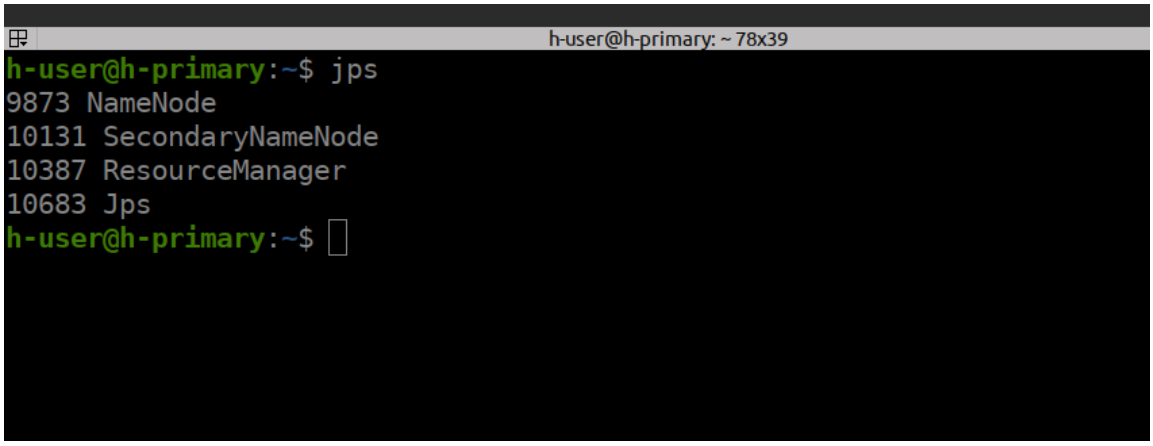
به نام خدا

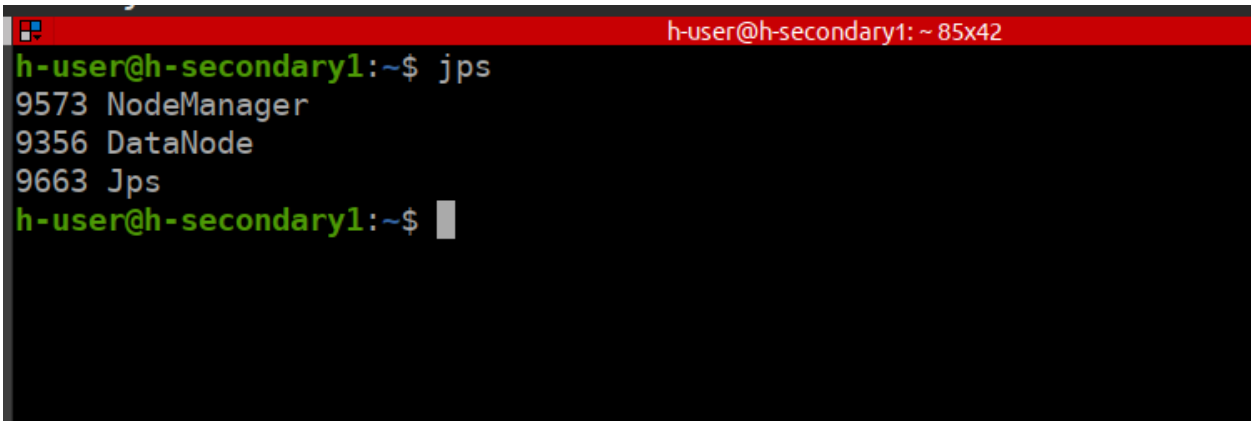
محمدعلی کشت پرور ۹۷۳۴۰۲۲ **گزارش تمرین سوم رایانش ابری**

گام اول

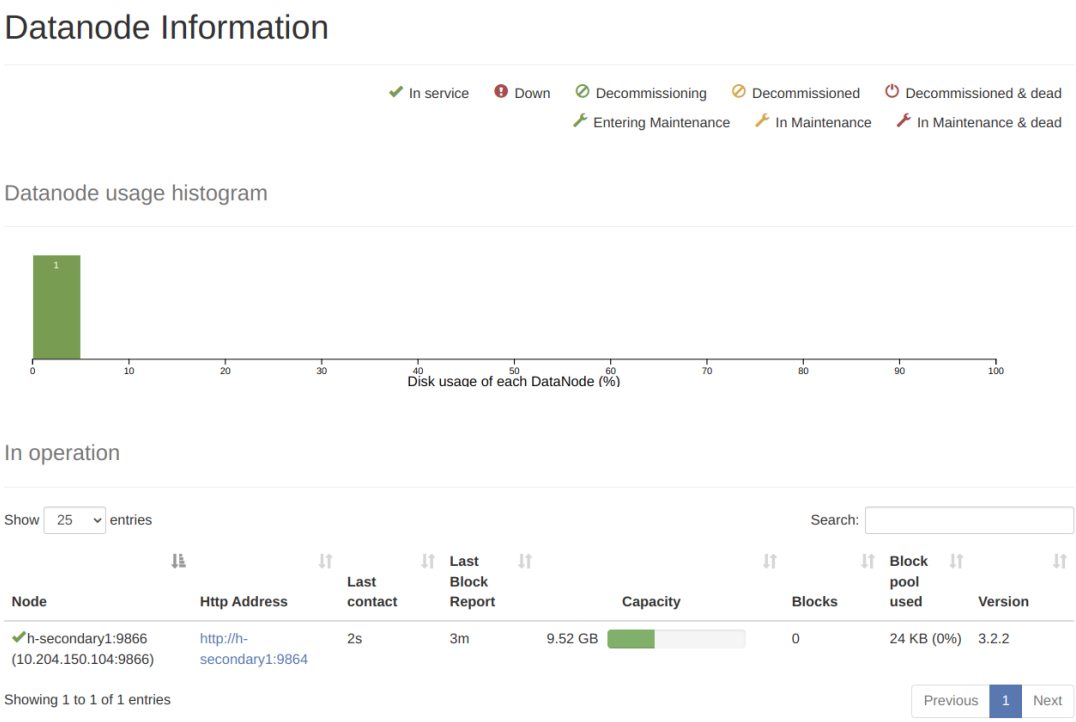
با توجه به محدود بودن منابع ، دو ماشین مجازی با استفاده از multipass می سازیم و کانفیگ های لازم را برای ایجاد کلاستر هدوپ انجام می دهیم. یکی از ماشین ها را به عنوان nameNode و ماشین دیگر به عنوان dataNode در نظر می گیریم . به هر ماشین ۱۰ گیگ حافظه و ۱ گیگ رم اختصاص می دهیم . در تصاویر زیر می توان صحت ایجاد کلاستر را با استفاده از دستور jps مشاهده کرد.



ماشین primary



ماشین secondary



همانطور که مشاهده می شود webGUI اطلاعات dataNode را به درستی نمایش می دهد. این اطلاعات با مقدار منابع اختصاص داده شده به ماشین تطابق دارد.

گام دوم

قرار دادن فایل ها روی HDFS

```
h-user@h-primary:/home/ubuntu/shared/light_dataset$ hadoop fs -mkdir -p /user/hadoop/input
h-user@h-primary:/home/ubuntu/shared/light_dataset$ hadoop fs -put light_dataset.csv  /user/hadoop/input
h-user@h-primary:/home/ubuntu/shared/light_dataset$ hadoop fs -ls /user/hadoop/input
Found 1 items
-rw-r--r--   1 h-user supergroup  103280184 2022-06-02 04:02 /user/hadoop/input/light_dataset.csv
h-user@h-primary:/home/ubuntu/shared/light_dataset$
```

```
h-user@h-primary: /home/ubuntu/shared 190x51
h-user@h-primary:/home/ubuntu/shared$ hadoop fs -ls /user/hadoop/input
Found 3 items
-rw-r--r--   1 h-user supergroup  103280184 2022-06-02 04:02 /user/hadoop/input/light_dataset.csv
-rw-r--r--   1 h-user supergroup  483855307 2022-06-02 13:47 /user/hadoop/input/new_hashtag_donaldtrump.csv
-rw-r--r--   1 h-user supergroup  380817742 2022-06-02 13:48 /user/hadoop/input/new_hashtag_joebiden.csv
h-user@h-primary:/home/ubuntu/shared$
```

همانطور که مشاهده می شود ورودی ها با replication 1 رو HDFS قرار گرفته اند.

۴-نمایش تعداد لایک ها ، retweetها و تعداد source ها

```
-------------------------------------------------------------------------------------------
Biden          : like=5416591   | retweet=1133359  | android=138536   | iPhone=166042  | web=142865
Both candidate : like=4178707   | retweet=882126   | android=131695   | iPhone=108239  | web=159726
Trump          : like=4661504   | retweet=1102474  | android=173390   | iPhone=167664  | web=201185
-------------------------------------------------------------------------------------------
0
h-user@h-primary:/home/ubuntu/shared$
```

۵- تعداد توییت ها در بازه زمانی مورد نظر در ایالت های ذکر شده

```
-------------------------------------------------------------------------
New York   0.26102359237205097   0.36330745458656816   0.37566895304138087   13267
Texas   0.2484779692498194   0.3655969456196471   0.3859250851305335   9691
California   0.21683616658006089   0.3887610422388835   0.3944027911810556   13471
Florida   0.2406596762699786   0.3635345617428484   0.395805761987173   9823
-------------------------------------------------------------------------
0
h-user@h-primary:/home/ubuntu/shared$
```

۶- تعداد توییت ها در بازه زمانی مورد نظر در ایالت های ذکر شده با استفاده از مختصات جغرافیایی

```
-------------------------------------------------------------------------
New York   0.2531429146132958   0.34707834657548026   0.399778738811224   19886
California   0.21906652066794416   0.3877634820695319   0.39316999726252394   14612
-------------------------------------------------------------------------
0
```

نتایج بخش ۵ و ۶ با هم متفاوت می باشد و زیرا در قسمت ۵ بسیاری از مقادیر فیلد state با null پر شده بود و این موضوع باعث می شود که خیلی از توییت های موجود در آن ایالت در نظر گرفته نشود . از طرفی مقدار مختصات هم تقریبی می باشد و ممکن است خطا داشته باشد . در واقع بهتر است قبل از پردازش داده ، پاکسازی داده صورت بگیرد .

تمامی نتایج بالا با ورودی دیتاست اصلی به دست آمده است .