













تمرین سوم – رایانش ابری

- گام اول

• نمایش کانتینرهای ایجاد شده

<input type="checkbox"/>	<div><div></div><div><div>hadoop</div><div>-</div></div></div> <div>Running (5/5)</div> <div>26 minutes ago</div> <div><div><div></div><div></div><div></div></div></div> <div><div></div><div></div><div></div></div>
<input type="checkbox"/>	<div><div></div><div><div>namenode</div><div>a70d919b772d </div></div></div> <div>arminzolfagharid/hadoop-namenode:v2-hadoop3</div> <div>Running</div> <div>9000:9000  Show all ports (2)</div> <div>27 minutes ago</div> <div><div><div></div><div></div><div></div></div></div> <div><div></div><div></div><div></div></div>
<input type="checkbox"/>	<div><div></div><div><div>nodemanager</div><div>7ea720a87beb </div></div></div> <div>arminzolfagharid/hadoop-nodemanager:v2-hadoo</div> <div>Running</div> <div></div> <div>27 minutes ago</div> <div><div><div></div><div></div><div></div></div></div> <div><div></div><div></div><div></div></div>
<input type="checkbox"/>	<div><div></div><div><div>historyserver</div><div>1de0dd598424 </div></div></div> <div>arminzolfagharid/hadoop-historyserver:v2-hadoo</div> <div>Running</div> <div></div> <div>27 minutes ago</div> <div><div><div></div><div></div><div></div></div></div> <div><div></div><div></div><div></div></div>
<input type="checkbox"/>	<div><div></div><div><div>resourcemanager</div><div>2db970b93083 </div></div></div> <div>arminzolfagharid/hadoop-resourcemanager:v2-ha</div> <div>Running</div> <div></div> <div>26 minutes ago</div> <div><div><div></div><div></div><div></div></div></div> <div><div></div><div></div><div></div></div>
<input type="checkbox"/>	<div><div></div><div><div>datanode</div><div>e778d66363dc </div></div></div> <div>arminzolfagharid/hadoop-datanode:v2-hadoop3.2</div> <div>Running</div> <div></div> <div>27 minutes ago</div> <div><div><div></div><div></div><div></div></div></div> <div><div></div><div></div><div></div></div>

• توضیح وظیفه هرکدام از کانتینرها در Hadoop

در یک خوشه Hadoop معمولاً از چند کانتینر استفاده می‌شود، هر کدام از این کانتینرها نقش و وظیفه خاصی در سیستم دارند. در زیر توضیحی کوتاه از وظایف هر کدام از کانتینرها در Hadoop آورده شده است:

۱. *****NameNode*****: کانتینر NameNode نقش مرکزی در سیستم Hadoop دارد. این کانتینر نقشی مشابه یک سرور مدیریتی ایفا می‌کند و اطلاعات مربوط به مکان و بلوک‌بندی داده‌ها را در خود ذخیره می‌کند. همچنین امکاناتی مانند مدیریت فضای ذخیره‌سازی و تخصیص منابع را در اختیار دارد.
۲. *****DataNode*****: هر کانتینر DataNode مربوط به یک سرور اجرا شده است که داده‌های واقعی را ذخیره می‌کند. این کانتینرها مسئول ذخیره و مدیریت بلوک‌های داده در خوشه Hadoop هستند و اطلاعات داده‌ها را بر اساس توزیع و بلوک‌بندی در نودهای مختلف نگهداری می‌کنند.
۳. *****NodeManager*****: هر کانتینر NodeManager مسئول مدیریت منابع محاسباتی (مانند پردازنده و حافظه) در یک نود است. این کانتینرها وظیفه زمان‌بندی و اجرای وظایف MapReduce را بر روی نودها بر عهده دارند و با مدیریت منابع، وظیفه اجرای تسک‌های Map و Reduce را به صورت موازی بر روی داده‌ها انجام می‌دهند.
۴. *****ResourceManager*****: کانتینر ResourceManager نقش مرکزی در مدیریت وظایف MapReduce در خوشه Hadoop دارد. این کانتینر مسئول تخصیص منابع، زمان‌بندی و مدیریت وظایف MapReduce است. ResourceManager با کمک NodeManagerها، وظایف را به نودها ارسال می‌کند و تخصیص منابع را بر اساس نیازهای هر وظیفه انجام می‌دهد.
۵. *****HistoryManager*****: کانتینر HistoryManager وظیفه ذخیره سازی و مدیریت لاگ‌ها و اطلاعات سابقه اجرای وظایف MapReduce را بر عهده دارد. این کانتینر اطلاعاتی مانند لاگ‌ها، زمان‌بندی و پیشرفت وظایف را در اختیار کاربران و مدیران قرار می‌دهد تا بتوانند عملکرد سیستم را بررسی و تجزیه و تحلیل کنند.

این وظایف برای خوشه Hadoop معمولی توضیح داده شده‌اند. در برخی موارد خاص، امکان استفاده از سایر کانتینرها و سرویس‌ها نیز وجود دارد تا وظایف خاصی مانند مدیریت صف‌ها، امنیت و بکاپ‌گیری را بر عهده بگیرند.

- با استفاده از دستور jps در هر کانتینر، صحت نقش آن کانتینر در Hadoop را بررسی کنید و اسکرین شات آن را بیاورید.

```
C:\Users\Samin>docker exec -it namenode bash
root@a70d919b772d:/# hps
bash: hps: command not found
root@a70d919b772d:/# exit
exit

C:\Users\Samin>docker exec -it namenode bash
root@a70d919b772d:/# jps
360 NameNode
1007 Jps
root@a70d919b772d:/# exit
exit

C:\Users\Samin>docker exec -it nodemanager bash
root@7ea720a87beb:/# jps
1077 Jps
363 NodeManager
root@7ea720a87beb:/# exit
exit

C:\Users\Samin>docker exec -it historyserver bash
root@1de0dd598424:/# jps
371 ApplicationHistoryServer
1006 Jps
root@1de0dd598424:/# exit
exit

C:\Users\Samin>docker exec -it resourcemanager bash
root@2db970b93083:/# jps
352 ResourceManager
1178 Jps
root@2db970b93083:/# exit
exit

C:\Users\Samin>docker exec -it datanode bash
root@e778d66363dc:/# jps
1065 Jps
366 DataNode
root@e778d66363dc:/#
```

• نمایش WebUI برای NameNode (localhost:9870) و فایل سیستم آن

The screenshot displays the Hadoop NameNode WebUI interface. The browser address bar shows the URL `http://localhost:9870/dfshealth.html#tab-overview`. The interface includes a navigation bar with tabs for Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is divided into two sections: Overview and Summary.

Overview 'namenode:9000' (active)

Started:	Sat Jun 10 21:26:42 +0330 2023
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 20:26:00 +0430 2019 by rohitsharmaks from branch-3.2.1
Cluster ID:	CID-c2898e54-979e-4cd7-a59e-f37de4bcc618
Block Pool ID:	BP-585507282-172.22.0.6-1686146153449

Summary

Security is off.
Safemode is off.

98 files and directories, 56 blocks (56 replicated blocks, 0 erasure coded block groups) = 154 total filesystem object(s).

Heap Memory used 74.11 MB of 177.5 MB Heap Memory. Max Heap Memory is 1.36 GB.

Non Heap Memory used 53.8 MB of 55.06 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	250.98 GB
Configured Remote Capacity:	0 B

NameNode Journal Status

Current transaction ID: 1033

Configured Remote Capacity:	0 B
DFS Used:	104.34 MB (0.04%)
Non DFS Used:	9.4 GB
DFS Remaining:	228.66 GB (91.11%)
Block Pool Used:	104.34 MB (0.04%)
DataNodes usages% (Min/Median/Max/stdDev):	0.04% / 0.04% / 0.04% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	56
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Sat Jun 10 21:26:42 +0330 2023
Last Checkpoint Time	Sat Jun 10 21:26:43 +0330 2023
Enabled Erasure Coding Policies	RS-6-3-1024k

Browser tabs: Namenode information, +

Address bar: http://localhost:9870/dfshealth.html#tab-overview

Bookmarks: Interesting things, CE, Mental Health, Tools, Me, Apply, French, ML & DL, Photography, Project, Pooyesh

Enabled Erasure Coding Policies: RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 1033

Journal Manager	State
FileJournalManager(root=/hadoop/dfs/name)	EditLogFileOutputStream(/hadoop/dfs/name/current/edits_inprogress_000000000000000990)

NameNode Storage

Storage Directory	Type	State
/hadoop/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	250.98 GB	104.34 MB (0.04%)	228.66 GB (91.11%)	104.34 MB	1

Hadoop, 2019.

Windows taskbar: Type here to search, 10:03 PM, 6/10/2023

گام دوم

گام دوم: توضیحات دیتاست

- این دیتاست شامل 200000 توییت با مضمون انتخابات امریکا است.
 - رکوردهای این دیتاست دارای 21 ستون هستند.
 - در برخی از رکوردهای دیتاست، ممکن است اطلاعات یک ستون وجود نداشته باشد (خالی یا مقدار Null باشد)
- برای استفاده از دیتاست کافی است فایل **dataset.csv** را در کانتینر **NameNode** قرار دهید.

ابتدا یک پوشه به نام CC تعریف کرده سپس فایل را قرار میدهم:

```
C:\Users\Samin>docker exec -it namenode bash
root@a70d919b772d:/# mkdir CC
root@a70d919b772d:/# exit
exit
```

```
C:\Users\Samin\Desktop\University\Term 7\Cloud Computing\Projects\Project3>docker cp dataset.csv namenode:/CC
C:\Users\Samin\Desktop\University\Term 7\Cloud Computing\Projects\Project3>docker exec -it namenode bash
root@a70d919b772d:/# cd CC
root@a70d919b772d:/CC# ls
dataset.csv
root@a70d919b772d:/CC#
```

سوال ۳:

```

2023-06-10 20:09:26,911 INFO mapreduce.Job: map 0% reduce 0%
2023-06-10 20:09:38,035 INFO mapreduce.Job: map 50% reduce 0%
2023-06-10 20:09:39,042 INFO mapreduce.Job: map 100% reduce 0%
2023-06-10 20:09:44,077 INFO mapreduce.Job: map 100% reduce 100%
2023-06-10 20:09:45,106 INFO mapreduce.Job: Job job_1686419840224_0008 completed successfully
2023-06-10 20:09:45,207 INFO mapreduce.Job: Counters: 54
File System Counters:

```

```

root@a70d919b772d:/CC/ms# hdfs dfs -cat /CC/ms/outputs0/part-00000
2023-06-10 20:12:31,215 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
Joe Biden      likes: 703237.0      retweets: 222133.0      Twitter Web App: 17108      Twitter for iPhone: 14594      Twitter for Android: 9553
Donald Trump   likes: 137246.0     retweets: 39405.0     Twitter Web App: 8493      Twitter for iPhone: 7057      Twitter for Android: 6074
Both Candidates likes: 74694.0 retweets: 28591.0    Twitter Web App: 4548      Twitter for iPhone: 3066      Twitter for Android: 2512

```

سوال ۴:

```

2023-06-10 22:05:25,849 INFO mapreduce.Job: map 100% reduce 100%
2023-06-10 22:05:26,883 INFO mapreduce.Job: Job job_1686419840224_0013 completed successfully
2023-06-10 22:05:26,997 INFO mapreduce.Job: Counters: 54
File System Counters:

```

```

2023-06-10 22:05:26,998 INFO streaming.StreamJob: Output directory: /CC/s2222/outputs0
root@a70d919b772d:/CC/s2222# hdfs dfs -cat /CC/s2222/outputs0/part-00000
2023-06-10 22:06:53,277 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
New York:    Joe Biden: 0.03139802408986331    Donald Trump: 0.05075111652456354    Both Candidates: 0.09392339964812559    total: 7389
California:  Joe Biden: 0.014669685930885103    Donald Trump: 0.03741262183715664    Both Candidates: 0.057891109579600274    total: 10157
Florida:     Joe Biden: 0.03185053380782918    Donald Trump: 0.036298932384341634    Both Candidates: 0.11797153024911032    total: 5620
Texas       Joe Biden: 0.02122946280228909    Donald Trump: 0.03950526121469448    Both Candidates: 0.10522429388960679    total: 5417

```

سوال ۵:

```

2023-06-10 22:19:33,174 INFO mapreduce.Job: map 100% reduce 100%
2023-06-10 22:19:48,934 INFO mapreduce.Job: Job job_1686419840224_0014 completed successfully
2023-06-10 22:19:49,093 INFO mapreduce.Job: Counters: 54
File System Counters:

```

```

2023-06-10 22:19:49,093 INFO streaming.StreamJob: Output directory: /CC/Q3/outputs0
root@a70d919b772d:/CC/Q3# hdfs dfs -cat /CC/Q3/outputs0/part-00000
2023-06-10 22:20:36,751 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
New York:    Joe Biden: 0.028826671578559587    Donald Trump: 0.055442991342788725    Both Candidates: 0.08896666052680051    total: 10858
California:  Joe Biden: 0.014183107555232295    Donald Trump: 0.03636694244931357    Both Candidates: 0.06018728975361397    total: 10999

```

همانطور که مشاهده میشود تعداد بیشتری از توییت ها موقعیت یابی شده اند که یعنی شناسایی ایالت با استفاده از موقعیت جغرافیایی موفقیت آمیز تر بوده است.