ECC Lab 2 Report

Name: Samir Sanyal

Rajat Sawant

Introduction:

In this lab, we were tasked with creating a Hadoop cluster with two instances using the prebuilt image provided.

Objective:

Develop MapReduce programs to analyze a large log file (sample.log) and extract specific statistics

Output the top K IP addresses for each hour

Make the program work like a parallel search, accepting a time period parameter and outputting the top K IP addresses for that period.

Process:

We created a Hadoop cluster with two instances using the prebuilt image "E516-Hadoop-Image-V4" on Jetstream2. We used our own SSH public key this time. The cluster consisted of:

Name * lab2_team_14_master Random Name

Image: E516-Hadoop-Image-V4

Flavor

General-purpose

Name	CPUs	RAM	Root Disk	Ephemeral Disk
m3.tiny	1	3 GB	20 GB	none
m3.small	2	6 GB	20 GB	none
m3.quad	4	15 GB	20 GB	none
m3.medium	8	30 GB	60 GB	none
m3.large	16	60 GB	60 GB	none
m3.xl	32	125 GB	60 GB	none
m3.2xl	64	250 GB	60 GB	none

Partial GPU

Name	CPUs	RAM	Root Disk	Ephemeral Disk
g3.small	4	15 GB	60 GB	none
g3.medium	8	30 GB	60 GB	none
g3.large	16	60 GB	60 GB	none

Full GPU

Name	CPUs	RAM	Root Disk	Ephemeral Disk
g3.xl	32	117 GB	60 GB	none

Choose a root disk size

- 20 GB (default for selected flavor)
- Custom disk size (volume-backed)

How many Instances?

Your quota supports up to 125 of these. Exosphere can create up to 25 at a time.

1

Enable web desktop?

No Yes

Choose an SSH public key

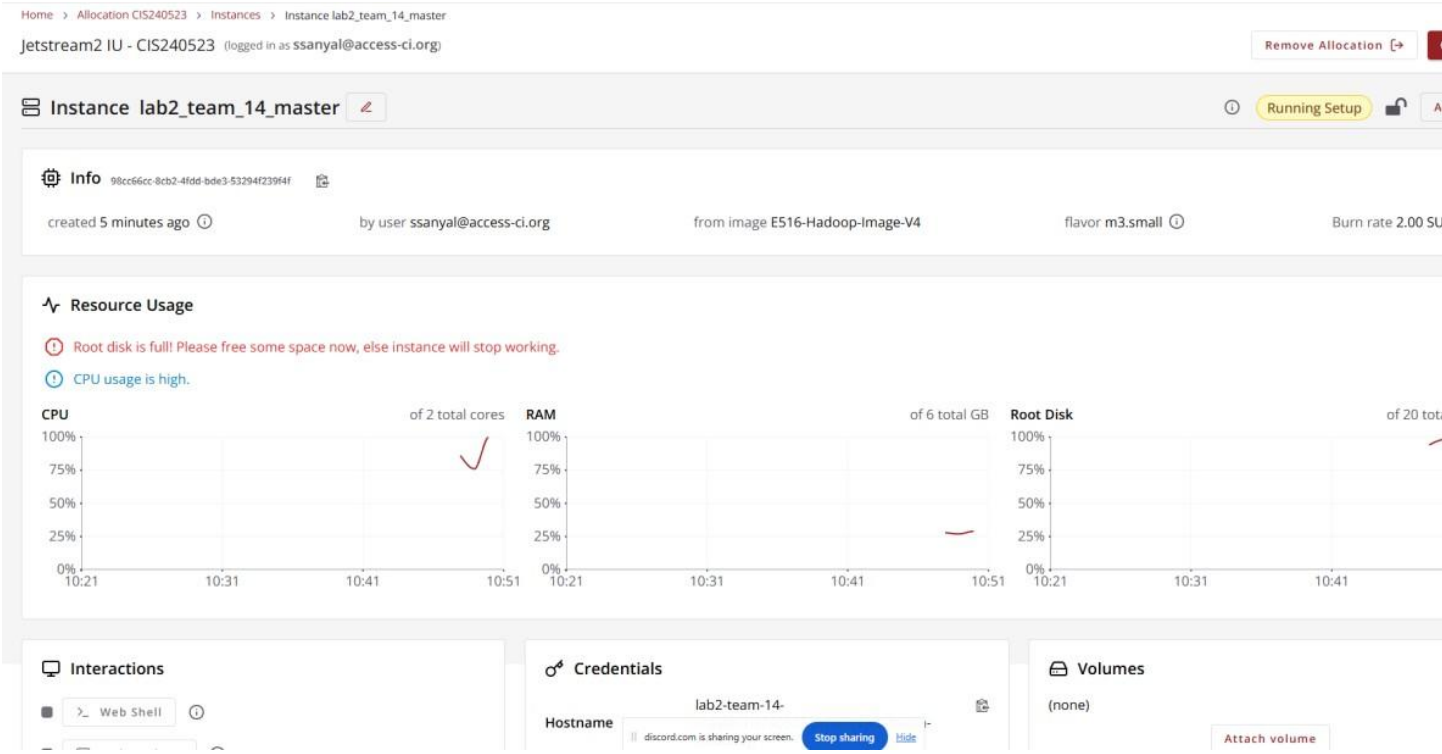
- None
- my_ssh_key

Upload a new SSH public key

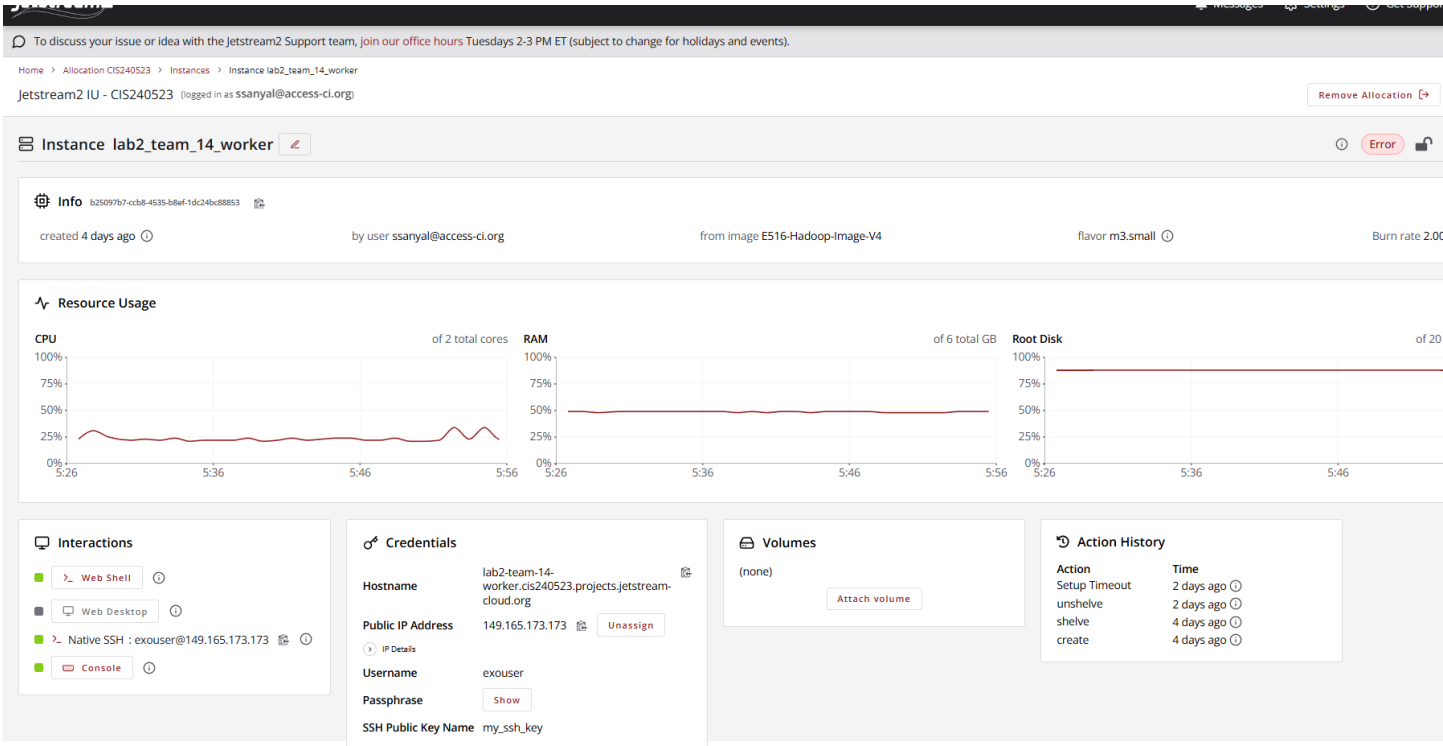
Advanced Options

Hide Show

Master Node: node-master



Worker Node: node-worker1



As instructed we first analyzed and understood the readme file to resolve and configure the Hadoop. We need to change the hostname and add the private IP's.

```
exouser@node-master:~$ cat How2Customize.README
This image assumes there are 3 instances. They are named node-master, node-worker1, and node-worker2.

After you create a number of instances based on this image, you need to do a few configurations. See th

On all instances:
===================================
Edit file /etc/hosts:
   [Private IP] node-master
   [Private IP] node-worker1
   [Private IP] node-worker2


Edit file /etc/hostname:
   Give the current instance a name as same as shown above.



On node-master only:
============================================
In order to monitor your HDFS(9870) and YARN(8088) status from a browser:
   $sudo ufw allow from [your-laptop-ip-address]
   Note: go to https://whatismyipaddress.com to see your IP address.


On all instances:
==========================
   $sudo ufw allow from 10.3.34.0/24
   Note: The above "10.3.34" should be the same as your instance's Internal IP address. If not, change


To customize your workers (e.g., you'd like to reduce 2 workers to 1 worker):
Edit file hadoop-3.4.0/etc/hadoop/workers



Remark:
=================================================================
*For the first time HDFS setup, remember to format your namenode (once!).

*If finding any problems, please let the Intructor and AI/TA know.

*ufw command format: ufw allow  from # port #   to #  port #
```

The /etc/hosts file was configured to ensure proper communication between the nodes:

exouser@lab2-team-14-worker: ~

```
DELL@DESKTOP-ODNUD4G MINGW64 ~
$ ssh exouser@149.165.173.173
The authenticity of host '149.165.173.173 (149.165.173.173)' can't be established.
ED25519 key fingerprint is SHA256:RKQBzxE5APJbzSVTdVe/opqvarD+GX1vprzLXKlidFI.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '149.165.173.173' (ED25519) to the list of known hosts.

 System information as of Fri Oct 11 15:42:03 UTC 2024

  System load:  0.52               Processes:             246
  Usage of /:   53.8% of 19.20GB   Users logged in:       0
  Memory usage: 40%                IPv4 address for ens3: 10.3.34.14
  Swap usage:   0%

=========================https://jetstream.status.io/=========================

Overall Jetstream2 Status:   Operational

Active Status Items:
 .   Scarce Availability of g3.large Instances


=============================================================================

Last login: Tue Feb 25 22:47:31 2025
exouser@lab2-team-14-worker:~$ ls -ltr
total 40
drwxr-xr-x 10 exouser exouser 4096 Mar  4  2024 hadoop-3.4.0
drwxr-xr-x  2 exouser exouser 4096 Oct 11 11:36 Videos
drwxr-xr-x  2 exouser exouser 4096 Oct 11 11:36 Templates
drwxr-xr-x  2 exouser exouser 4096 Oct 11 11:36 Public
drwxr-xr-x  2 exouser exouser 4096 Oct 11 11:36 Pictures
drwxr-xr-x  2 exouser exouser 4096 Oct 11 11:36 Music
drwxr-xr-x  2 exouser exouser 4096 Oct 11 11:36 Downloads
drwxr-xr-x  2 exouser exouser 4096 Oct 11 11:36 Documents
drwxr-xr-x  2 exouser exouser 4096 Oct 11 11:36 Desktop
-rw-rw-r--  1 exouser exouser 1391 Oct 24 23:06 How2Customize.README
exouser@lab2-team-14-worker:~$
```

exouser@lab2-team-14-worker: ~

```
exouser@lab2-team-14-worker:~$ cat /etc/hosts
127.0.0.1 localhost
10.3.34.203 node-master
10.3.34.225 node-worker1
149.165.169.180 node-worker2

# The following lines are desirable for IPv6 capable hosts
::1 ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
ff02::3 ip6-allhosts
exouser@lab2-team-14-worker:~$
```

We also configured the ufw to allow connectivity from our local machine to the instances and between the 2 instances. We used ifconfig to get the private IP's.

```
exouser@node-master:~$ cat /etc/hostname
node-master
exouser@node-master:~$ cat /etc/hosts
127.0.0.1 localhost
10.3.34.203 node-master
10.3.34.225 node-worker1
149.165.169.180 node-worker2

# The following lines are desirable for IPv6 capable hosts
::1 ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
ff02::3 ip6-allhosts
exouser@node-master:~$
```

```
exouser@node-worker1:~$ hostname
node-worker1
exouser@node-worker1:~$ cat /etc/hostname
node-worker1
exouser@node-worker1:~$ cat /etc/hosts
127.0.0.1 localhost
10.3.34.203 node-master
10.3.34.225 node-worker1
149.165.169.180 node-worker2

# The following lines are desirable for IPv6 capable hosts
::1 ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
ff02::3 ip6-allhosts
exouser@node-worker1:~$ |
```

```
exouser@lab2-team-14-master:~$ sudo ufw allow from 10.3.34.0/24
Rule added
exouser@lab2-team-14-master:~$ |
```

```
exouser@node-master:~$ sudo ufw status verbose
Status: active
Logging: on (low)
Default: deny (incoming), allow (outgoing), deny (routed)
New profiles: skip

To                         Action      From
--                         ------      ----
22/tcp (OpenSSH)           ALLOW IN    Anywhere
22/tcp                     ALLOW IN    Anywhere
49528/tcp                  ALLOW IN    Anywhere
Anywhere                   ALLOW IN    68.50.13.204
Anywhere                   ALLOW IN    10.3.34.0/24
8032                       ALLOW IN    10.3.34.225
22/tcp (OpenSSH (v6))      ALLOW IN    Anywhere (v6)
22/tcp (v6)                ALLOW IN    Anywhere (v6)
49528/tcp (v6)             ALLOW IN    Anywhere (v6)
```

Made this one change as we are using only 1 worker node.

```
exouser@node-master:~$ cat hadoop-3.4.0/etc/hadoop/workers
node-worker1
#node-worker2
exouser@node-master:~$ |
```

After setting up the cluster, we ensured that Hadoop was properly configured and running. We used the command hdfs dfsadmin -report to verify the same. The output confirmed that the cluster was running with the expected capacity and that the data nodes were active. We faced some issues as the datanode was not getting registered which was causing our cluster to fail. After lots of troubleshooting steps we were finally able to resolve it. We also formatted the namenode before using it.

exouser@lab2-team-14-master: ~

```
2025-02-25 23:31:40,793 INFO namenode.FSNamesystem: Detailed lock hold time metrics enabled: false
2025-02-25 23:31:40,826 INFO namenode.FSNamesystem: fsOwner                = exouser (auth:SIMPLE)
2025-02-25 23:31:40,828 INFO namenode.FSNamesystem: supergroup             = supergroup
2025-02-25 23:31:40,828 INFO namenode.FSNamesystem: isPermissionEnabled    = true
2025-02-25 23:31:40,828 INFO namenode.FSNamesystem: isStoragePolicyEnabled = true
2025-02-25 23:31:40,828 INFO namenode.FSNamesystem: HA Enabled: false
2025-02-25 23:31:40,915 INFO common.Util: dfs.datanode.fileio.profiling.sampling.percentage set to 0. Disabling file IO profiling
2025-02-25 23:31:41,136 INFO blockmanagement.DatanodeManager: Slow peers collection thread shutdown
2025-02-25 23:31:41,152 INFO blockmanagement.DatanodeManager: dfs.block.invalidate.limit : configured=1000, counted=60, effected=1000
2025-02-25 23:31:41,152 INFO blockmanagement.DatanodeManager: dfs.namenode.datanode.registration.ip-hostname-check=true
2025-02-25 23:31:41,154 INFO blockmanagement.BlockManager: dfs.namenode.startup.delay.block.deletion.sec is set to 000:00:00:00.000
2025-02-25 23:31:41,155 INFO blockmanagement.BlockManager: The block deletion will start around 2025 Feb 25 23:31:41
2025-02-25 23:31:41,156 INFO util.GSet: Computing capacity for map BlocksMap
2025-02-25 23:31:41,156 INFO util.GSet: VM type       = 64-bit
2025-02-25 23:31:41,157 INFO util.GSet: 2.0% max memory 1.4 GB = 29.6 MB
2025-02-25 23:31:41,157 INFO util.GSet: capacity      = 2^22 = 4194304 entries
2025-02-25 23:31:41,194 INFO blockmanagement.BlockManager: Storage policy satisfier is disabled
2025-02-25 23:31:41,194 INFO blockmanagement.BlockManager: dfs.block.access.token.enable = false
2025-02-25 23:31:41,201 INFO blockmanagement.BlockManagerSafeMode: Using 1000 as SafeModeMonitor Interval
2025-02-25 23:31:41,201 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.threshold-pct = 0.999
2025-02-25 23:31:41,201 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.min.datanodes = 0
2025-02-25 23:31:41,201 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.extension = 30000
2025-02-25 23:31:41,203 INFO blockmanagement.BlockManager: defaultReplication         = 1
2025-02-25 23:31:41,203 INFO blockmanagement.BlockManager: maxReplication             = 512
2025-02-25 23:31:41,203 INFO blockmanagement.BlockManager: minReplication             = 1
2025-02-25 23:31:41,203 INFO blockmanagement.BlockManager: maxReplicationStreams      = 2
2025-02-25 23:31:41,204 INFO blockmanagement.BlockManager: redundancyRecheckInterval  = 3000ms
2025-02-25 23:31:41,204 INFO blockmanagement.BlockManager: encryptDataTransfer        = false
2025-02-25 23:31:41,204 INFO blockmanagement.BlockManager: maxNumBlocksToLog          = 1000
2025-02-25 23:31:41,263 INFO namenode.FSDirectory: GLOBAL serial map: bits=29 maxEntries=536870911
2025-02-25 23:31:41,263 INFO namenode.FSDirectory: USER serial map: bits=24 maxEntries=16777215
2025-02-25 23:31:41,263 INFO namenode.FSDirectory: GROUP serial map: bits=24 maxEntries=16777215
2025-02-25 23:31:41,263 INFO namenode.FSDirectory: XATTR serial map: bits=24 maxEntries=16777215
2025-02-25 23:31:41,279 INFO util.GSet: Computing capacity for map INodeMap
2025-02-25 23:31:41,279 INFO util.GSet: VM type       = 64-bit
2025-02-25 23:31:41,279 INFO util.GSet: 1.0% max memory 1.4 GB = 14.8 MB
2025-02-25 23:31:41,280 INFO util.GSet: capacity      = 2^21 = 2097152 entries
2025-02-25 23:31:41,301 INFO namenode.FSDirectory: ACLs enabled? true
2025-02-25 23:31:41,301 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2025-02-25 23:31:41,301 INFO namenode.FSDirectory: XAttrs enabled? true
2025-02-25 23:31:41,301 INFO namenode.NameNode: Caching file names occurring more than 10 times
2025-02-25 23:31:41,305 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDescendant:
2025-02-25 23:31:41,305 INFO snapshot.SnapshotManager: dfs.namenode.snapshot.deletion.ordered = false
2025-02-25 23:31:41,307 INFO snapshot.SnapshotManager: SkipList is disabled
2025-02-25 23:31:41,310 INFO util.GSet: Computing capacity for map cachedBlocks
2025-02-25 23:31:41,310 INFO util.GSet: VM type       = 64-bit
2025-02-25 23:31:41,310 INFO util.GSet: 0.25% max memory 1.4 GB = 3.7 MB
2025-02-25 23:31:41,310 INFO util.GSet: capacity      = 2^19 = 524288 entries
2025-02-25 23:31:41,328 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2025-02-25 23:31:41,331 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2025-02-25 23:31:41,331 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2025-02-25 23:31:41,335 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2025-02-25 23:31:41,337 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2025-02-25 23:31:41,341 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2025-02-25 23:31:41,341 INFO util.GSet: VM type       = 64-bit
2025-02-25 23:31:41,341 INFO util.GSet: 0.029999999329447746% max memory 1.4 GB = 455.3 KB
2025-02-25 23:31:41,341 INFO util.GSet: capacity      = 2^16 = 65536 entries
2025-02-25 23:31:41,373 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1443718422-10.3.34.203-1740544301362
2025-02-25 23:31:41,404 INFO common.Storage: Storage directory /home/exouser/hadoop-3.4.0/data/nameNode has been successfully formatted.
2025-02-25 23:31:41,466 INFO namenode.FSImageFormatProtobuf: Saving image file /home/exouser/hadoop-3.4.0/data/nameNode/current/fsimage.ckpt_0000000000000000000 using no c
2025-02-25 23:31:41,641 INFO namenode.FSImageFormatProtobuf: Image file /home/exouser/hadoop-3.4.0/data/nameNode/current/fsimage.ckpt_0000000000000000000 of size 402 bytes
2025-02-25 23:31:41,674 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2025-02-25 23:31:41,679 INFO blockmanagement.DatanodeManager: Slow peers collection thread shutdown
2025-02-25 23:31:41,712 INFO namenode.FSNamesystem: Stopping services started for active state
2025-02-25 23:31:41,712 INFO namenode.FSNamesystem: Stopping services started for standby state
2025-02-25 23:31:41,721 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2025-02-25 23:31:41,721 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at lab2-team-14-master/10.3.34.203
************************************************************/
exouser@lab2-team-14-master:~$
```

exouser@node-master: ~

```
exouser@node-master:~$ hdfs dfsadmin -report
Configured Capacity: 20617822208 (19.20 GB)
Present Capacity: 2532724736 (2.36 GB)
DFS Remaining: 2521133056 (2.35 GB)
DFS Used: 11591680 (11.05 MB)
DFS Used%: 0.46%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (1):

Name: 10.3.34.225:9866 (node-worker1)
Hostname: node-worker1
Decommission Status : Normal
Configured Capacity: 20617822208 (19.20 GB)
DFS Used: 11591680 (11.05 MB)
Non DFS Used: 18068320256 (16.83 GB)
DFS Remaining: 2521133056 (2.35 GB)
DFS Used%: 0.06%
DFS Remaining%: 12.23%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Mar 02 18:35:06 EST 2025
Last Block Report: Sun Mar 02 16:58:03 EST 2025
Num of Blocks: 99
```

```
exouser@node-worker1: ~
exouser@node-worker1:~$ hdfs dfsadmin -report
Configured Capacity: 20617822208 (19.20 GB)
Present Capacity: 2532691968 (2.36 GB)
DFS Remaining: 2521100288 (2.35 GB)
DFS Used: 11591680 (11.05 MB)
DFS Used%: 0.46%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (1):

Name: 10.3.34.225:9866 (node-worker1)
Hostname: node-worker1
Decommission Status : Normal
Configured Capacity: 20617822208 (19.20 GB)
DFS Used: 11591680 (11.05 MB)
Non DFS Used: 18068353024 (16.83 GB)
DFS Remaining: 2521100288 (2.35 GB)
DFS Used%: 0.06%
DFS Remaining%: 12.23%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Mar 02 18:35:15 EST 2025
Last Block Report: Sun Mar 02 16:58:03 EST 2025
Num of Blocks: 99


exouser@node-worker1:~$
```

Used jps to check that all the processes are running on both master and worker node.

```
exouser@node-master:~$ jps
130593 SecondaryNameNode
131011 ResourceManager
1475 Bootstrap
316545 Jps
130104 NameNode
exouser@node-master:~$
```

```
exouser@node-worker1:~$ jps
259605 Jps
133386 DataNode
128699 NodeManager
1436 Bootstrap
exouser@node-worker1:~$ |
```

We used the hdfs command to create the directories and then put the sample.log file.

```
exouser@node-master: ~
exouser@node-master:~$ hdfs dfs -mkdir -p /user/hduser/logs
exouser@node-master:~$ hdfs dfs -ls /user/hduser/
Found 1 items
drwxr-xr-x   - exouser supergroup          0 2025-02-27 20:55 /user/hduser/logs
exouser@node-master:~$ |
```

```
exouser@node-master:~$ hdfs dfs -put sample.log /user/hduser/logs/
exouser@node-master:~$ hdfs dfs -ls /user/hduser/logs/
Found 1 items
-rw-r--r--   1 exouser supergroup     102400 2025-02-27 20:55 /user/hduser/logs/sample.log
exouser@node-master:~$ |
```

This is one of the error which we were facing which we were able to resolve.

```
exouser@lab2-team-14-master: ~
exouser@lab2-team-14-master:~$ hdfs dfs -ls /user/hduser/
Found 1 items
drwxr-xr-x   - exouser supergroup          0 2025-02-25 23:44 /user/hduser/logs
exouser@lab2-team-14-master:~$ hdfs dfs -put sample.log /user/hduser/logs/
2025-02-25 23:48:52,672 WARN hdfs.DataStreamer: Exception in createBlockOutputStream blk_1073741826_1002
java.io.EOFException: Unexpected EOF while trying to read response from server
        at org.apache.hadoop.hdfs.protocolPB.PBHelperClient.vintPrefixed(PBHelperClient.java:529)
        at org.apache.hadoop.hdfs.DataStreamer.createBlockOutputStream(DataStreamer.java:1905)
        at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1822)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:751)
2025-02-25 23:48:52,674 WARN hdfs.DataStreamer: Abandoning BP-1443718422-10.3.34.203-1740544301362:blk_1073741826_1002
2025-02-25 23:48:52,680 WARN hdfs.DataStreamer: Excluding datanode DatanodeInfoWithStorage[10.3.34.225:9866,DS-4c24022a-715c-49c2-baa7-4538539e3c81,DISK]
2025-02-25 23:48:52,732 WARN hdfs.DataStreamer: DataStreamer Exception
org.apache.hadoop.ipc.RemoteException(java.io.IOException): File /user/hduser/logs/sample.log._COPYING_ could only be written to 0 of the 1 minReplication nodes. There are 1 datanode(s) running and 1 node(s) are excluded in this operation.
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.chooseTarget4NewBlock(BlockManager.java:2473)
        at org.apache.hadoop.hdfs.server.namenode.FSDirWriteFileOp.chooseTargetForNewBlock(FSDirWriteFileOp.java:293)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getAdditionalBlock(FSNamesystem.java:3075)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.addBlock(NameNodeRpcServer.java:932)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.addBlock(ClientNamenodeProtocolServerSideTranslatorPB.java:603)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:621)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:589)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:573)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:1227)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:1246)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:1169)
        at java.base/java.security.AccessController.doPrivileged(Native Method)
        at java.base/javax.security.auth.Subject.doAs(Subject.java:423)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1953)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:3203)

        at org.apache.hadoop.ipc.Client.getRpcResponse(Client.java:1584)
        at org.apache.hadoop.ipc.Client.call(Client.java:1529)
        at org.apache.hadoop.ipc.Client.call(Client.java:1426)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Invoker.invoke(ProtobufRpcEngine2.java:258)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Invoker.invoke(ProtobufRpcEngine2.java:139)
        at com.sun.proxy.$Proxy12.addBlock(Unknown Source)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.lambda$addBlock$11(ClientNamenodeProtocolTranslatorPB.java:500)
        at org.apache.hadoop.ipc.internal.ShadedProtobufHelper.ipc(ShadedProtobufHelper.java:160)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.addBlock(ClientNamenodeProtocolTranslatorPB.java:500)
        at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at java.base/jdk.internal.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.base/java.lang.reflect.Method.invoke(Method.java:566)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:437)
        at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invokeMethod(RetryInvocationHandler.java:170)
        at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invoke(RetryInvocationHandler.java:162)
        at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invokeOnce(RetryInvocationHandler.java:100)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:366)
        at com.sun.proxy.$Proxy13.addBlock(Unknown Source)
        at org.apache.hadoop.hdfs.DFSOutputStream.addBlock(DFSOutputStream.java:1143)
        at org.apache.hadoop.hdfs.DataStreamer.locateFollowingBlock(DataStreamer.java:2009)
        at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1811)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:751)
put: File /user/hduser/logs/sample.log._COPYING_ could only be written to 0 of the 1 minReplication nodes. There are 1 datanode(s) running and 1 node(s) are excluded in this operation.
exouser@lab2-team-14-master:~$
```

The goal of **Part 1** was to count the number of visits from each IP address and output the top K IP addresses . The following steps were taken:

Map Stage: The input log file (sample.log) was processed line by line. Each line was parsed to extract the IP address and the hour from the timestamp. The output of the Map stage was a key-value pair in the format: (IP_Hour, 1). This **mapper.py** script is part of the Hadoop MapReduce program and is responsible for reading each line of the log file, extracting the **IP address** and **hour** from the timestamp, and emitting them as key-value pairs.

```
exouser@node-master:~$ cat mapper.py
#!/usr/bin/env python3
import sys
import re

for line in sys.stdin:
    parts = line.strip().split(" ")
    if len(parts) < 4:
        continue
    ip = parts[0]  # Extract IP address
    timestamp = parts[3].strip("[").split(":")  # Extract timestamp
    if len(timestamp) >= 2:
        hour = timestamp[1]  # Extract hour
        print(f"{ip} {hour}\t1")  # Emit (IP Hour, 1)
```

Reduce Stage: The Reduce stage aggregated the counts for each (IP_Hour) key. The output was a list of IP addresses with their corresponding visit counts for each hour. This reducer script is designed for a Hadoop Streaming job to calculate the **Top K IP addresses grouped by hour** based on their frequency in the log data.

```
exouser@node-master:~$ cat reducer_k.py
import sys
import os
import heapq
from collections import defaultdict

# Read K from environment variable (default to 5 if not set)
K = int(os.getenv("K", 5))

# Dictionary to store counts per (hour, IP)
ip_hour_count = defaultdict(lambda: defaultdict(int))

# Process input from Hadoop streaming
for line in sys.stdin:
    line = line.strip()
    if not line:
        continue

    key, count = line.split("\t")
    ip, hour = key.split()
    ip_hour_count[hour][ip] += int(count)

# Collect results and sort them
result_list = []
for hour, ip_counts in ip_hour_count.items():
    top_k_ips = heapq.nlargest(K, ip_counts.items(), key=lambda x: x[1])
    for ip, count in top_k_ips:
        result_list.append((count, ip, hour))  # Store as tuple (Count, IP, Hour)

# Sort results by count in descending order
result_list.sort(reverse=True, key=lambda x: x[0])

# Print the formatted output
print("Count\tIP\tHour")
for count, ip, hour in result_list:
    print(f"{count}\t{ip}\t{hour}")
```

Top K Selection: After the Reduce stage, the results were sorted by the number of visits, and the top K IP addresses for each hour were selected.

exouser@node-master: ~

```
exouser@node-master:~$ hadoop jar /home/exouser/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
-files /home/exouser/mapper.py,/home/exouser/reducer.py \
-input /user/hduser/logs/sample.log \
-output /user/hduser/output \
-mapper "/usr/bin/env python3 mapper.py" \
-reducer "/usr/bin/env python3 reducer.py"
packageJobJar: [/tmp/hadoop-unjar1395001736168815358 2/] [] /tmp/streamjob1240022287147765577 0.jar tmpDir=null
2025-02-27 21:30:35,043 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at node-master/10.3.34.203:8032
2025-02-27 21:30:35,174 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at node-master/10.3.34.203:8032
2025-02-27 21:30:35,377 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/exouser/.staging/job_1740706676773_0008
2025-02-27 21:30:36,081 INFO mapred.FileInputFormat: Total input files to process : 1
2025-02-27 21:30:36,122 INFO mapreduce.JobSubmitter: number of splits:2
2025-02-27 21:30:36,707 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1740706676773_0008
2025-02-27 21:30:36,707 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-02-27 21:30:36,838 INFO conf.Configuration: resource-types.xml not found
2025-02-27 21:30:36,838 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-02-27 21:30:36,884 INFO impl.YarnClientImpl: Submitted application application_1740706676773_0008
2025-02-27 21:30:36,904 INFO mapreduce.Job: The url to track the job: http://node-master:8088/proxy/application_1740706676773_0008/
2025-02-27 21:30:36,905 INFO mapreduce.Job: Running job: job_1740706676773_0008
2025-02-27 21:30:42,027 INFO mapreduce.Job: Job job_1740706676773_0008 running in uber mode : false
2025-02-27 21:30:42,028 INFO mapreduce.Job:  map 0% reduce 0%
2025-02-27 21:30:49,103 INFO mapreduce.Job:  map 100% reduce 0%
2025-02-27 21:30:54,131 INFO mapreduce.Job:  map 100% reduce 100%
2025-02-27 21:30:55,143 INFO mapreduce.Job: Job job_1740706676773_0008 completed successfully
2025-02-27 21:30:55,210 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=6459
                FILE: Number of bytes written=950567
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=106702
                HDFS: Number of bytes written=853
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=17468
                Total time spent by all reduces in occupied slots (ms)=4980
                Total time spent by all map tasks (ms)=8734
                Total time spent by all reduce tasks (ms)=2490
                Total vcore-milliseconds taken by all map tasks=8734
                Total vcore-milliseconds taken by all reduce tasks=2490
                Total megabyte-milliseconds taken by all map tasks=2235904
                Total megabyte-milliseconds taken by all reduce tasks=637440
        Map-Reduce Framework
                Map input records=320
                Map output records=319
                Map output bytes=5815
                Map output materialized bytes=6465
                Input split bytes=206
                Combine input records=0
                Combine output records=0
                Reduce input groups=46
                Reduce shuffle bytes=6465
                Reduce input records=319
                Reduce output records=46
                Spilled Records=638
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=131
                CPU time spent (ms)=1920
                Physical memory (bytes) snapshot=763154432
                Virtual memory (bytes) snapshot=6169100288
                Total committed heap usage (bytes)=513802240
                Peak Map Physical memory (bytes)=295010304
```

```
        File System Counters
                FILE: Number of bytes read=6459
                FILE: Number of bytes written=950567
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=106702
                HDFS: Number of bytes written=853
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=17468
                Total time spent by all reduces in occupied slots (ms)=4980
                Total time spent by all map tasks (ms)=8734
                Total time spent by all reduce tasks (ms)=2490
                Total vcore-milliseconds taken by all map tasks=8734
                Total vcore-milliseconds taken by all reduce tasks=2490
                Total megabyte-milliseconds taken by all map tasks=2235904
                Total megabyte-milliseconds taken by all reduce tasks=637440
        Map-Reduce Framework
                Map input records=320
                Map output records=319
                Map output bytes=5815
                Map output materialized bytes=6465
                Input split bytes=206
                Combine input records=0
                Combine output records=0
                Reduce input groups=46
                Reduce shuffle bytes=6465
                Reduce input records=319
                Reduce output records=46
                Spilled Records=638
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=131
                CPU time spent (ms)=1920
                Physical memory (bytes) snapshot=763154432
                Virtual memory (bytes) snapshot=6169100288
                Total committed heap usage (bytes)=513802240
                Peak Map Physical memory (bytes)=295010304
                Peak Map Virtual memory (bytes)=2058031104
                Peak Reduce Physical memory (bytes)=186777600
                Peak Reduce Virtual memory (bytes)=2053275648
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=106496
        File Output Format Counters
                Bytes Written=853
2025-02-27 21:30:55,210 INFO streaming.StreamJob: Output directory: /user/hduser/output
exouser@node-master:~$
```

```
exouser@node-master:~$ hdfs dfs -ls /user/hduser/output
Found 2 items
-rw-r--r--   1 exouser supergroup          0 2025-02-27 21:30 /user/hduser/output/_SUCCESS
-rw-r--r--   1 exouser supergroup        853 2025-02-27 21:30 /user/hduser/output/part-00000
exouser@node-master:~$ hdfs dfs -cat /user/hduser/output/part-00000
104.194.24.33 03        1
157.55.39.245 03        3
17.58.102.43 03 3
172.20.2.174 03 1
173.249.54.67 03        6
178.253.33.51 03        6
2.177.12.140 03 18
2.179.141.98 03 6
2.185.221.79 03 1
204.18.198.248 03       10
207.46.13.104 03        6
207.46.13.115 03        7
207.46.13.136 03        11
31.56.96.51 03  22
34.247.132.53 03        1
40.77.167.129 03        10
46.224.77.32 03 4
5.112.52.254 03 2
5.160.157.20 03 2
5.209.200.218 03        21
5.211.97.39 03  36
5.62.206.249 03 1
5.78.180.75 03  1
5.78.198.52 03  16
51.15.15.54 03  2
54.36.148.10 03 1
54.36.148.117 03        1
54.36.148.161 03        1
54.36.148.17 03 1
54.36.148.18 03 1
54.36.148.232 03        1
54.36.148.32 03 1
54.36.148.87 03 1
54.36.149.17 03 1
54.36.149.35 03 1
54.36.149.41 03 1
54.36.149.58 03 1
54.36.149.63 03 1
54.36.149.70 03 1
54.36.149.92 03 1
66.111.54.249 03        38
66.249.64.66 03 1
66.249.66.194 03        31
66.249.66.91 03 20
89.199.193.251 03       1
91.99.72.15 03  16
exouser@node-master:~$ |
```

exouser@node-master: ~

```
exouser@node-master:~$ hadoop jar /home/exouser/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
-files /home/exouser/mapper.py,/home/exouser/reducer_k.py \
-input /user/hduser/logs/sample.log \
-output /user/hduser/output \
-mapper "/usr/bin/env python3 mapper.py" \
-reducer "/usr/bin/env python3 reducer_k.py 5"
packageJobJar: [/tmp/hadoop-unjar7833611503750863041/] [] /tmp/streamjob10488770208325745237.jar tmpDir=null
2025-02-27 21:43:19,502 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at node-master/10.3.34.203:8032
2025-02-27 21:43:19,606 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at node-master/10.3.34.203:8032
2025-02-27 21:43:19,820 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/exouser/.staging/job_1740706676773_0011
2025-02-27 21:43:20,093 INFO mapred.FileInputFormat: Total input files to process : 1
2025-02-27 21:43:20,132 INFO mapreduce.JobSubmitter: number of splits:2
2025-02-27 21:43:20,685 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1740706676773_0011
2025-02-27 21:43:20,685 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-02-27 21:43:20,805 INFO conf.Configuration: resource-types.xml not found
2025-02-27 21:43:20,805 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-02-27 21:43:20,848 INFO impl.YarnClientImpl: Submitted application application_1740706676773_0011
2025-02-27 21:43:20,868 INFO mapreduce.Job: The url to track the job: http://node-master:8088/proxy/application_1740706676773_0011/
2025-02-27 21:43:20,869 INFO mapreduce.Job: Running job: job_1740706676773_0011
2025-02-27 21:43:26,935 INFO mapreduce.Job: Job job_1740706676773_0011 running in uber mode : false
2025-02-27 21:43:26,935 INFO mapreduce.Job:  map 0% reduce 0%
2025-02-27 21:43:33,008 INFO mapreduce.Job:  map 100% reduce 0%
2025-02-27 21:43:38,034 INFO mapreduce.Job:  map 100% reduce 100%
2025-02-27 21:43:40,048 INFO mapreduce.Job: Job job_1740706676773_0011 completed successfully
2025-02-27 21:43:40,114 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=6459
                FILE: Number of bytes written=950597
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=106702
                HDFS: Number of bytes written=101
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=17074
                Total time spent by all reduces in occupied slots (ms)=4570
                Total time spent by all map tasks (ms)=8537
                Total time spent by all reduce tasks (ms)=2285
                Total vcore-milliseconds taken by all map tasks=8537
                Total vcore-milliseconds taken by all reduce tasks=2285
                Total megabyte-milliseconds taken by all map tasks=2185472
                Total megabyte-milliseconds taken by all reduce tasks=584960
        Map-Reduce Framework
                Map input records=320
                Map output records=319
                Map output bytes=5815
                Map output materialized bytes=6465
                Input split bytes=206
                Combine input records=0
                Combine output records=0
                Reduce input groups=46
                Reduce shuffle bytes=6465
                Reduce input records=319
                Reduce output records=5
                Spilled Records=638
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=134
                CPU time spent (ms)=1800
                Physical memory (bytes) snapshot=775999488
                Virtual memory (bytes) snapshot=6171312128
                Total committed heap usage (bytes)=513802240
                Peak Map Physical memory (bytes)=300359680
```

```
exouser@node-master: ~
2025-02-27 21:43:20,848 INFO impl.YarnClientImpl: Submitted application application_1740706676773_0011
2025-02-27 21:43:20,868 INFO mapreduce.Job: The url to track the job: http://node-master:8088/proxy/application_1740706676773_0011/
2025-02-27 21:43:20,869 INFO mapreduce.Job: Running job: job_1740706676773_0011
2025-02-27 21:43:26,935 INFO mapreduce.Job: Job job_1740706676773_0011 running in uber mode : false
2025-02-27 21:43:26,935 INFO mapreduce.Job:  map 0% reduce 0%
2025-02-27 21:43:33,008 INFO mapreduce.Job:  map 100% reduce 0%
2025-02-27 21:43:38,034 INFO mapreduce.Job:  map 100% reduce 100%
2025-02-27 21:43:40,048 INFO mapreduce.Job: Job job_1740706676773_0011 completed successfully
2025-02-27 21:43:40,114 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=6459
                FILE: Number of bytes written=950597
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=106702
                HDFS: Number of bytes written=101
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=17074
                Total time spent by all reduces in occupied slots (ms)=4570
                Total time spent by all map tasks (ms)=8537
                Total time spent by all reduce tasks (ms)=2285
                Total vcore-milliseconds taken by all map tasks=8537
                Total vcore-milliseconds taken by all reduce tasks=2285
                Total megabyte-milliseconds taken by all map tasks=2185472
                Total megabyte-milliseconds taken by all reduce tasks=584960
        Map-Reduce Framework
                Map input records=320
                Map output records=319
                Map output bytes=5815
                Map output materialized bytes=6465
                Input split bytes=206
                Combine input records=0
                Combine output records=0
                Reduce input groups=46
                Reduce shuffle bytes=6465
                Reduce input records=319
                Reduce output records=5
                Spilled Records=638
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=134
                CPU time spent (ms)=1800
                Physical memory (bytes) snapshot=775999488
                Virtual memory (bytes) snapshot=6171312128
                Total committed heap usage (bytes)=513802240
                Peak Map Physical memory (bytes)=300359680
                Peak Map Virtual memory (bytes)=2058452992
                Peak Reduce Physical memory (bytes)=191688704
                Peak Reduce Virtual memory (bytes)=2058334208
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=106496
        File Output Format Counters
                Bytes Written=101
2025-02-27 21:43:40,114 INFO streaming.StreamJob: Output directory: /user/hduser/output
exouser@node-master:~$ |
```

```
exouser@node-master:~$ hdfs dfs -ls /user/hduser/output
Found 2 items
-rw-r--r--   1 exouser supergroup          0 2025-02-27 21:43 /user/hduser/output/_SUCCESS
-rw-r--r--   1 exouser supergroup        101 2025-02-27 21:43 /user/hduser/output/part-00000
exouser@node-master:~$ hdfs dfs -cat /user/hduser/output/part-00000
03 66.111.54.249 38
03 5.211.97.39 36
03 66.249.66.194 31
03 31.56.96.51 22
03 5.209.200.218 21
exouser@node-master:~$ |
```

After correct formatting so that it makes more sense.

```
2025-02-27 21:49:05,065 INFO streaming.StreamJob: Output directory: /user/hduser/output
exouser@node-master:~$ hdfs dfs -ls /user/hduser/output
Found 2 items
-rw-r--r--   1 exouser supergroup          0 2025-02-27 21:49 /user/hduser/output/_SUCCESS
-rw-r--r--   1 exouser supergroup        110 2025-02-27 21:49 /user/hduser/output/part-00000
exouser@node-master:~$ hdfs dfs -cat /user/hduser/output/part-00000
Count   IP      Hour
38      66.111.54.249   03
36      5.211.97.39     03
31      66.249.66.194   03
22      31.56.96.51     03
21      5.209.200.218   03
exouser@node-master:~$
```

Before starting the part 2 we deleted the output directory.

```
exouser@node-master:~$ hdfs dfs -rm -r /user/hduser/output
Deleted /user/hduser/output
exouser@node-master:~$
```

In **Part 2**, the program was extended to accept a time period parameter and output the top K IP addresses for that period. The following steps were taken:

Map Stage: The input log file was processed line by line. Each line was parsed to extract the IP address and the hour from the timestamp. If the hour fell within the specified time period, the output was a key-value pair in the format: (IP, 1). The mapper filters log entries within a specific time period and emits the IP addresses for those entries.

```
exouser@node-master:~$ cat mapper_deep.py
import os
import sys
import re

TIME_PERIOD = os.environ.get("TIME_PERIOD", "0-1")
start_hour, end_hour = map(int, TIME_PERIOD.split("-"))

def extract_hour(timestamp):
    match = re.search(r"\[(\d{2})/.*:(\d{2}):\d{2}:\d{2}", timestamp)
    if match:
        return int(match.group(2))
    return -1

data_found = False  # Track if any data is found

for line in sys.stdin:
    parts = line.strip().split()
    if len(parts) < 4:
        continue

    ip_address = parts[0]
    timestamp = parts[3]

    hour = extract_hour(timestamp)

    if hour == -1:
        print(f"DEBUG: Invalid timestamp format in line -> {line}", file=sys.stderr)
        continue

    if start_hour <= hour < end_hour:
        print(f"{ip_address}\t1")
        data_found = True
    else:
        print(f"DEBUG: Skipped IP {ip_address} at hour {hour}", file=sys.stderr)

if not data_found:
    print(f"WARNING: No matching data found for TIME_PERIOD={TIME_PERIOD}", file=sys.stderr)
```

Reduce Stage: The Reduce stage aggregated the counts for each IP address. The output was a list of IP addresses with their corresponding visit counts for the specified time period. The reducer_time.py script is part of the Hadoop Streaming job. It takes the (IP, 1) pairs emitted by the mapper and calculates the **Top K most frequent IP addresses** within the specified time period.

```
exouser@node-master:~$ cat reducer_time.py
#!/usr/bin/env python3
import sys
import os
from collections import defaultdict

# Read environment variable for K
K = int(os.environ.get("K", 5))  # Default K=5 if not provided

ip_counts = defaultdict(int)

# Read input from Mapper
for line in sys.stdin:
    ip, count = line.strip().split("\t")
    ip_counts[ip] += int(count)

# Sort IPs based on visit count (descending order)
sorted_ips = sorted(ip_counts.items(), key=lambda x: x[1], reverse=True)

# Print the top K IP addresses
for i in range(min(K, len(sorted_ips))):
    print(f"{sorted_ips[i][0]}\t{sorted_ips[i][1]}")
```

Top K Selection: The results were sorted by the number of visits, and the top K IP addresses for the specified time period were selected.

The same Hadoop Streaming command was used, but the mapper and reducer scripts were modified to accept the time period parameter

exouser@node-master: ~

```
exouser@node-master:~$ hadoop jar /home/exouser/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar -files /home/exouser/mapper_deep.py,/home/exouser/reducer_time.py -
mapper_deep.py" -reducer "/usr/bin/env python3 reducer_time.py" -cmdenv K=5 -cmdenv TIME_PERIOD=3-4
packageJobJar: [/tmp/hadoop-unjar5377916111452919906/] [] /tmp/streamjob1837559787500888706 5.jar tmpDir=null
2025-03-02 18:54:54,609 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at node-master/10.3.34.203:8032
2025-03-02 18:54:54,805 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at node-master/10.3.34.203:8032
2025-03-02 18:54:55,086 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/exouser/.staging/job_1740706676773_0035
2025-03-02 18:54:55,566 INFO mapred.FileInputFormat: Total input files to process : 1
2025-03-02 18:54:55,632 INFO mapreduce.JobSubmitter: number of splits:2
2025-03-02 18:54:55,928 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1740706676773_0035
2025-03-02 18:54:55,928 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-03-02 18:54:56,131 INFO conf.Configuration: resource-types.xml not found
2025-03-02 18:54:56,131 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-03-02 18:54:56,210 INFO impl.YarnClientImpl: Submitted application application_1740706676773_0035
2025-03-02 18:54:56,250 INFO mapreduce.Job: The url to track the job: http://node-master:8088/proxy/application_1740706676773_0035/
2025-03-02 18:54:56,251 INFO mapreduce.Job: Running job: job_1740706676773_0035
2025-03-02 18:55:02,339 INFO mapreduce.Job: Job job_1740706676773_0035 running in uber mode : false
2025-03-02 18:55:02,340 INFO mapreduce.Job:  map 0% reduce 0%
2025-03-02 18:55:10,434 INFO mapreduce.Job:  map 100% reduce 0%
2025-03-02 18:55:16,467 INFO mapreduce.Job:  map 100% reduce 100%
2025-03-02 18:55:16,475 INFO mapreduce.Job: Job job_1740706676773_0035 completed successfully
2025-03-02 18:55:16,551 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=5502
                FILE: Number of bytes written=948833
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=106702
                HDFS: Number of bytes written=81
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=24402
                Total time spent by all reduces in occupied slots (ms)=5614
                Total time spent by all map tasks (ms)=12201
                Total time spent by all reduce tasks (ms)=2807
                Total vcore-milliseconds taken by all map tasks=12201
                Total vcore-milliseconds taken by all reduce tasks=2807
                Total megabyte-milliseconds taken by all map tasks=3123456
                Total megabyte-milliseconds taken by all reduce tasks=718592
        Map-Reduce Framework
                Map input records=320
                Map output records=319
                Map output bytes=4858
                Map output materialized bytes=5508
                Input split bytes=206
                Combine input records=0
                Combine output records=0
                Reduce input groups=46
                Reduce shuffle bytes=5508
                Reduce input records=319
                Reduce output records=5
                Spilled Records=638
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=235
                CPU time spent (ms)=3320
                Physical memory (bytes) snapshot=799428608
                Virtual memory (bytes) snapshot=6161952768
                Total committed heap usage (bytes)=532676608
                Peak Map Physical memory (bytes)=303493120
                Peak Map Virtual memory (bytes)=2053038080
                Peak Reduce Physical memory (bytes)=195493888
                Peak Reduce Virtual memory (bytes)=2058575872
        Shuffle Errors
```

```
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=106496
        File Output Format Counters
                Bytes Written=81
2025-03-02 18:55:16,552 INFO streaming.StreamJob: Output directory: /user/hduser/output
exouser@node-master:~$
```

exouser@node-master: ~

```
exouser@node-master:~$ hdfs dfs -ls /user/hduser/output
Found 2 items
-rw-r--r--   1 exouser supergroup          0 2025-03-02 18:55 /user/hduser/output/_SUCCESS
-rw-r--r--   1 exouser supergroup         81 2025-03-02 18:55 /user/hduser/output/part-00000
exouser@node-master:~$ hdfs dfs -cat /user/hduser/output/part-00000
66.111.54.249   38
5.211.97.39     36
66.249.66.194   31
31.56.96.51     22
5.209.200.218   21
exouser@node-master:~$
```