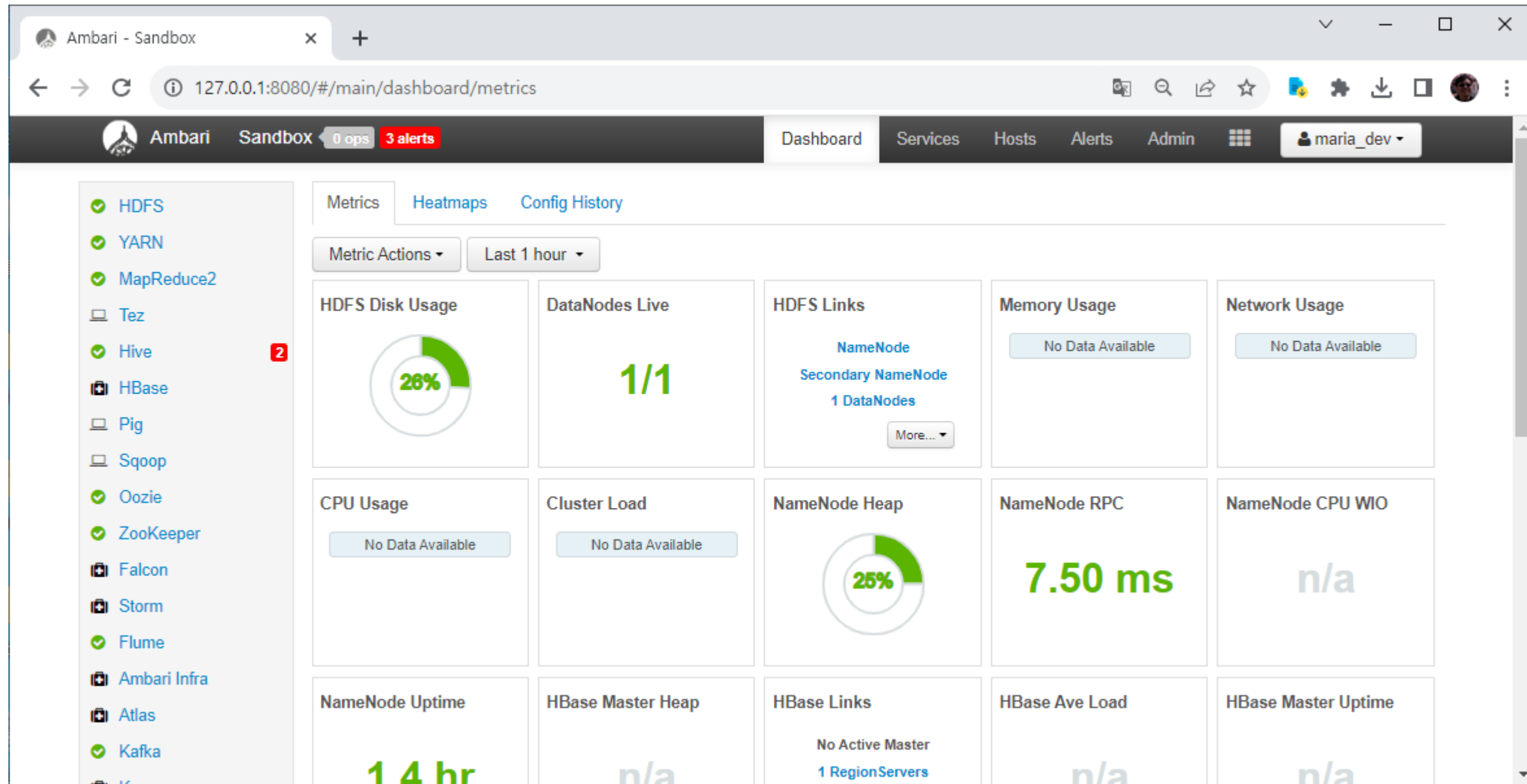


The screenshot shows a web browser window titled "Ambari - Chrome" with the address bar displaying "127.0.0.1:8080/#/login". The page features a dark header with the Ambari logo and name. The main content area contains a "Sign in" form with the following elements:

- 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".

Ambari Dashboard



Ambari Service

The screenshot shows the Ambari web interface for a 'Sandbox' environment. The browser address bar shows 'localhost:8080/#/main/services/HDFS/summary'. The top navigation bar includes links for Dashboard, Services, Hosts, Alerts, and Admin. The left sidebar lists various services: HDFS, YARN, MapReduce2, Tez, Hive, HBase, Pig, Sqoop, Oozie, ZooKeeper, Falcon, Storm, Flume, Ambari Infra, Atlas, Kafka, and Knox. The main content area is titled 'Summary' and shows the status of the HDFS service. It indicates that the NameNode is started, SNameNode is started, and DataNodes are 1/1 started. It also shows disk usage and other metrics. The 'Metrics' section at the bottom shows five metrics: NameNode GC count, NameNode GC time, NN Connection Load, NameNode Heap, and NameNode Host Load, all of which are currently showing 'No Data No available data for the time period.'

Ambari - Sandbox

localhost:8080/#/main/services/HDFS/summary

Ambari Sandbox 0 ops 0 alerts

Dashboard Services Hosts Alerts Admin

Summary Heatmaps Configs Quick Links Service Actions

Summary No alerts

[NameNode](#) Started No alerts

[SNameNode](#) Started No alerts

[DataNodes](#) 1/1 Started

DataNodes Status 1 live / 0 dead / 0 decommissioning

[JournalNodes](#) 1/1 JournalNodes Live

[NFSGateways](#) 0/0 Started

NameNode Uptime 24.76 mins

NameNode Heap 27.7 MB / 240.0 MB (11.5% used)

Disk Usage (DFS Used) 2.1 GB / 106.0 GB (1.95%)

Disk Usage (Non DFS Used) 25.5 GB / 106.0 GB (24.08%)

Disk Remaining 78.4 GB / 106.0 GB (73.97%)

Blocks (total) 1121

Block Errors 0 corrupt replica / 0 missing / 0 under replicated

Total Files + Directories 1352

Upgrade Status No pending upgrade

Safe Mode Status Not in safe mode

Metrics Last 1 hour

NameNode GC count	NameNode GC time	NN Connection Load	NameNode Heap	NameNode Host Load
No Data No available data for the time period.	No Data No available data for the time period.	No Data No available data for the time period.	No Data No available data for the time period.	No Data No available data for the time period.

Ambari Host

Ambari - Sandbox

localhost:8080/#/main/hosts

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

Filter by host and component attributes or search by keyword ...

<input type="checkbox"/>	Name	IP Address	Rack	Cores	RAM	Disk Usage	Load Avg	Versions	Components
<input type="checkbox"/>	sandbox-hdp.hortonworks....	172.18.0.2	/default-rack	4 (4)	7.79GB			HDP-2.6.5.0	56 Components

Show: 10 1 - 1 of 1

Licensed under the Apache License, Version 2.0.
See [third-party tools/resources that Ambari uses and their respective authors](#)

Ambari Alerts

The screenshot shows the Ambari Alerts page in a web browser. The browser's address bar displays the URL `127.0.0.1:8080/#/main/alerts`. The Ambari interface includes a top navigation bar with the Ambari logo, the name 'Sandbox', and a red badge indicating '3 alerts'. The main navigation menu contains links for Dashboard, Services, Hosts, Alerts (which is the active tab), and Admin. A user profile dropdown for 'maria_dev' is visible on the right.

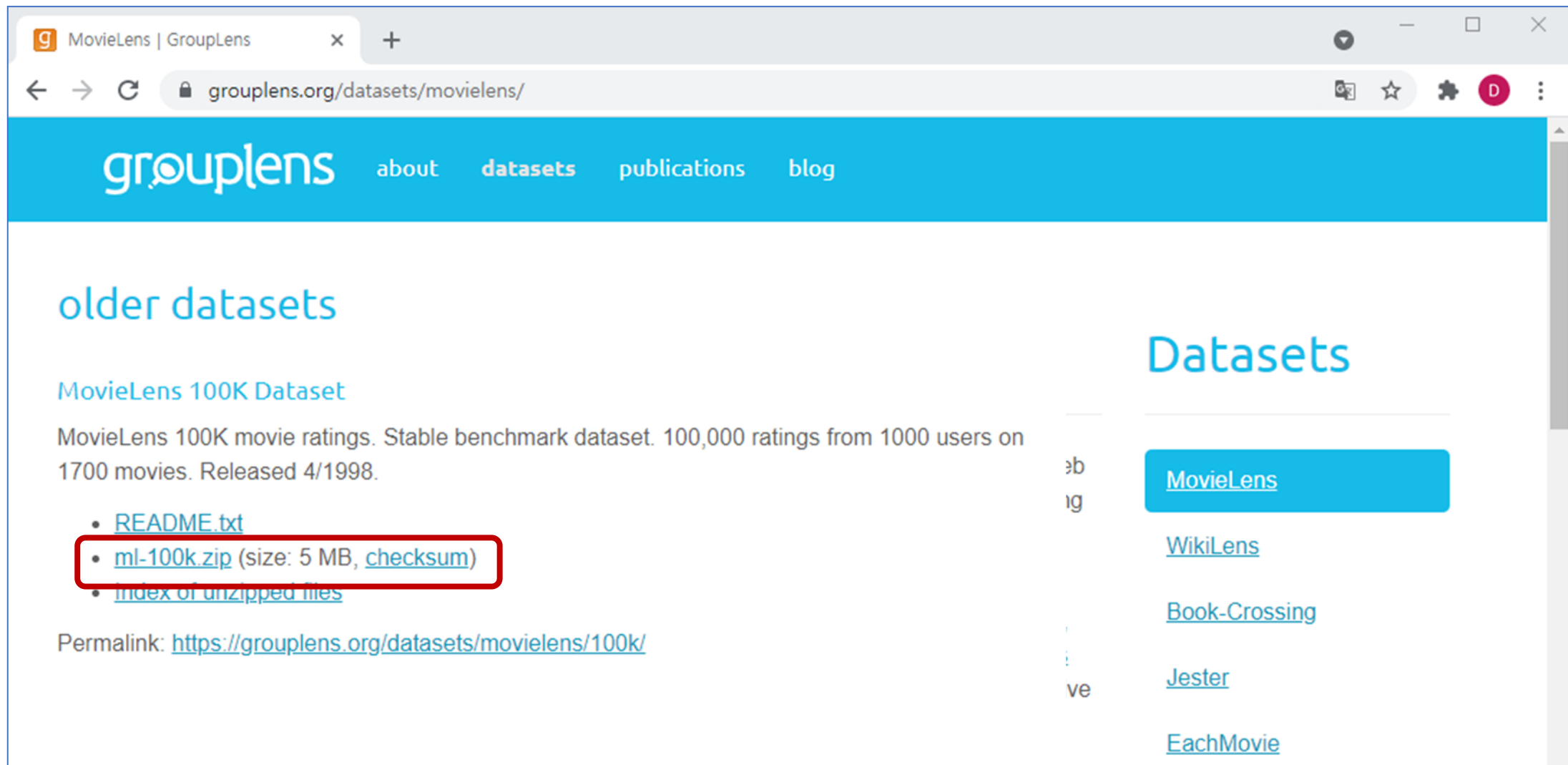
Below the navigation bar, there is a filter section with a dropdown menu set to 'Groups: All (85)'. The main content area is a table of alert definitions with the following columns: Alert Definition Name, Status, Service, Last Status Changed, and State.

Alert Definition Name	Status	Service	Last Status Changed	State
Falcon Server Web UI	CRIT	Falcon	5 years ago	Enabled
Falcon Server Process	CRIT	Falcon	5 years ago	Enabled
Metadata Server Web UI	CRIT	Atlas	5 years ago	Enabled
HBase Master Process	CRIT	HBase	5 years ago	Enabled
HBase RegionServer Process	CRIT	HBase	5 years ago	Enabled
Knox Gateway Process	CRIT	Knox	5 years ago	Enabled
Hive Metastore Process	CRIT	Hive	2 hours ago	Enabled
HiveServer2 Process	CRIT	Hive	2 hours ago	Enabled
Infra Solr Web UI	CRIT	Ambari Infra	5 years ago	Enabled
Storm Web UI	CRIT	Storm	5 years ago	Enabled

HDFS 실습

MovieLens 데이터 다운로드

<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' entry 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](#)

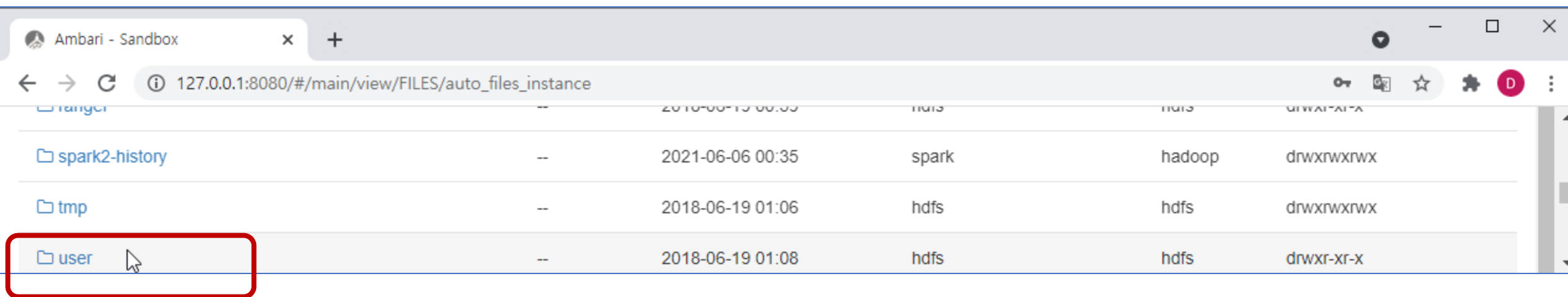
HDFS

The screenshot displays the Ambari Sandbox interface for the HDFS service. The left sidebar (labeled 1) lists various services, with HDFS selected. The top navigation bar (labeled 2) shows the user profile dropdown menu open, displaying options such as Files View, Hive View, Hive View 2.0, Pig View, Tez View, and Workflow Manager. The main content area shows the HDFS Summary page, including details about NameNode, SNameNode, DataNodes, JournalNodes, NFS Gateways, NameNode Uptime, NameNode Heap, Disk Usage (DFS Used), and Disk Usage (Non DFS Used).

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
NFS Gateways	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%)			

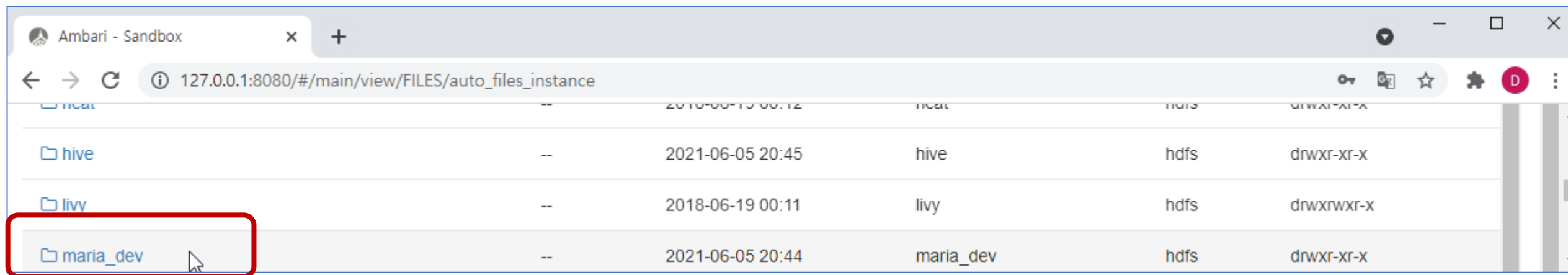
HDFS



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



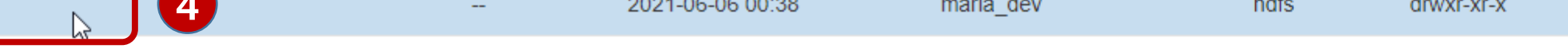
Ambari - Sandbox

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



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 main content area shows a list of files and directories. The 'ml-100k' directory is highlighted with a red box, and a red circle with the number '4' is placed next to it, indicating the next step in the process.

File/Directory	Permissions	Created	Owner	Group	Mode
hive	--	2021-06-05 20:44	maria_dev	hdfs	drwxr-xr-x
ml-100k	--	2021-06-06 00:38	maria_dev	hdfs	drwxr-xr-x

HDFS

The screenshot shows the Ambari Sandbox web interface for managing HDFS. The browser address bar shows 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 for `maria_dev`. The main content area displays the file system path `/ > user > maria_dev > ml-100k` and indicates that there are 0 files or folders in the current directory. A red circle with the number 1 highlights the **Upload** button in the top right corner of the file view area. An upload dialog box is open in the center, titled **Upload file to /user/maria_dev/ml-100k**. Inside the dialog, a red circle with the number 2 highlights a button with a cloud upload icon and the text **u.data** and **u.item**. Below this button, the text **Drag file to upload or click to browse** is displayed. At the bottom right of the dialog is a **Cancel** button. A red note at the bottom of the dialog states *Currently supports single file upload*.

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 file and 0 folders are 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 preview shows a table of data with four columns. At the bottom of the modal, there are 'Cancel' and 'Download' buttons.

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

Ambari - Sandbox

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

Ambari Sandbox 0 ops 0 alerts

Dashboard Services Hosts Alerts Admin

0 Files, 1 Folders selected

+ Select All New Folder Upload

Open Rename Permissions **Delete** Copy Move Download concatenate

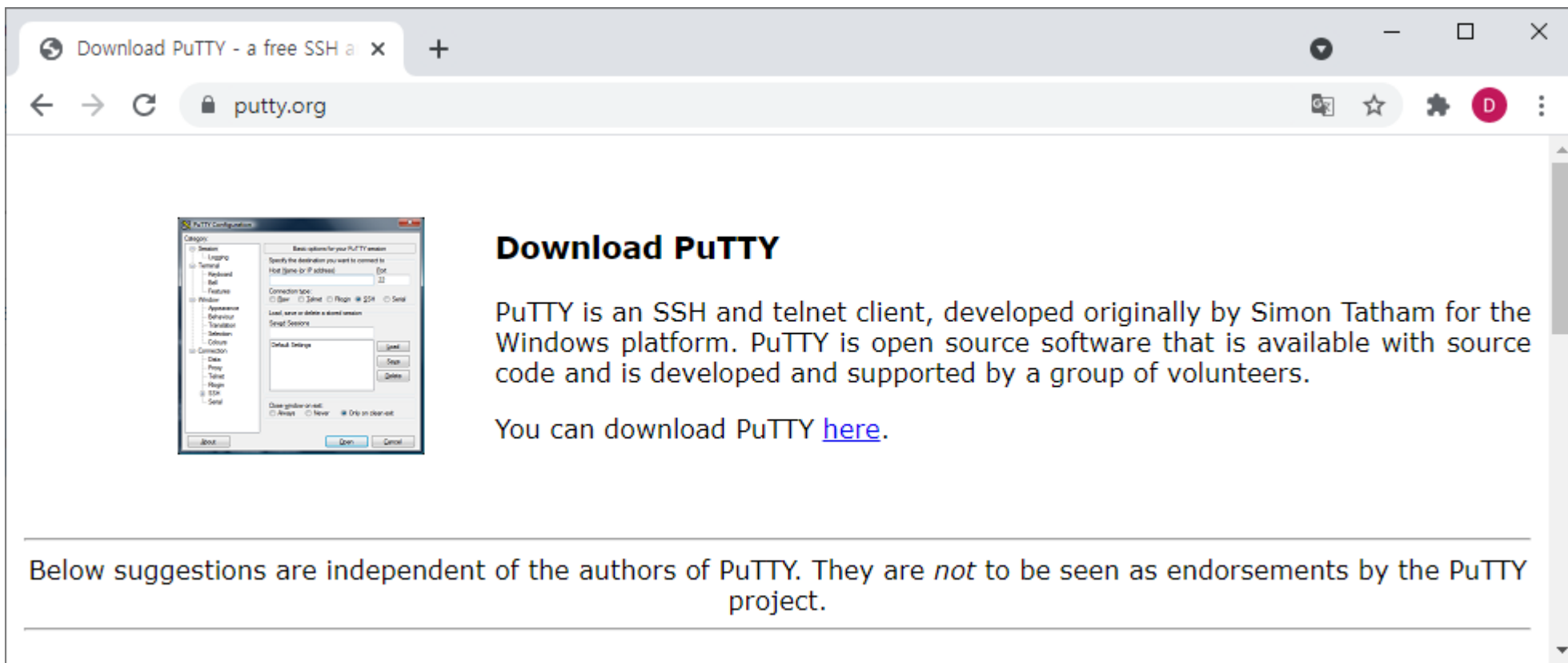
Search in current directory...

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.

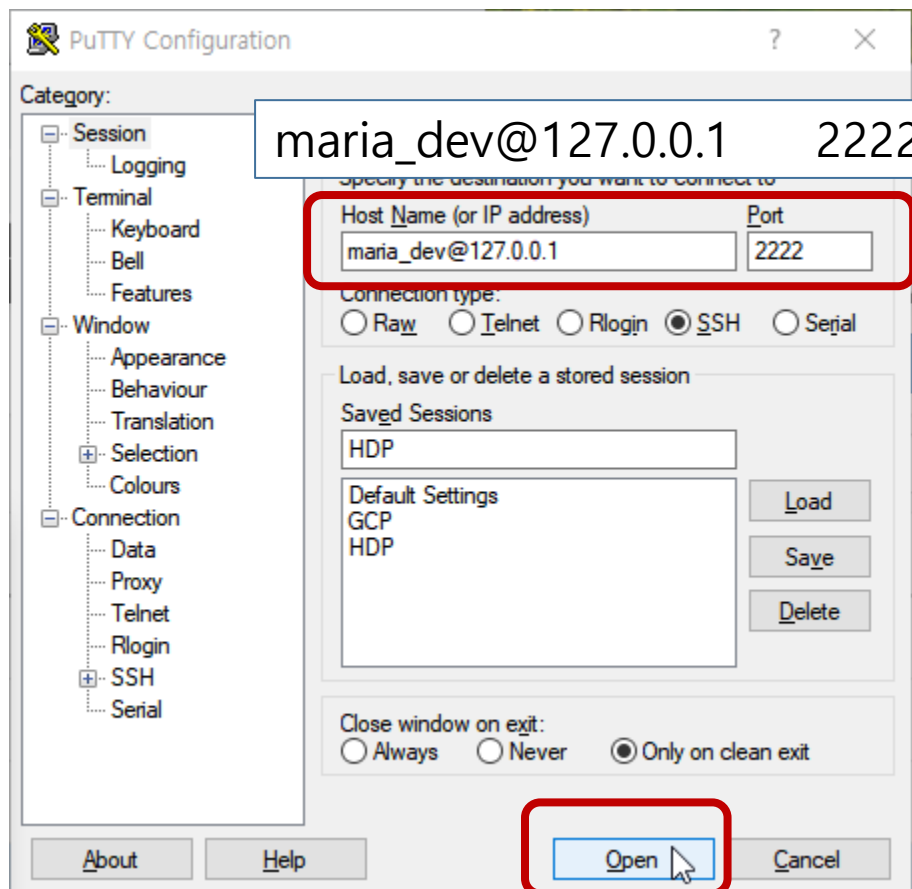
You can download PuTTY [here](#).

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

HDFS(터미널 환경)

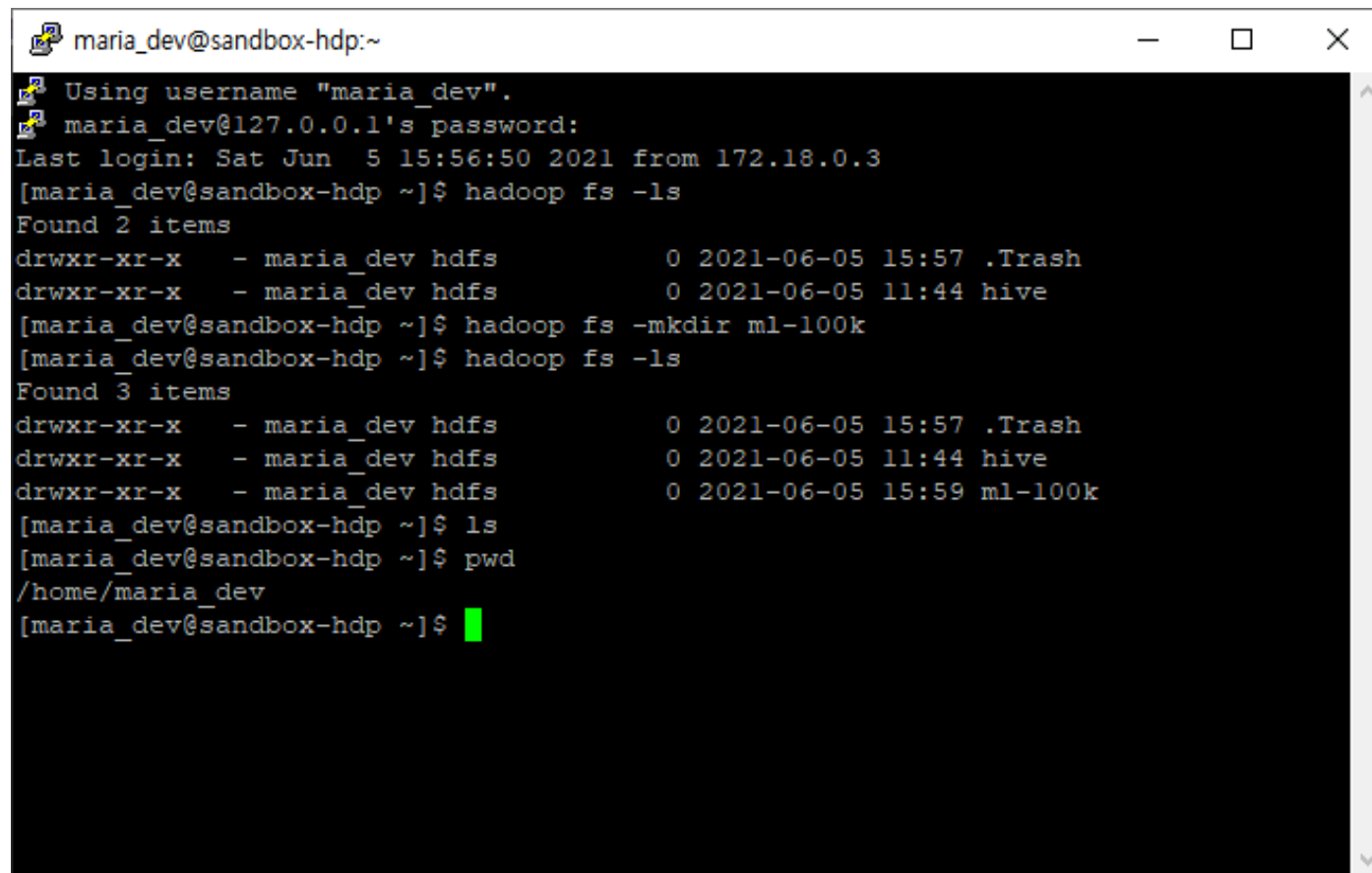


putty 실행



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)

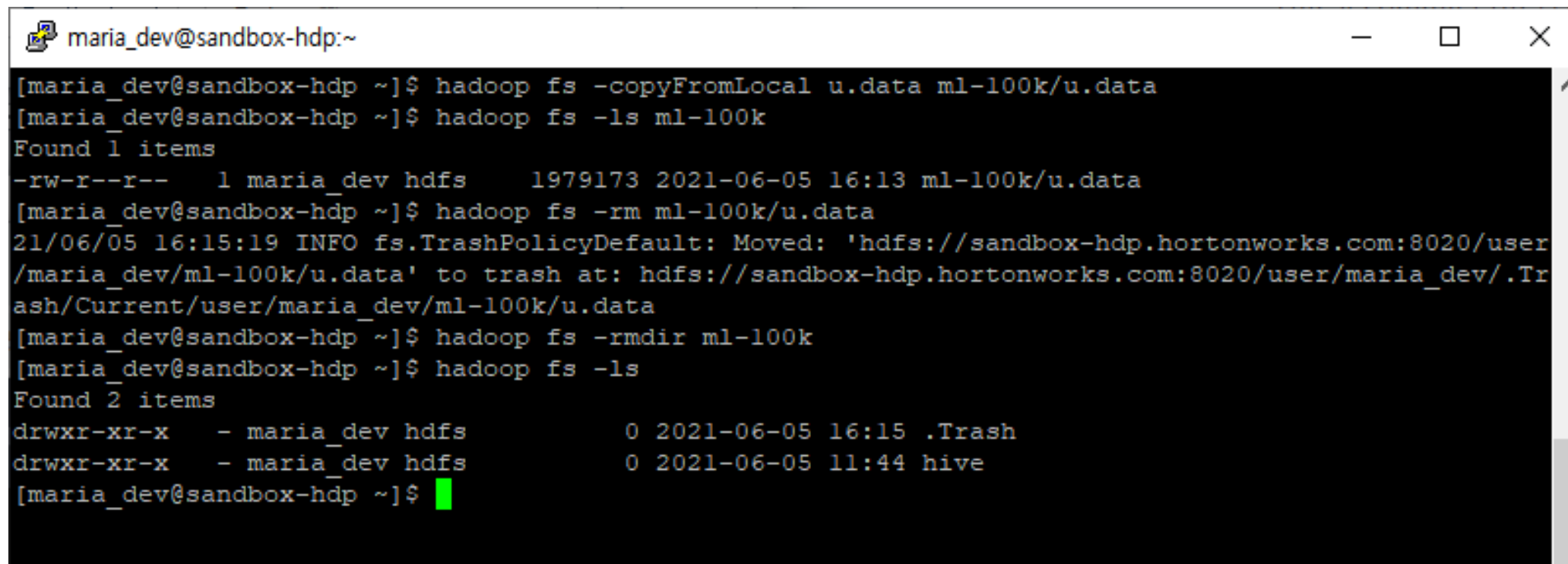
hadoop fs -copyFromLocal u.data ml-100k/u.data

hadoop fs -ls ml-100k

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

hadoop fs -rmdir ml-100k

hadoop fs -ls



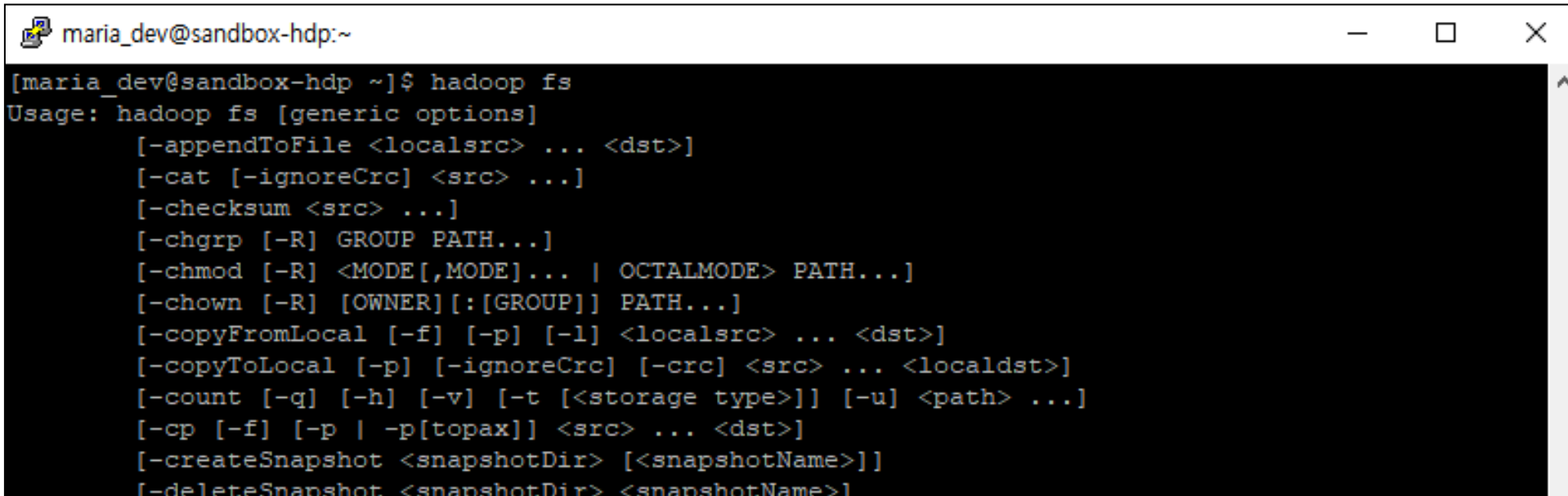
```
maria_dev@sandbox-hdp:~  
[maria_dev@sandbox-hdp ~]$ hadoop fs -copyFromLocal u.data ml-100k/u.data  
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls ml-100k  
Found 1 items  
-rw-r--r--  1 maria_dev hdfs      1979173 2021-06-05 16:13 ml-100k/u.data  
[maria_dev@sandbox-hdp ~]$ hadoop fs -rm ml-100k/u.data  
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)

■ HDFS Commands Guide

<https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html>

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>]
```

MapReduce

MapReduce

Mapper는 데이터를 변환(Transform)하고, Reducer는 데이터를 집계(Aggregate) 하는 것입니다.



How many of each movie rating exist?



MapReduce

Map each input line to (rating, 1)
Redude each rating with the sum of all the 1's

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



```
def mapper_get_ratings(self, _, line):  
    (userID, movieID, rating, timestamp) = line.split('\t')  
    yield rating, 1
```

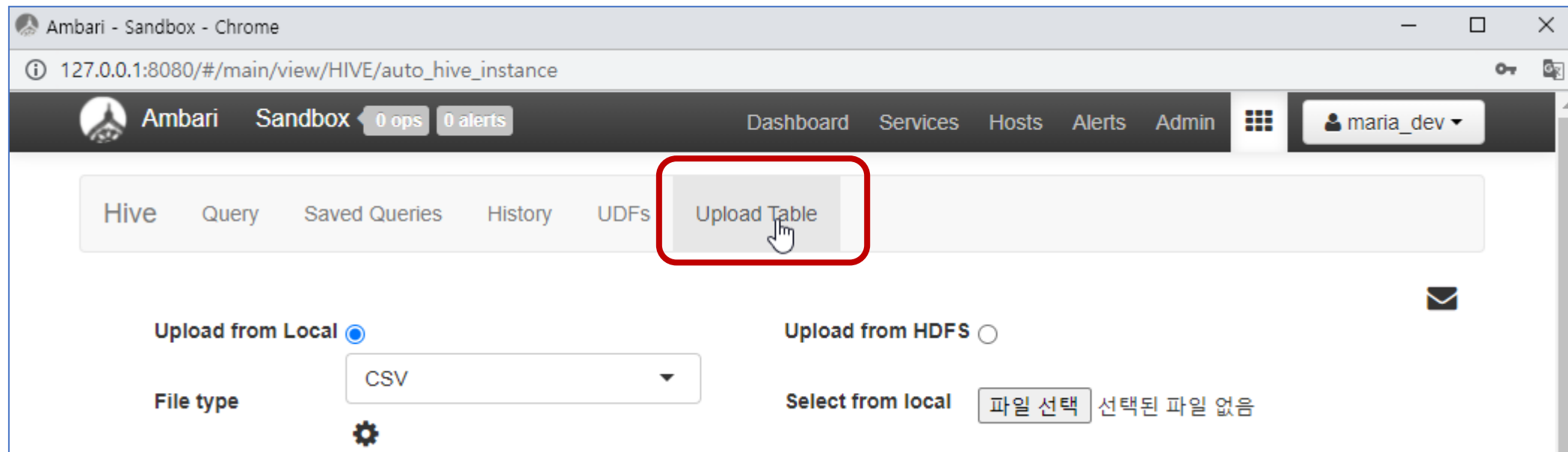
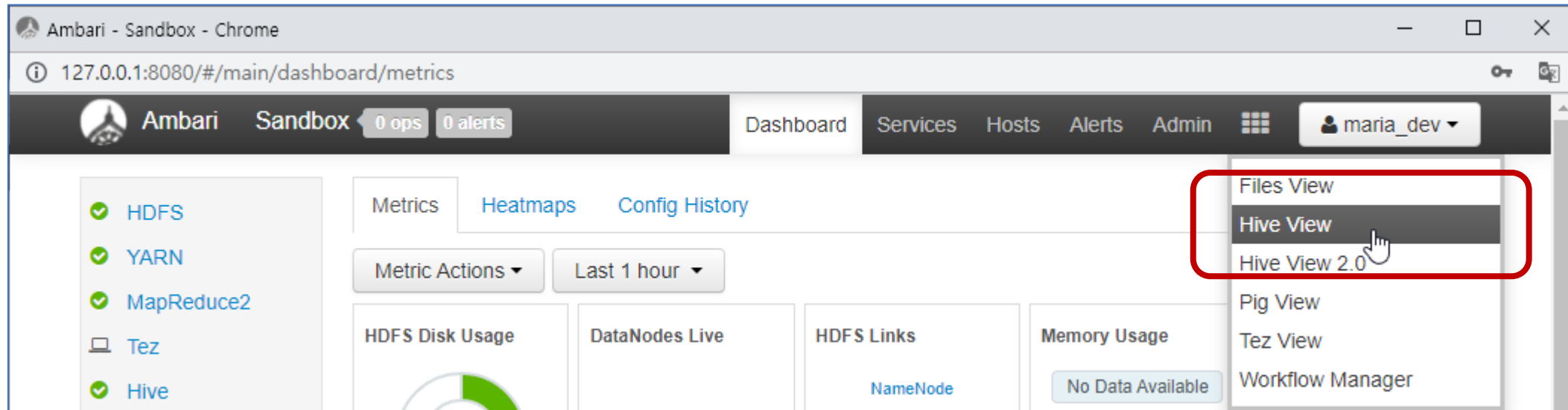
```
def reducer_count_ratings(self, key, values):  
    yield key, sum(values)
```

MapReduce

```
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()
```

HIVE 실습

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 1 2 3 4 6

Database default

Stored as ORC

Field Delimiter: | **TAB**

Escape Character: 2

Quote Character:

Is first row header?

Select from local file 3 파일 선택 u.data **u.data**

Table name 4 ratings **ratings**

Contains endlines? ☐

5 user_id **user_id** movie_id **movie_id** rating **rating** rating_time **rating_time**

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 Sandbox web interface. The browser address bar displays 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 highlighted with a red box. On the right side of this bar, a dropdown menu is open, listing 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 databases with 'default' selected, containing a table named 'movie_names' with columns: movie_id (INT), name (STRING), column3 (STRING), column4 (STRING), and column5 (STRING). The right panel, titled 'Query Editor', shows a 'Worksheet' with the text '1 |'.

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

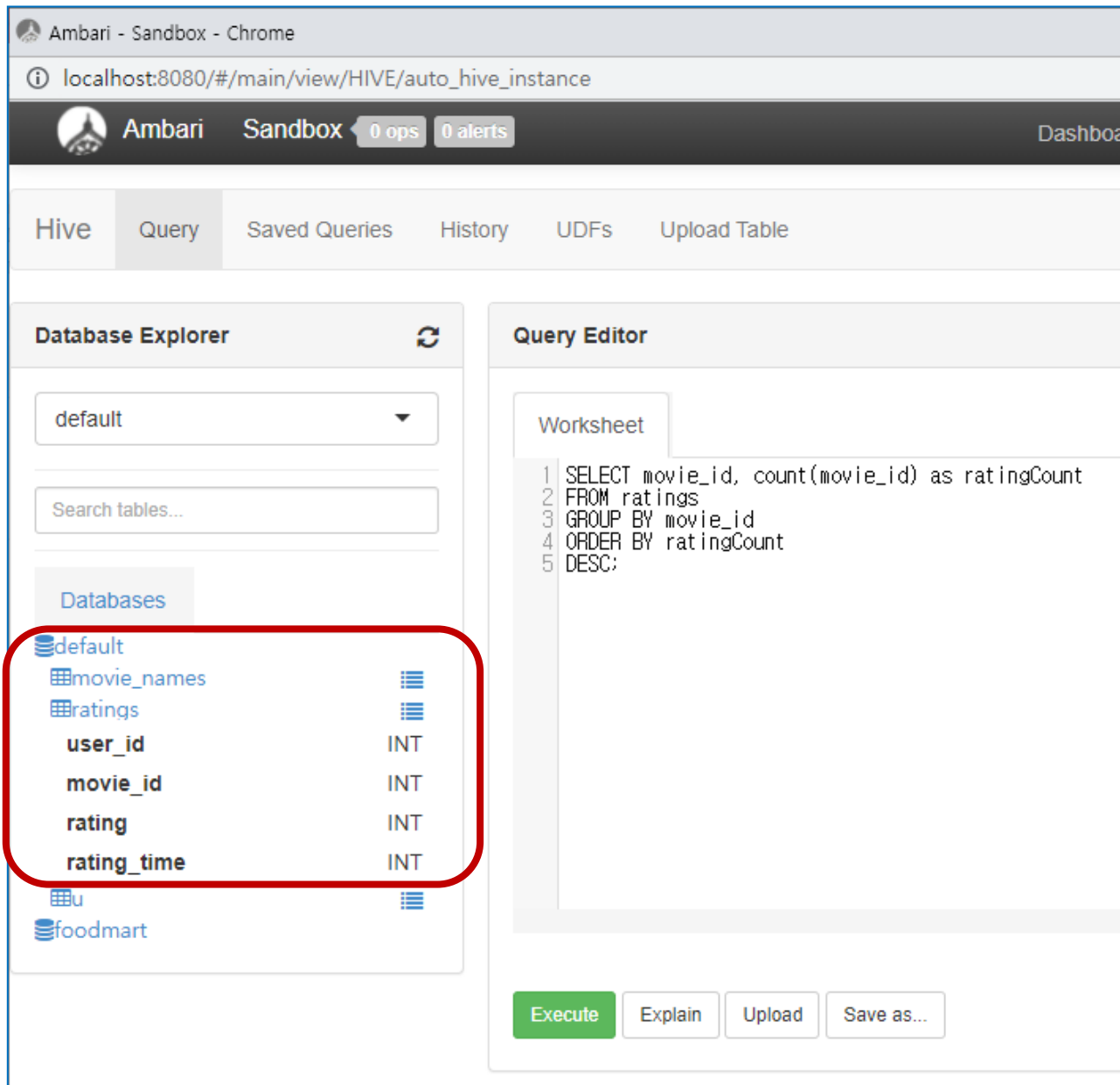
1 |

SQL

TEZ

1

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

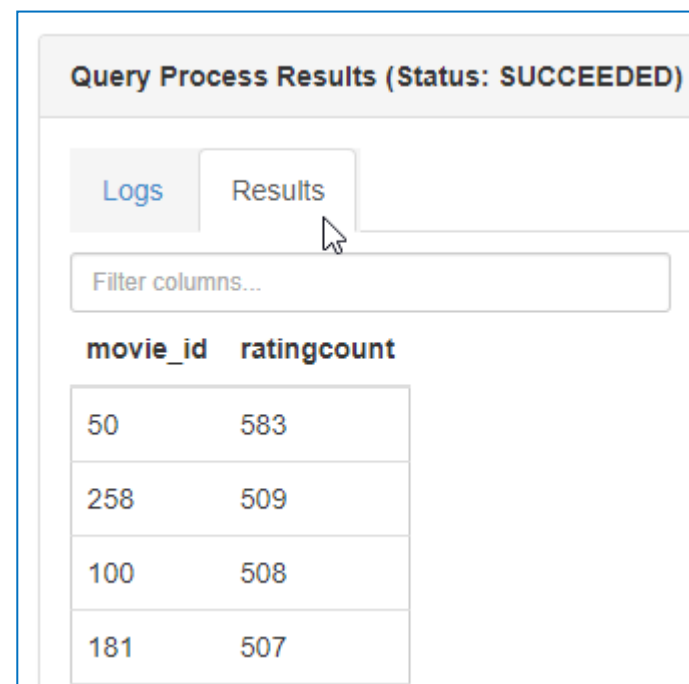
Query Editor

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;
```



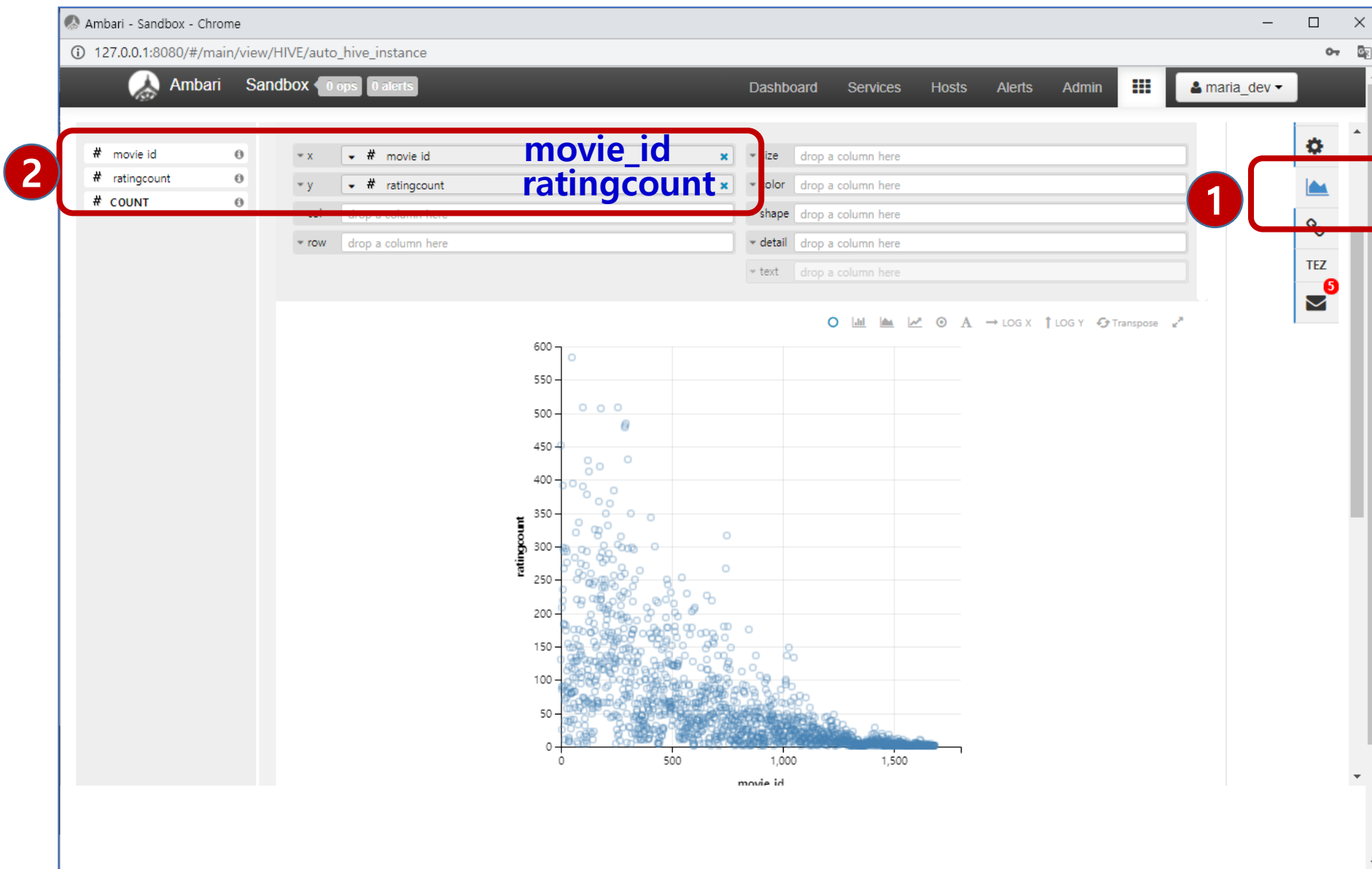
Query Process Results (Status: SUCCEEDED)

Logs Results

Filter columns...

movie_id	ratingcount
50	583
258	509
100	508
181	507

Hive



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...

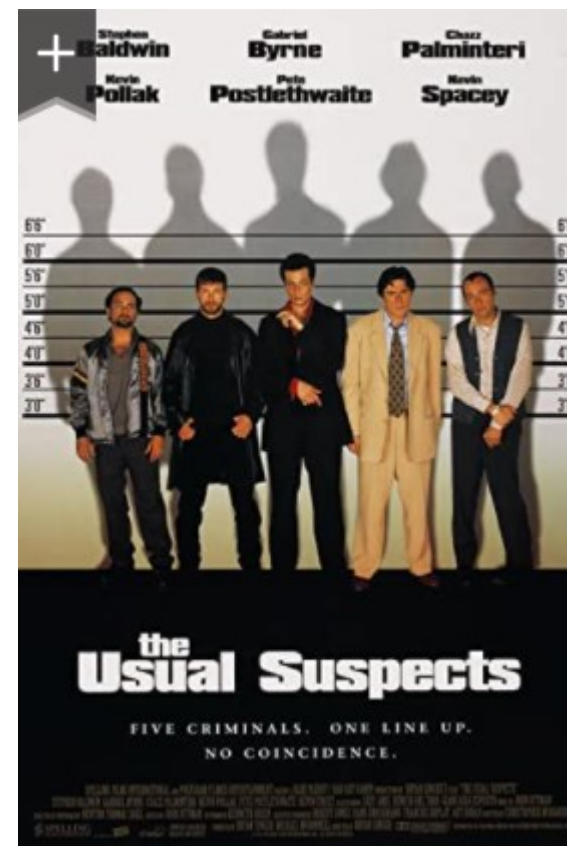
previous next

name

Star Wars (1977)

Pig 실습

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 links for Dashboard, Services, Hosts, Alerts, and Admin, along with a user profile dropdown for `maria_dev`. On the left sidebar, there are icons and labels for Scripts, UDFs, and History. The main content area is titled "Scripts" and features a table with columns: Name, Last Executed, Last Results, and Actions. The table is currently empty, with a message stating: "No pig scripts have been created. To get started, click New Script." A context menu is open over the Actions column, listing options: 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

Scripts

Name	Last Executed	Last Results	Actions
No pig scripts have been created. To get started, click New Script.			

Files View

Hive View

Hive View 2.0

Pig View

Tez View

Workflow Manager

localhost:8080/#

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 links for Dashboard, Services, Hosts, Alerts, and Admin, along with a user profile dropdown for `maria_dev`. On the left, a sidebar contains icons for Scripts, UDFs, and History. The main content area is titled "Scripts" and features a table with columns: Name, Last Executed, Last Results, and Actions. The table is currently empty, with a message stating: "No pig scripts have been created. To get started, click New Script." A context menu is open over the Actions column, listing options: 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

Scripts

Name	Last Executed	Last Results	Actions
No pig scripts have been created. To get started, click New Script.			

Files View

Hive View

Hive View 2.0

Pig View

Tez View

Workflow Manager

localhost:8080/#

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

1 + 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

Ambari - Sandbox - Chrome

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

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)
```

HBase 실습

HBase 실습

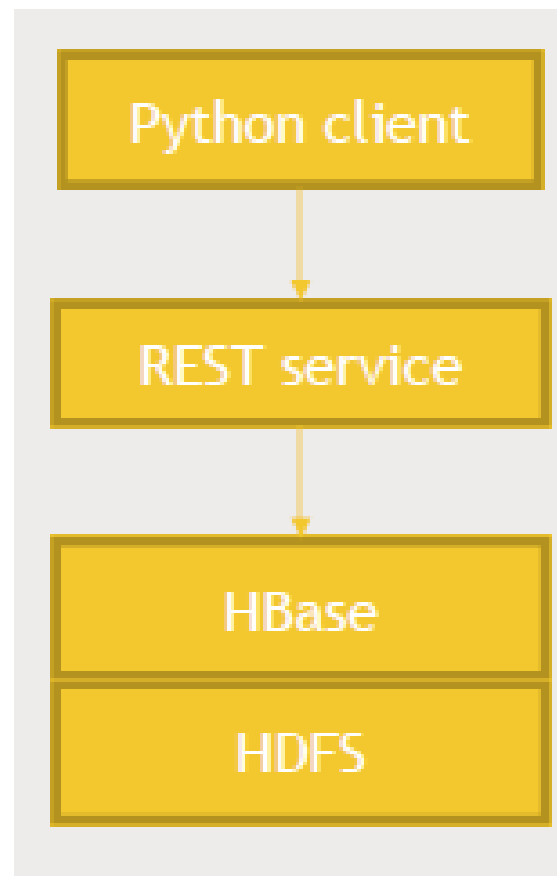
■ 실습내용

사용자별 영화 등급에 대한 Hbase 테이블 생성

사용자별 영화 등급 개수 집계



■ 구조

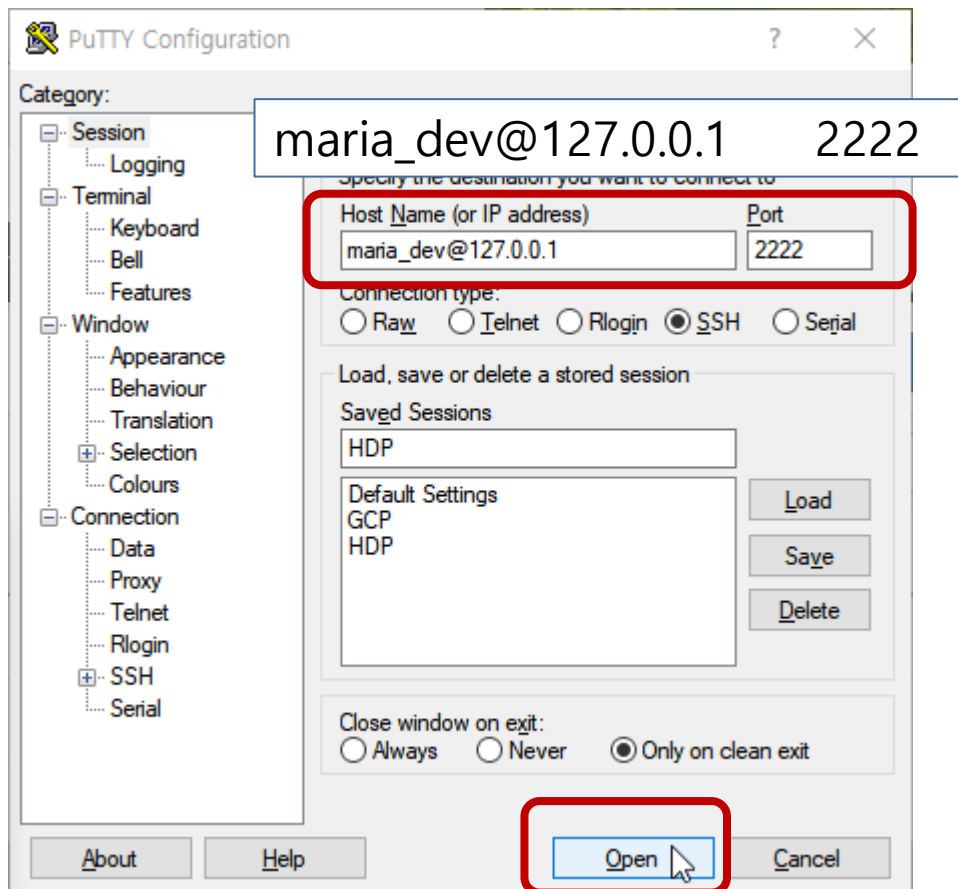


HBase 실습

■ HDP 접속



putty 실행



■ HBase 서버 실행

- `su root #root` 초기 비밀번호는 `hadoop`
- `/usr/hdp/current/hbase-master/bin/hbase-daemon.sh start rest -p 8000`

HBase 실습

■ starbase 패키지를 설치

pip install starbase

■ HBaseExamples.py

```
from starbase import Connection

c = Connection("127.0.0.1", "8000")

ratings = c.table('ratings')

if (ratings.exists()):
    print("Dropping existing ratings table\n")
    ratings.drop()

ratings.create('rating')

print("Parsing the ml-100k ratings data...\n")
ratingFile = open("u.data", "r")

batch = ratings.batch()
print(batch)

for line in ratingFile:
    (userID, movieID, rating, timestamp) = line.split()
    batch.update(userID, {'rating': {movieID: rating}})

ratingFile.close()

print ("Committing ratings data to HBase via REST service\n")
batch.commit(finalize=True)

print ("Get back ratings for some users...\n")
print ("Ratings for user ID 1:\n")
print (ratings.fetch("1"))
print ("Ratings for user ID 33:\n")
print (ratings.fetch("33"))

ratings.drop()
```

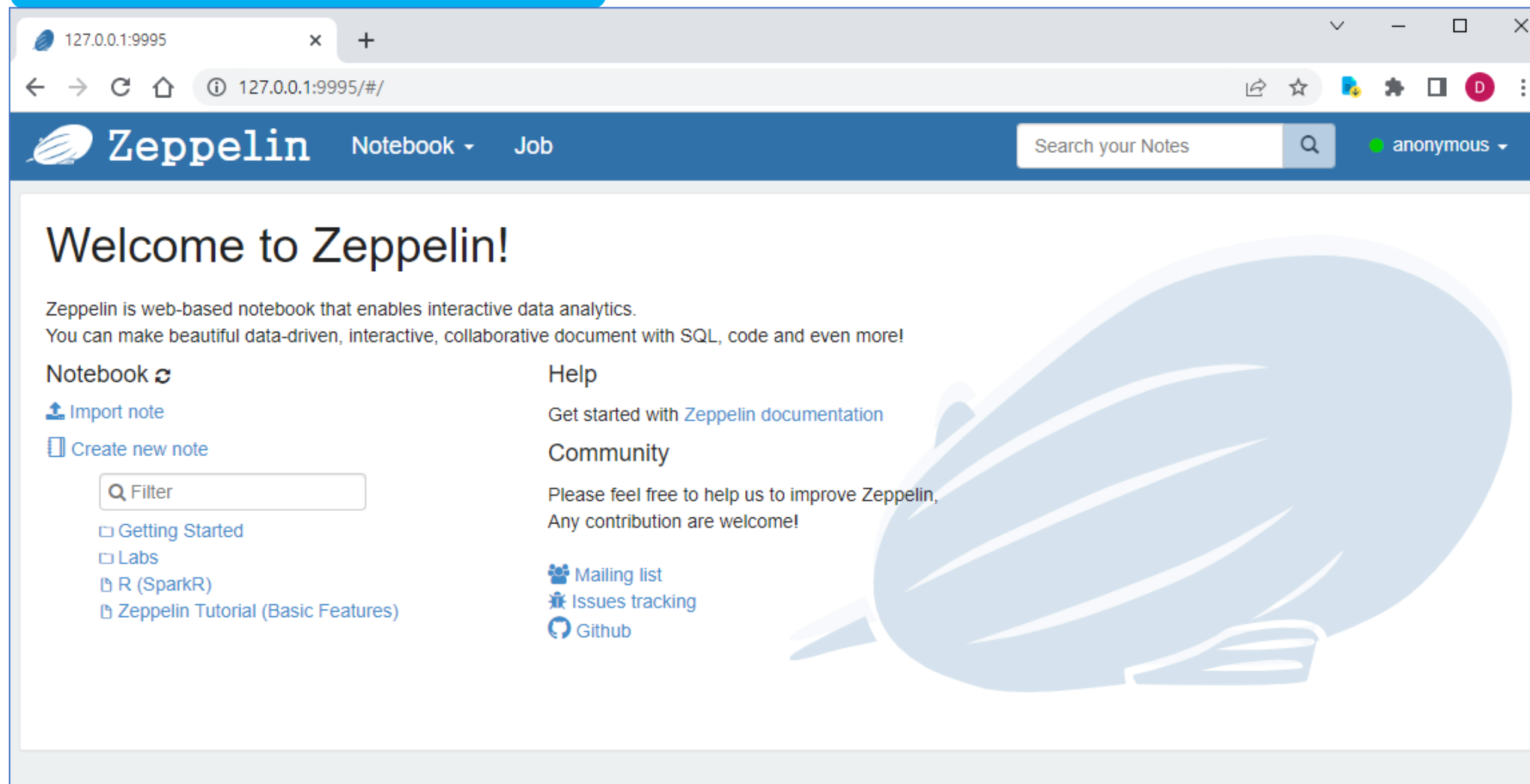
■ HBaseExamples 실행

Python HBaseExamples.py

Zeppelin 실습

Zeppelin

<http://127.0.0.1:9995/>



The screenshot shows the Zeppelin web interface in a browser window. The browser's address bar displays '127.0.0.1:9995/#/'. The Zeppelin header is dark blue with the logo, 'Notebook' dropdown, 'Job' dropdown, a search bar for notes, and a user profile labeled 'anonymous'. The main content area has a large 'Welcome to Zeppelin!' heading, followed by a description of the tool as a web-based notebook for interactive data analytics. On the left, there's a 'Notebook' section with links to 'Import note' and 'Create new note', a search filter, and a list of categories: 'Getting Started', 'Labs', 'R (SparkR)', and 'Zeppelin Tutorial (Basic Features)'. On the right, there's a 'Help' section with a link to 'Zeppelin documentation' and a 'Community' section with links to 'Mailing list', 'Issues tracking', and 'Github'. A large, faint blue illustration of a zeppelin is in the background.

127.0.0.1:9995

127.0.0.1:9995/#/

Zeppelin Notebook Job

Search your Notes

anonymous

Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!

Notebook

Import note

Create new note

Filter

- Getting Started
- Labs
- R (SparkR)
- Zeppelin Tutorial (Basic Features)

Help

Get started with [Zeppelin documentation](#)

Community

Please feel free to help us to improve Zeppelin,
Any contribution are welcome!

- Mailing list
- Issues tracking
- Github

Zeppelin

```
sc.version
```

```
%sh
```

```
wget http://media.sundog-soft.com/hadoop/ml-100k/u.data -O /tmp/u.data  
wget http://media.sundog-soft.com/hadoop/ml-100k/u.item -O /tmp/u.item
```

```
%sh
```

```
hadoop fs -rm -r -f /tmp/m1-100k
```

```
hadoop fs mkdir /tmp/m1-100k
```

```
hadoop fs -put /tmp/u.data /tmp/m1-100k
```

```
hadoop fs -put /tmp/u.item item /tmp/m1-100k
```


Zeppelin

The screenshot shows the Zeppelin Notebook web interface. The browser address bar indicates the URL `127.0.0.1:9995/#/notebook/2J8MM35T5`. The Zeppelin logo and navigation tabs for 'Notebook' and 'Job' are visible. A search bar and a user dropdown menu (showing 'anonymous') are on the right. The notebook title 'zeppelin_demo' is on the left, followed by a toolbar with icons for running, saving, and other actions. The main area contains three code blocks, each with a 'READY' status and control icons on the right.

sc.version

%sh

```
wget https://raw.githubusercontent.com/kgpark88/bigdata/main/ml-100k/u.data -O /tmp/u.data
wget https://raw.githubusercontent.com/kgpark88/bigdata/main/ml-100k/u.item -O /tmp/u.item
```

%sh

```
hadoop fs -rm -r -f /tmp/m1-100k
hadoop fs mkdir /tmp/m1-100k
hadoop fs -put /tmp/u.data /tmp/m1-100k
hadoop fs -put /tmp/u.item item /tmp/m1-100k
```

Zeppelin

127.0.0.1:9995/#/notebook/2A94M5J1Z

Zeppelin Notebook Job

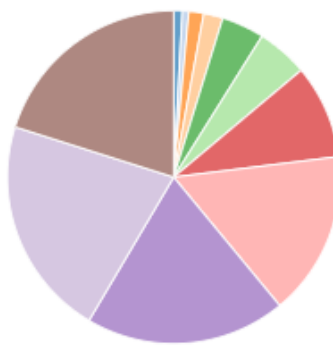
Search your Notes anonymous

Zeppelin Tutorial (Basic Features)

```
%spark.sql
select age, count(1) value
from bank
where age < 30
group by age
order by age
```

FINISHED

19 20 21 22 23 24 25 26 27 28 29



Took 51 sec. Last updated by anonymous at December 18 2016, 12:31:04 AM. (outdated)

```
%spark.sql
select age, count(1) value
from bank
where age < ${maxAge=30}
group by age
order by age
```

FINISHED

maxAge

35

age	value
19	4
20	3
21	7
22	9
23	20
24	24
25	44
26	77
27	94

Took 9 sec. Last updated by anonymous at December 18 2016, 12:31:07 AM. (outdated)

```
%spark.sql
select age, count(1) value
from bank
where marital=${marital=single,single|divorced|married}
group by age
order by age
```

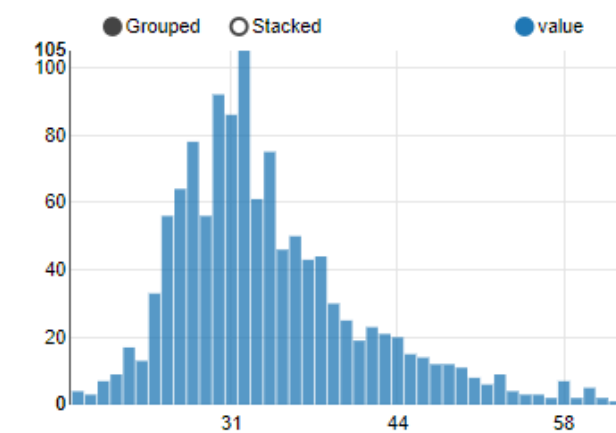
FINISHED

marital

single

Grouped Stacked

value



Took 4 sec. Last updated by anonymous at December 18 2016, 12:31:09 AM. (outdated)

50

Thank you