

به نام خدا
دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

مبانی پردازش ابری

گزارش کار تمرین ۳
آشنایی عملیاتی با Hadoop و MapReduce

استاد درس: دکتر جوادی

محمد رضا شهرستانی

۹۷۲۸۰۵۴

نیم سال دوم ۱۴۰۰-۱۴۰۱

گام اول

(۱) ساخت ۳ ماشین مجازی:

```
C:\Users\m shahr>multipass list
Name                State      IPv4             Image
h-primary           Running    172.30.50.203    Ubuntu 20.04 LTS
h-secondary1        Running    172.30.54.120    Ubuntu 20.04 LTS
h-secondary2        Running    172.30.59.72     Ubuntu 20.04 LTS
```

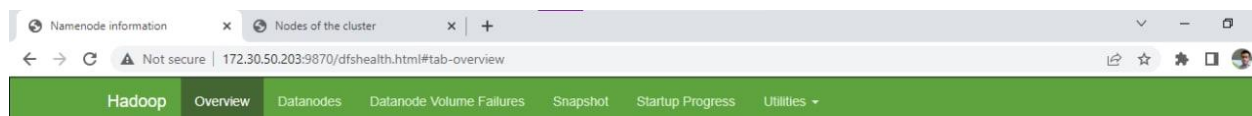
(۲ و ۳) دستور jps و گرفتن نقش node ها:

```
h-user@h-primary:~$ jps
13120 ResourceManager
12952 SecondaryNameNode
15341 Jps
12701 NameNode
```

```
h-user@h-secondary1:~$ jps
12960 NodeManager
13079 Jps
11551 DataNode
h-user@h-secondary1:~$
```

```
h-user@h-secondary2:~$ jps
12914 Jps
11723 DataNode
12795 NodeManager
h-user@h-secondary2:~$
```

(۴) نشان دادن WebGUI:

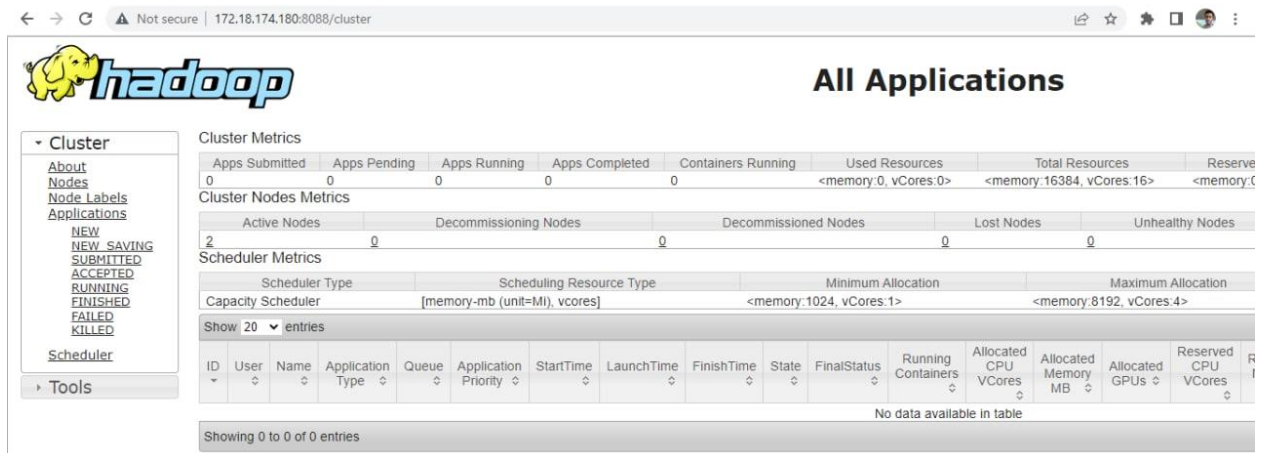


Overview 'h-primary:9000' (active)

Started:	Sun Jun 05 21:27:32 +0430 2022
Version:	3.2.2, r7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled:	Sun Jan 03 12:56:00 +0330 2021 by hexiaoqiao from branch-3.2.2
Cluster ID:	CID-cc29017c-ade9-49c7-995b-ca0944ee7098
Block Pool ID:	BP-2113101152-172.21.156.219-1654333196352

Summary

Security is off.
Safemode is off

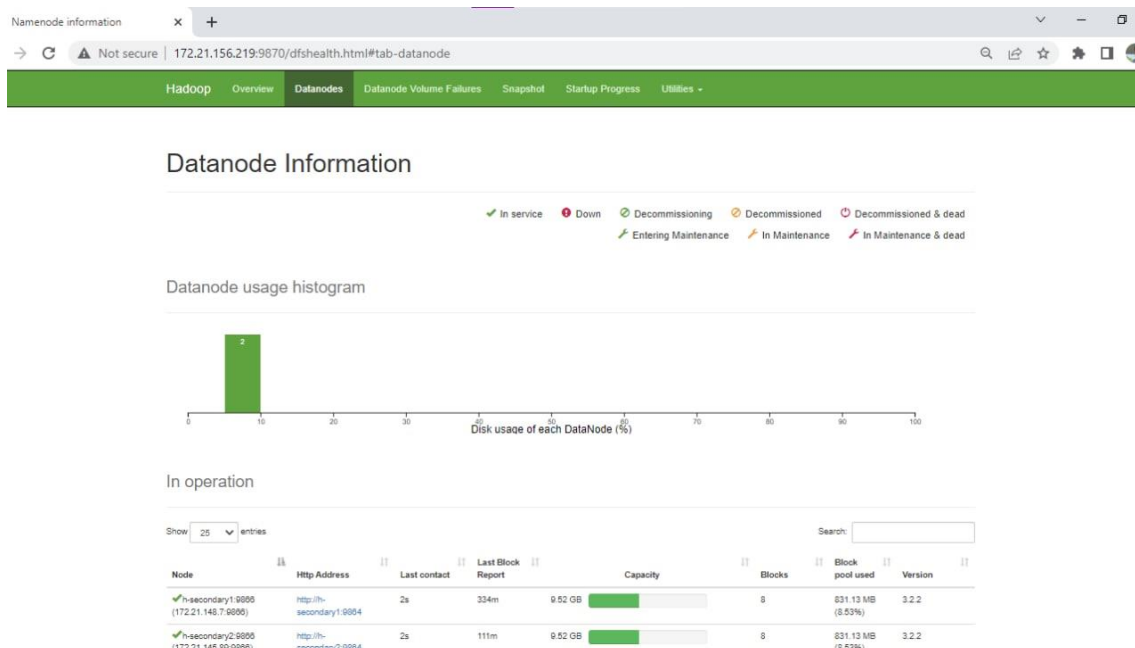


The screenshot shows the Hadoop All Applications page. On the left is a navigation menu with options like Cluster, About, Nodes, Node Labels, Applications, NEW, NEW SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED, Scheduler, and Tools. The main content area displays various metrics:

- Cluster Metrics:** Apps Submitted (0), Apps Pending (0), Apps Running (0), Apps Completed (0), Containers Running (0), Used Resources (memory:0, vCores:0), Total Resources (memory:16384, vCores:16), Reserve (memory:0).
- Cluster Nodes Metrics:** Active Nodes (0), Decommissioning Nodes (0), Decommissioned Nodes (0), Lost Nodes (0), Unhealthy Nodes (0).
- Scheduler Metrics:** Scheduler Type (Capacity Scheduler), Scheduling Resource Type (memory-mb (unit=Mi), vcores), Minimum Allocation (memory:1024, vCores:1), Maximum Allocation (memory:8192, vCores:4).

Below these metrics is a table with columns: ID, User, Name, Application Type, Queue, Application Priority, StartTime, LaunchTime, FinishTime, State, FinalStatus, Running Containers, Allocated CPU Vcores, Allocated Memory MB, Allocated GPUs, Reserved CPU Vcores, and Reserved Memory MB. The table currently shows "No data available in table".

active nodes (۵)



The screenshot shows the Hadoop Datanode Information page. It includes a legend for node states: In service (green check), Down (red X), Decommissioning (green circle), Decommissioned (orange circle), Decommissioned & dead (red circle), Entering Maintenance (green circle), In Maintenance (orange circle), and In Maintenance & dead (red circle).

Datanode usage histogram: A bar chart showing disk usage of each DataNode (%). The x-axis ranges from 0 to 100, and the y-axis shows a single bar at approximately 8.53%.

In operation: A table showing details for two data nodes:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
h-secondary1-9888 (172.21.148.7:9888)	http://h-secondary1-9884	2s	334m	9.52 GB (8.53%)	8	831.13 MB (8.53%)	3.2.2
h-secondary2-9888 (172.21.145.80:9888)	http://h-secondary2-9884	2s	111m	9.52 GB (8.53%)	8	831.13 MB (8.53%)	3.2.2

همان طور که در تصاویر قبلی نشان داده شد دو data node با ظرفیت ۱۰ گیگابایت حافظه و دو گیگابایت رم و دو core پردازنده با multipass ساخته شده است. ولی در این قسمت برای هر نود ۸ گیگابایت در نظر گرفته شده است و cpu در نظر گرفته نشده است و به طور پیش فرض در تنظیمات hadoop این اعداد ذکر شده اند.

گام دوم

(۱) ساخت پوشه در file system هدوپ:

```
h-user@h-primary:~$ hdfs dfs -mkdir /user/hadoop
h-user@h-primary:~$ hdfs dfs -mkdir /user/hadoop/input
h-user@h-primary:~$ hdfs dfs -copyFromLocal a.txt /user/hadoop/input
h-user@h-primary:~$ hdfs dfs -copyFromLocal new_hashtag_donaldtrump.csv /user/hadoop/input
```

(۲) خارج کردن از زیپ:

```
h-user@h-primary:~$ ls
LR_count      LR_count_reducer.py  geo_state.log      new_hashtag_donaldtrump.csv  state.log          y
LR_count.log   datasets.zip          geo_state_mapper.py  new_hashtag_joebiden.csv    state_mapper.py    y.pub
```

(۳) انتقال دو فایل به file system هدوپ:

```
h-user@h-primary:~$ hdfs dfs -ls /user/hadoop/input
Found 2 items
-rw-r--r--  2 h-user  supergroup  483855307  2022-06-04 13:37 /user/hadoop/input/new_hashtag_donaldtrump.csv
-rw-r--r--  2 h-user  supergroup  380817742  2022-06-04 13:38 /user/hadoop/input/new_hashtag_joebiden.csv
h-user@h-primary:~$
```

نتایج زیر از فایل‌های نتایج اجرای هر برنامه از output برداشته شده است.

(۴) برنامه‌ای که تعداد لایک، ریتوییت و سورس‌ها را نمایش دهد:

Both Candidate	likes	retweets	Twitter Web App	Twitter for iPhone	Twitter for Android
Donald Trump	likes	retweets	Twitter Web App	Twitter for iPhone	Twitter for Android
Joe Biden	likes	retweets	Twitter Web App	Twitter for iPhone	Twitter for Android

```
h-user@h-primary:~$ hdfs dfs -cat /user/hadoop/output/LR_count/part-00000
biden  5416591 1133359 142708 166038 138071
both   4178707 882126 159573 108233 131152
trump  4661504 1102474 200978 167659 172318
```

(۵) برنامه‌ای که توپیت‌های ساعت ۹ تا ۱۷ ایالت‌ها را نشان می‌دهد:

```
h-user@h-primary:~$ hdfs dfs -cat /user/hadoop/output/state/part-00000
california 0.21683616658006113 0.38876104223888297 0.3944027911810558 13471
florida 0.2406596762699777 0.36353456174284765 0.39580576198717243 9823
new york 0.2610235923720512 0.36330745458656877 0.3756689530413817 13267
texas 0.24847796924981966 0.3655969456196469 0.38592508513053075 9691
```

(۶) برنامه‌ای که توپیت‌های ساعت ۹ تا ۱۷ ایالت‌ها را با مختصات جغرافیایی نشان می‌دهد:

```
h-user@h-primary:~$ hdfs dfs -cat /user/hadoop/output/geo_state/part-00000
california 0.21906652066794438 0.3877634820695319 0.3931699972625257 14612
new york 0.2531429146132955 0.34707834657548037 0.399778738811222 19886
```

مقایسه نتایج ۵ و ۶ به نظرم اختلاف تعداد میان این دو بخش، این است که احتمالا متادیتا مربوط به موقعیت مکانی (عرض و طول جغرافیایی) از طریق جی پی اس جمع آوری می‌شود اما چیزی که به عنوان ایالت ثبت می‌شود از روی IP افراد به دست می‌آید و اگر افراد از سرویس‌های تغییر IP استفاده کنند مانند VPN، میان این دو مورد تفاوت دیده می‌شود.

نکته فایل‌های mapper و reducer و log اجرا و فایل‌های نتایج همگی در فایل files قرار دارند.