Peer-graded Assignment: Distributed File Systems

You passed!
Congratulations. You earned 8 / 8 points. Review the feedback below and continue the course when you are ready. You can also help more classmates by reviewing their submissions.

Review Classmates' Work

Instructions

My submission

Peer-review DFS

Submitted on November 13, 2018

Discussions

Shareable Link

PROMPT

Estimate minimum Namenode RAM size for HDFS with 1 PB capacity, block size 64 MB, average metadata size for each block is 300 B, replication factor is 3. Provide the formula for calculations and the result.

=1000000000000/(64000000 * 3) * 300 = 1 562 500 000 b ~= 1.56Gb

RUBRIC

Check the formula and the result, they should both be correct.

1 PB / 64 MB / 3 * 300 B = 1024 * 1024 * 1024 / 64 / 3 * 300 = 1600 MB

The result may be not exactly the same, rounding and other units are possible. So 1600 MB, 1.6 GB, 1.56 GB are all allowed.

0 points
No

1 point
Yes

PROMPT

HDDs in your cluster have the following characteristics: average reading speed is 60 MB/s, seek time is 5 ms. You want to spend 0.5 % time for seeking the block, i.e. seek time should be 200 times less than the time to read the block. Estimate the minimum block size.

5 ms seek rate / 0.005 (i.e., 0.5% time for seeking block) = 1000 ms (1s) transfer time. In 1 second, given a transfer rate of 60 MB/s, a block size of 60MB can be read. 60 MB is the minimum block size.

RUBRIC

Check the calculations and the result, they should both be correct.

block_size / 60 MB/s * 0.5 / 100 >= 5 ms

block_size >= 60 MB/s * 0.005 s / 0.005 = 60 MB

So, the minimum block size is 60 MB or 64 MB.

0 points
No

1 point
Yes

PROMPT

To complete this task use the 'HDFS CLI Playground' item.

Create text file ‘test.txt’ in a local fs. Use HDFS CLI to make the following operations:

- create directory ‘assignment1’ in your home directory in HDFS (you can use a relative path or prescribe it explicitly "/user/jovyan/...")
- put test.txt in it
- output the size and the owner of the file
- revoke ‘read’ permission for ‘other users’
- read the first 10 lines of the file
- rename it to ‘test2.txt’.

Provide all the commands to HDFS CLI.

!hdfs dfs -mkdir ./assignment1

!hdfs dfs -copyFromLocal test.txt ./assignment1/test.txt

!hdfs dfs -ls -h ./assignment1/test.txt

!hdfs dfs -chmod 640 assignment1/test.txt

!hdfs dfs -cat assignment1/test.txt | head

!hdfs dfs -mv assignment1/test.txt assignment1/test2.txt

RUBRIC

Check the commands, they should be like these:

$ hdfs dfs -mkdir assignment1

$ hdfs dfs -put test.txt assignment1/

$ hdfs dfs -ls assignment1/test.txt ***or*** hdfs dfs -stat "%b %u" assignment1/test.txt

$ hdfs dfs -chmod o-r assignment1/test.txt

$ hdfs dfs -cat assignment1/test.txt | head -10

$ hdfs dfs -mv assignment1/test.txt assignment1/test2.txt

There can be the following differences:

- ‘hdfs dfs’ and ‘hadoop fs’ are the same
- absolute paths are also allowed: ‘/user/<username>/assignment1/test.txt’ instead of ‘assignment1/test.txt’

- the permissions argument can be in an octal form, like 0640
- the 'text' command can be used instead of 'cat'

0 points
0 commands are correct.

1 point
Less than 3 commands are correct.

2 points
Less than 6 commands are correct.

3 points
All the commands are correct.

PROMPT

To complete this task use the 'HDFS CLI Playground' item.

Use HDFS CLI to investigate the file '/data/wiki/en_articles_part/articles-part' in HDFS:

- get blocks and their locations in HDFS for this file, show the command without an output
- get the information about any block of the file, show the command and the block locations from the output

!hdfs fsck /data/wiki/en_articles_part/articles-part -files -blocks -locations

Using output from above command, I am using blk_1073741825 as input below:
!hdfs fsck -blockId blk_1073741825

Block locations from output would be:
Block belongs to: /data/wiki/en_articles_part/articles-part

RUBRIC

Blocks and locations of '/data/wiki/en_articles_part/articles-part':

```
1  $ hdfs fsck /data/wiki/en_articles_part/articles-part -files -blocks -locations
```

Block information (block id may be different):

```
$ hdfs fsck -blockId blk_1073971670
```

It outputs the block locations, example (nodes list will be different):

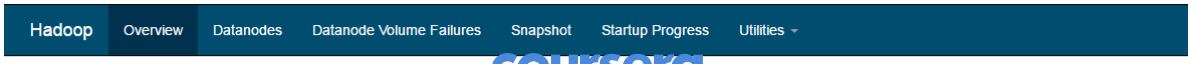
Block replica on datanode/rack: some_node_hostname/default-rack is HEALTHY

- 0 points
  No correct answers.
- 1 point
  Some answers are correct.
- **2 points**
  Everything is correct.

PROMPT

Look at the picture of Namenode web interface from a real Hadoop cluster.

Show the total capacity of this HDFS installation, used space and total data nodes in the cluster.



Overview '[illegible]:8020' (active)

Started:	Sun Jul 09 10:41:38 +0300 2017
Version:	2.6.0-cdh5.11.0, r91a488f2c5abb3de0e6ee74080dbc439c7576fb4
Compiled:	Thu Apr 06 06:07:00 +0300 2017 by jenkins from Unknown
Cluster ID:	cluster12
Block Pool ID:	BP-1135830572-93.175.29.106-1493426010892

Summary

Security is off.

Safemode is off.

3 601 files and directories, 1 735 blocks = 5 336 total filesystem object(s).

Heap Memory used 498.31 MB of 1.97 GB Heap Memory. Max Heap Memory is 1.97 GB.

Non Heap Memory used 62.02 MB of 62.31 MB Commited Non Heap Memory. Max Non Heap Memory is 130 MB.

Configured Capacity:	2.14 TB
DFS Used:	242.12 GB (11.03%)
Non DFS Used:	35.51 GB
DFS Remaining:	1.77 TB (82.45%)
Block Pool Used:	242.12 GB (11.03%)
DataNodes usages% (Min/Median/Max/stdDev):	3.49% / 14.60% / 14.63% / 4.55%
Live Nodes	4 (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	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	Sun Jul 09 10:41:38 +0300 2017
Last Checkpoint Time	Thu Jul 27 12:49:22 +0300 2017

Total capacity of this HDFS installation: 2.14TB
Used space: 242.12GB
Total data nodes in the cluster:4

RUBRIC

Total capacity: 2.14 TB

Used space: 242.12 GB (=DFS Used) or 242.12+35.51 = 277.63 GB (=DFS Used + Non DFS Used) - the latter answer is more precise, but the former is also possible

Data nodes in the cluster: 4

0 points
No

1 point
Yes

Edit submission

Comments

Comments left for the learner are visible only to that learner and the person who left the comment.

share your thoughts...