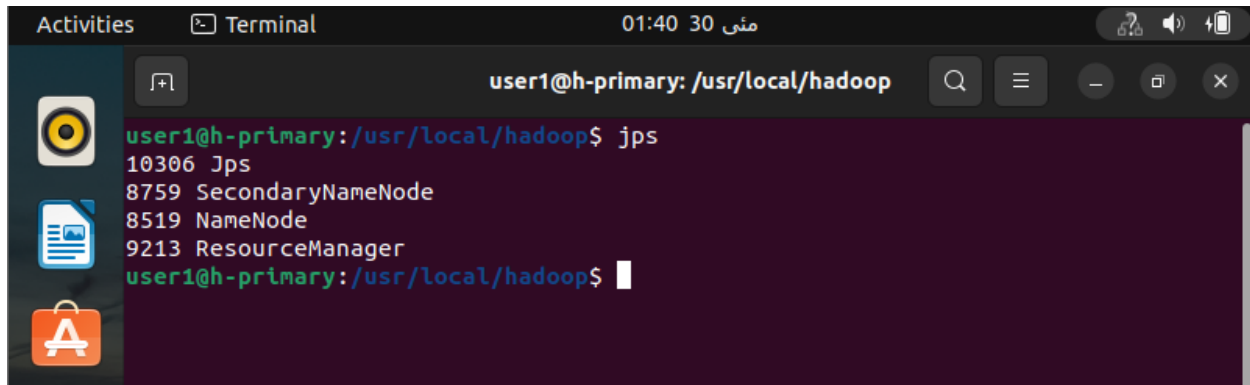


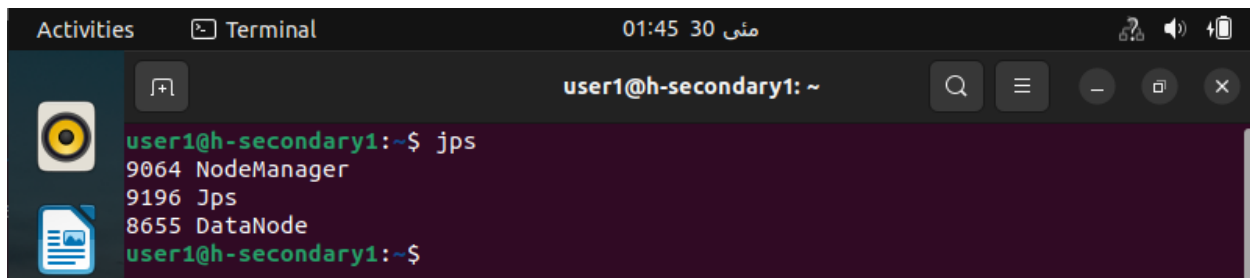
## گام اول: نصب و راه اندازی خوشه‌ی Hadoop

## نقش ماشین h-primary



```
Activities Terminal 01:40 30 مئی user1@h-primary: /usr/local/hadoop
user1@h-primary:/usr/local/hadoop$ jps
10306 Jps
8759 SecondaryNameNode
8519 NameNode
9213 ResourceManager
user1@h-primary:/usr/local/hadoop$
```

## نقش ماشین h-secondary1



```
Activities Terminal 01:45 30 مئی user1@h-secondary1: ~
user1@h-secondary1:~$ jps
9064 NodeManager
9196 Jps
8655 DataNode
user1@h-secondary1:~$
```

طبق تصاویر مشخص است ماشین h-primary نقش های Jps, SecondaryNameNode, NameNode, ResourceManager و ماشین h-secondary1 نقش های Jps, NodeManager, DataNode را به خود گرفته است.

**\* توجه:** به دلیل نکشیدن ران کردن سه ماشین مجازی، ماشین مجازی سوم خاموش شده و کار با دو ماشین مجازی h-primary و h-secondary1 انجام شده است یعنی یک worker داریم.

پس از چک کردن نقش های گرفته شده توسط ماشین مجازی حالت باید بالا آمدن WebGUI را چک کنیم که ۱۹۲.۱۶۸.۰.۱۰۱:۹۸۷۰ بریم که همان آدرس IP ماشین h-primary می باشد.

WebGUI به درستی بالا می آید که تصاویر آن به صورت زیر میباشد:

## Overview 'h-primary:9000' (active)

<b>Started:</b>	Mon Jun 06 12:16:37 +0430 2022
<b>Version:</b>	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
<b>Compiled:</b>	Tue Sep 10 20:26:00 +0430 2019 by rohithsharmaks from branch-3.2.1
<b>Cluster ID:</b>	CID-9edd6692-1f53-43cb-bfc0-0a6f1165b47f
<b>Block Pool ID:</b>	BP-646140595-127.0.1.1-1653847152911

## Summary

Security is off.

Safemode is off.

15 files and directories, 5 blocks (5 replicated blocks, 0 erasure coded block groups) = 20 total filesystem object(s).

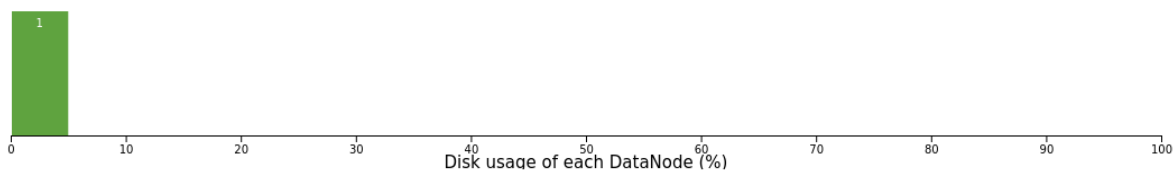
Heap Memory used 166.67 MB of 281.5 MB Heap Memory. Max Heap Memory is 873 MB.

Non Heap Memory used 71.78 MB of 73.61 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

<b>Configured Capacity:</b>	19.02 GB
<b>Configured Remote Capacity:</b>	0 B
<b>DFS Used:</b>	99.34 MB (0.51%)
<b>Non DFS Used:</b>	10.28 GB
<b>DFS Remaining:</b>	7.65 GB (40.22%)
<b>Block Pool Used:</b>	99.34 MB (0.51%)
<b>DataNodes usages% (Min/Median/Max/stdDev):</b>	0.51% / 0.51% / 0.51% / 0.00%
<b>Live Nodes</b>	1 (Decommissioned: 0, In Maintenance: 0)
<b>Dead Nodes</b>	0 (Decommissioned: 0, In Maintenance: 0)

که در قسمت Live Node میبینیم یک نود فعال داریم و قسمت Dead Nodes خالی است و این نشان میدهد نود به درستی بالا آمده است.

Datanode usage histogram



که نشان میدهد یک نود فعال داریم.

## In operation

Show 25 entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓h-secondary1:9866 (192.168.0.103:9866)	http://h-secondary1:9866	0s	157m	19.02 GB	5	99.34 MB (0.51%)	3.2.1

Showing 1 to 1 of 1 entries

Previous 1 Next

این تصویر نیز نشان میدهد نود ما h-secondary1 فعال شده است.

همچنین Yarn نیز فعال شده است که باید به آدرس ۱۹۲.۱۶۸.۰.۱۰۱:۸۰۸۸ بریم که عکس آن به صورت زیر است:

Cluster Metrics																													
Apps Submitted		Apps Pending		Apps Running		Apps Completed		Containers Running		Memory Used		Memory Total		Memory Reserved		VCores Used		VCores Total		VCores Reserved									
0		0		0		0		0		0 B		8 GB		0 B		0		8		0									
Cluster Nodes Metrics																													
Active Nodes			Decommissioning Nodes			Decommissioned Nodes			Lost Nodes			Unhealthy Nodes			Rebooted Nodes			Shutdown Nodes											
1			0			0			0			0			0			0											
Scheduler Metrics																													
Scheduler Type		Scheduling Resource Type				Minimum Allocation				Maximum Allocation				Maximum Cluster Application Priority															
Capacity Scheduler		[memory-mb (unit=Mi), vcores]				<memory:1024, vCores:1>				<memory:8192, vCores:4>				0															
Show 20 ▾ entries																		Search:											
ID ▾	User ▾	Name ▾	Application Type ▾	Queue ▾	Application Priority ▾	StartTime ▾	LaunchTime ▾	FinishTime ▾	State ▾	FinalStatus ▾	Running Containers	Allocated CPU Vcores ▾	Allocated Memory MB ▾	Reserved CPU Vcores ▾	Reserved Memory MB ▾	% of Queue ▾	% of Cluster ▾	Progress	Tracking UI ▾	Blacklisted Nodes ▾									
No data available in table																													
Showing 0 to 0 of 0 entries																													
																		First			Previous			Next			Last		

که میبینیم در قسمت active nodes یک نود فعال داریم که درواقع این قسمت تعداد NodeManager های ما را نشان میدهد که مقدار ۱ گرفته است زیرا ماشین h-secondary1 نقش NodeManager را گرفته است.

## گام دوم: توسعه و اجرای برنامه ی Mapreducer

پس از اجرای مراحل ۱ تا ۳ فایل سیستم به صورت زیر در می آید:

Show  entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	user1	supergroup	0 B	May 30 23:18	0	0 B	user	

Showing 1 to 1 of 1 entries Previous **1** Next

Show  entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	user1	supergroup	0 B	May 30 16:05	0	0 B	hadoop	
<input type="checkbox"/>	drwxr-xr-x	user1	supergroup	0 B	Jun 06 15:20	0	0 B	user1	

Showing 1 to 2 of 2 entries Previous **1** Next

Show  entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	user1	supergroup	22 B	May 30 23:18	1	128 MB	input.txt	
<input type="checkbox"/>	-rw-r--r--	user1	supergroup	98.5 MB	Jun 06 12:25	1	128 MB	light_dataset.csv	

\* توجه: از دیتاست سبک یعنی light\_dataset.csv استفاده شده است.

حال وقت آن رسیده است که کد هارا به کلاستر هدوپ ببریم. برای اینکار کد هارا به دایرکتوری اصلی user1 منتقل میکنیم.

سپس دستور زیر را برای اجرا شدن کد قسمت ۴ اجرا میکنیم:

```
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar
-mapper "python3 mapper4.py" -reducer "python3 reducer4.py" -input
/user/user1/light_dataset.csv -output /user/user1/output1
```

که خروجی آن را در دایرکتوری output1 میریزیم که میبینم دایرکتوری زیر ظاهر میشود:

	<a href="#">drwxr-xr-x</a>	<a href="#">user1</a>	<a href="#">supergroup</a>	0 B	Jun 06 15:20	0	0 B	<a href="#">output1</a>	
--	----------------------------	-----------------------	----------------------------	-----	--------------	---	-----	-------------------------	--

که اگر وارد آن شویم فایل های زیر ایجاد شده اند:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
	-rw-r--r--	<a href="#">user1</a>	<a href="#">supergroup</a>	0 B	Jun 06 15:20	1	128 MB	<a href="#">_SUCCESS</a>	
	-rw-r--r--	<a href="#">user1</a>	<a href="#">supergroup</a>	134 B	Jun 06 15:20	1	128 MB	<a href="#">part-00000</a>	

که نتایج در فایل part-00000 ذخیره شده اند که محتویات آن به صورت زیر می باشد:

File contents						
Both Candidate	204750	78843	17541	9479	7916	
Donald Trump	370451	126000	24885	17710	15396	
Joe Biden	801355	267920	24737	21786	14031	

حال کد قسمت ۵ را با دستور زیر اجرا میکنیم:

```
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar
-mapper "python3 mapper5.py" -reducer "python3 reducer5.py" -input
/user/user1/light_dataset.csv -output /user/user1/output2
```

که خروجی آن را در دایرکتوری output1 میریزیم که میبینم دایرکتوری زیر ظاهر میشود:

	<a href="#">drwxr-xr-x</a>	<a href="#">user1</a>	<a href="#">supergroup</a>	0 B	Jun 06 15:20	0	0 B	<a href="#">output2</a>	
--	----------------------------	-----------------------	----------------------------	-----	--------------	---	-----	-------------------------	--

که اگر وارد آن شویم فایل های زیر ایجاد شده اند:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
	-rw-r--r--	<a href="#">user1</a>	<a href="#">supergroup</a>	0 B	Jun 06 15:20	1	128 MB	<a href="#">_SUCCESS</a>	
	-rw-r--r--	<a href="#">user1</a>	<a href="#">supergroup</a>	304 B	Jun 06 15:20	1	128 MB	<a href="#">part-00000</a>	

که نتایج در فایل part-00000 ذخیره شده اند که محتویات آن به صورت زیر می باشد:

File contents					
new york	0.22832780672666983	0.41544291804831834	0.35622927522501185	2111	
texas	0.2068519715578539	0.45765998707175176	0.3354880413703943	1547	
california	0.1897407611693326	0.419746276889134	0.39051296194153334	1813	
florida	0.24957651044607565	0.4748729531338227	0.27555053642010163	1771	

حال کد قسمت ۶ را با دستور زیر اجرا میکنیم:

```
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar
-mapper "python3 mapper6.py" -reducer "python3 reducer6.py" -input
/user/user1/light_dataset.csv -output /user/user1/output3
```

که خروجی آن را در دایرکتوری output1 میریزیم که میبینم دایرکتوری زیر ظاهر میشود:

	<a href="#">drwxr-xr-x</a>	<a href="#">user1</a>	<a href="#">supergroup</a>	0 B	Jun 06 15:20	<a href="#">0</a>	0 B	<a href="#">output3</a>	
--	----------------------------	-----------------------	----------------------------	-----	--------------	-------------------	-----	-------------------------	--

که اگر وارد آن شویم فایل های زیر ایجاد شده اند:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">user1</a>	<a href="#">supergroup</a>	0 B	Jun 06 15:20	<a href="#">1</a>	128 MB	<a href="#">_SUCCESS</a>	
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">user1</a>	<a href="#">supergroup</a>	153 B	Jun 06 15:20	<a href="#">1</a>	128 MB	<a href="#">part-00000</a>	

که نتایج در فایل part-00000 ذخیره شده اند که محتویات آن به صورت زیر می باشد:

File contents					
new york	0.24517374517374518	0.3915701415701416	0.36325611325611323	3108	
california	0.19300911854103345	0.42806484295846	0.3789260385005066	1974	

نتایج قسمت های ۵ و ۶ برای ایالت های new York و California کمی متفاوت است به این دلیل که درس سوال ۶ طول و عرض جغرافیایی داده شده به صورت تقریبی هستند و مقادیری دقیقی نیستند از این جهت نتایج این دو قسمت کمی متفاوت است.