Hadoop HDFS / HDFS-5837

dfs.namenode.replication.considerLoad does not consider decommissioned nodes

Details

Type:	Bug	Status:	CLOSED
Priority:	Major	Resolution:	Fixed
Affects Version/s:	2.0.0-alpha, 2.0.6-alpha, ... (2)	Fix Version/s:	2.3.0
Component/s:	namenode		
Labels:	None		
Hadoop Flags:	Reviewed		

Description

In DefaultBlockPlacementPolicy, there is a setting dfs.namenode.replication.considerLoad which tries to balance the load of the cluster when choosing replica locations. This code does not take into account decommissioned nodes.

The code for considerLoad calculates the load by doing: TotalClusterLoad / numNodes. However, numNodes includes decommissioned nodes (which have 0 load). Therefore, the average load is artificially low. Example:

TotalLoad = 250
numNodes = 100
decommissionedNodes = 70
remainingNodes = numNodes - decommissionedNodes = 30

avgLoad = 250/100 = 2.50
trueAvgLoad = 250 / 30 = 8.33

If the real load of the remaining 30 nodes is (on average) 8.33, this is more than 2x the calculated average load of 2.50. This causes these nodes to be rejected as replica locations. The final result is that all nodes are rejected, and no replicas can be placed.

See exceptions printed from client during this scenario: https://gist.github.com/bbeaudreault/49c8aa4bb231de54e9c1

Attachments

HDFS-5837_B.patch	9 kB	06/Feb/14 20:10
HDFS-5837_branch_2.2.0.patch	10 kB	07/Feb/14 00:39
HDFS-5837_C_branch_2.2.0.patch	10 kB	10/Feb/14 19:00
HDFS-5837_C.patch	10 kB	08/Feb/14 06:52
HDFS-5837.patch	9 kB	05/Feb/14 23:51

Issue Links

breaks

HDFS-6599 2.4 addBlock is 10 to 20 times slower compared to 0.23	CLOSED	

Activity

Tao Luo added a comment - 06/Feb/14 00:02

Regarding the number of DNS that are counted, I am thinking of adding getNumDatanodesInService() to FSClusterStats to get the number of live and in service datanodes.

Hadoop QA added a comment - 06/Feb/14 03:28

-1 overall. Here are the results of testing the latest attachment
http://issues.apache.org/jira/secure/attachment/12627251/HDFS-5837.patch
against trunk revision .

+1 @author. The patch does not contain any @author tags.

+1 tests included. The patch appears to include 1 new or modified test files.

+1 javac. The applied patch does not increase the total number of javac compiler warnings.

+1 javadoc. There were no new javadoc warning messages.

+1 eclipse:eclipse. The patch built with eclipse:eclipse.

-1 findbugs. The patch appears to introduce 1 new Findbugs (version 1.3.9) warnings.

+1 release audit. The applied patch does not increase the total number of release audit warnings.

-1 core tests. The patch failed these unit tests in hadoop-hdfs-project/hadoop-hdfs:

org.apache.hadoop.hdfs.server.namenode.TestCacheDirectives

+1 contrib tests. The patch passed contrib unit tests.

Test results: https://builds.apache.org/job/PreCommit-HDFS-Build/6046//testReport/
Findbugs warnings: https://builds.apache.org/job/PreCommit-HDFS-Build/6046//artifact/trunk/patchprocess/newPatchFindbugsWarningshadoop-hdfs.html
Console output: https://builds.apache.org/job/PreCommit-HDFS-Build/6046//console

This message is automatically generated.

Hadoop QA added a comment - 06/Feb/14 19:53

-1 overall. Here are the results of testing the latest attachment
http://issues.apache.org/jira/secure/attachment/12627401/HDFS-5837_B.patch
against trunk revision .

-1 patch. Trunk compilation may be broken.

Console output: https://builds.apache.org/job/PreCommit-HDFS-Build/6059//console

This message is automatically generated.

Hadoop QA added a comment - 06/Feb/14 22:45

+1 overall. Here are the results of testing the latest attachment
http://issues.apache.org/jira/secure/attachment/12627409/HDFS-5837_B.patch
against trunk revision .

+1 @author. The patch does not contain any @author tags.

+1 tests included. The patch appears to include 1 new or modified test files.

+1 javac. The applied patch does not increase the total number of javac compiler warnings.

+1 javadoc. There were no new javadoc warning messages.

+1 eclipse:eclipse. The patch built with eclipse:eclipse.

+1 findbugs. The patch does not introduce any new Findbugs (version 1.3.9) warnings.

+1 release audit. The applied patch does not increase the total number of release audit warnings.

+1 core tests. The patch passed unit tests in hadoop-hdfs-project/hadoop-hdfs.

+1 contrib tests. The patch passed contrib unit tests.

Test results: https://builds.apache.org/job/PreCommit-HDFS-Build/6060//testReport/
Console output: https://builds.apache.org/job/PreCommit-HDFS-Build/6060//console

This message is automatically generated.

Hadoop QA added a comment - 07/Feb/14 01:25

-1 overall. Here are the results of testing the latest attachment
http://issues.apache.org/jira/secure/attachment/12627518/HDFS-5837_branch_2.2.0.patch

against trunk revision .

+1 @author. The patch does not contain any @author tags.

+1 tests included. The patch appears to include 1 new or modified test files.

-1 javac. The patch appears to cause the build to fail.

Console output: https://builds.apache.org/job/PreCommit-HDFS-Build/6062//console

This message is automatically generated.

Konstantin Shvachko added a comment - 07/Feb/14 23:39

Makes sense to me. Couple nits:

1. Add comment `// FSClusterStats` to `@ Override`
2. In the test it is better to set up and destroy NN using BeforeClass and AfterClass rather than in a try-catch block. As of now you will miss shutting down NN when an exception is thrown before try{}.

Tao Luo added a comment - 08/Feb/14 00:17

Thanks Konstantin.
~~HDFS-5837~~_C.patch addresses the above comments.

Hadoop QA added a comment - 08/Feb/14 02:44

-1 overall. Here are the results of testing the latest attachment
http://issues.apache.org/jira/secure/attachment/12627748/HDFS-5837_C.patch
against trunk revision .

+1 @author. The patch does not contain any @author tags.

+1 tests included. The patch appears to include 1 new or modified test files.

+1 javac. The applied patch does not increase the total number of javac compiler warnings.

+1 javadoc. There were no new javadoc warning messages.

+1 eclipse:eclipse. The patch built with eclipse:eclipse.

+1 findbugs. The patch does not introduce any new Findbugs (version 1.3.9) warnings.

+1 release audit. The applied patch does not increase the total number of release audit warnings.

-1 core tests. The patch failed these unit tests in hadoop-hdfs-project/hadoop-hdfs:

org.apache.hadoop.hdfs.server.namenode.TestCacheDirectives

+1 contrib tests. The patch passed contrib unit tests.

Test results: https://builds.apache.org/job/PreCommit-HDFS-Build/6084//testReport/
Console output: https://builds.apache.org/job/PreCommit-HDFS-Build/6084//console

This message is automatically generated.

Hadoop QA added a comment - 08/Feb/14 09:24

+1 overall. Here are the results of testing the latest attachment
http://issues.apache.org/jira/secure/attachment/12627785/HDFS-5837_C.patch
against trunk revision .

+1 @author. The patch does not contain any @author tags.

+1 tests included. The patch appears to include 1 new or modified test files.

+1 javac. The applied patch does not increase the total number of javac compiler warnings.

+1 javadoc. There were no new javadoc warning messages.

+1 eclipse:eclipse. The patch built with eclipse:eclipse.

+1 findbugs. The patch does not introduce any new Findbugs (version 1.3.9) warnings.

+1 release audit. The applied patch does not increase the total number of release audit warnings.

+1 core tests. The patch passed unit tests in hadoop-hdfs-project/hadoop-hdfs.

+1 contrib tests. The patch passed contrib unit tests.

Test results: https://builds.apache.org/job/PreCommit-HDFS-Build/6088//testReport/
Console output: https://builds.apache.org/job/PreCommit-HDFS-Build/6088//console

This message is automatically generated.

Konstantin Shvachko added a comment - 09/Feb/14 19:56

+1

Konstantin Shvachko added a comment - 09/Feb/14 20:55

I just committed this. Thank you Tao.

Hudson added a comment - 09/Feb/14 20:56

SUCCESS: Integrated in Hadoop-trunk-Commit #5134 (See https://builds.apache.org/job/Hadoop-trunk-Commit/5134/)
~~HDFS-5837~~. dfs.namenode.replication.considerLoad should consider decommissioned nodes. Contributed by Tao Luo. (shv: http://svn.apache.org/viewcvs.cgi/?root=Apache-SVN&view=rev&rev=1566410)

- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/CHANGES.txt
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/blockmanagement/BlockPlacementPolicyDefault.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/namenode/FSClusterStats.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/namenode/FSNamesystem.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/test/java/org/apache/hadoop/hdfs/server/blockmanagement/TestReplicationPolicyConsiderLoad.java

---

Bryan Beaudreault added a comment - 09/Feb/14 22:26

Thanks!

Hudson added a comment - 10/Feb/14 11:14

SUCCESS: Integrated in Hadoop-Yarn-trunk #477 (See https://builds.apache.org/job/Hadoop-Yarn-trunk/477/)
~~HDFS-5837~~. dfs.namenode.replication.considerLoad should consider decommissioned nodes. Contributed by Tao Luo. (shv: http://svn.apache.org/viewcvs.cgi/?root=Apache-SVN&view=rev&rev=1566410)

- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/CHANGES.txt
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/blockmanagement/BlockPlacementPolicyDefault.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/namenode/FSClusterStats.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/namenode/FSNamesystem.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/test/java/org/apache/hadoop/hdfs/server/blockmanagement/TestReplicationPolicyConsiderLoad.java

---

Hudson added a comment - 10/Feb/14 13:41

SUCCESS: Integrated in Hadoop-Hdfs-trunk #1669 (See https://builds.apache.org/job/Hadoop-Hdfs-trunk/1669/)
~~HDFS-5837~~. dfs.namenode.replication.considerLoad should consider decommissioned nodes. Contributed by Tao Luo. (shv: http://svn.apache.org/viewcvs.cgi/?root=Apache-SVN&view=rev&rev=1566410)

- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/CHANGES.txt
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/blockmanagement/BlockPlacementPolicyDefault.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/namenode/FSClusterStats.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/namenode/FSNamesystem.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/test/java/org/apache/hadoop/hdfs/server/blockmanagement/TestReplicationPolicyConsiderLoad.java

---

Hudson added a comment - 10/Feb/14 14:51

SUCCESS: Integrated in Hadoop-Mapreduce-trunk #1694 (See https://builds.apache.org/job/Hadoop-Mapreduce-trunk/1694/) ~~HDFS-5837~~. dfs.namenode.replication.considerLoad should consider decommissioned nodes. Contributed by Tao Luo. (shv: http://svn.apache.org/viewcvs.cgi/?root=Apache-SVN&view=rev&rev=1566410)

- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/CHANGES.txt
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/blockmanagement/BlockPlacementPolicyDefault.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/namenode/FSClusterStats.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/main/java/org/apache/hadoop/hdfs/server/namenode/FSNamesystem.java
- /hadoop/common/trunk/hadoop-hdfs-project/hadoop-hdfs/src/test/java/org/apache/hadoop/hdfs/server/blockmanagement/TestReplicationPolicyConsiderLoad.java

Tao Luo added a comment - 10/Feb/14 19:18

Thanks Konstantin!

People

Assignee:

Tao Luo

Reporter:

Bryan Beaudreault

Votes:

0 Vote for this issue

Watchers:

9 Start watching this issue

Dates

Created:

27/Jan/14 18:45

Updated:

12/May/16 18:18

Resolved:

09/Feb/14 20:55