TAJO-118

jhkim

2013-08-04

Contents

Chapter 1

Root issue TAJO-118

1.1 Summary

Refactor and Improve text file Scanner

1.2 Description

This will be reduce the execution time
* Implement column projection
* Change the split function to the apache StringUtils
* Improve slow codes

1.3 Attachments

1. TAJO-118_2.patch
2. TAJO-118_3.patch
3. TAJO-118.patch

1.4 Commits

1. Commit **301da59** by **jinossy** (2013-08-12): TAJO-118: Refactor and Improve text file Scanner. (fixed wrong index)
2. Commit **cf6bd4b** by **jinossy** (2013-08-12): TAJO-118: Refactor and Improve text file Scanner. (jinho)

1.5 Comments

1. **jhkim:** h5. Generated TPC-H DBGEN 100G

   * Before

   ```
   select count(*) from lineitem;
   total response time: 216.671 sec
   ```

* After

```
select count(*) from lineitem;
total response time: 146.245 sec
```

2. **jhkim:** I verified 'mvn clean install'.

3. **hyunsik:** That's really great work! Thank you for your contribution.

   However, the first patch has some bug in the case where some columns are pushed down into scan node. So, I added the unit test that causes that case and fixed the bug. I've uploaded the updated patch.

   Thank you!

4. **hyunsik:** I missed +1.

   +1 Ship it!

5. **jhkim:** Hyunsik,
   Thank you for your review.
   I have a problem running integration tests on linux(CentOS 6.3).So, I've changed thread configuration of MiniYARNCluster
   Getting following error

```
2013-08-12 11:46:00,617 INFO  pb.RpcServerFactoryPBImpl (RpcServerFactoryPBImpl.java:createServer(172)) - Adding protocol org.apache.hadoop.yarn.api.AMRMProtocolPB to the server
2013-08-12 11:46:00,617 INFO  ipc.Server (Server.java:run(782)) - IPC Server Responder: starting
2013-08-12 11:46:00,617 INFO  ipc.Server (Server.java:run(613)) - IPC Server listener on 55859: starting
2
013-08-12 11:46:00,620 ERROR service.CompositeService (CompositeService.java:start(72)) - Error starting services ResourceManager
java.lang.OutOfMemoryError: unable to create new native thread
        at java.lang.Thread.start0(Native Method)
        at java.lang.Thread.start(Thread.java:640)
        at org.apache.hadoop.ipc.Server.start(Server.java:2045)
        at org.apache.hadoop.yarn.server.r
esourcemanager.ApplicationMasterService.start(ApplicationMasterService.java:119)
        at org.apache.hadoop.yarn.service.CompositeService.start(CompositeService.java:68)
        at org.apache.hadoop.yarn.server.resourcemanager.ResourceManager.start(ResourceManager.java:567)
        at org.apache.hadoop.yarn.server.MiniYARNCluster$ResourceManagerWrapper$2.run(MiniYARNCluster.java:164)
\end{lstlis
ting} \ \newline%
\begin{lstlisting}
ulimit -a
max user processes              (-u) 1024
```

   For integration test

```
//default is 50
conf.setInt(YarnConfiguration.RM_CLIENT_THREAD_COUNT, 2);
conf.setInt(YarnConfiguration.RM_SCHEDULER_CLIENT_THREAD_COUNT, 2);
conf.setInt(YarnConfiguration.RM_RESOURCE_TRACKER_CLIENT_THREAD_COUNT, 2);
conf.setInt(YarnConfiguration.NM_CONTAINER_MGR_THREAD_COUNT, 2);
```

6. **hyunsik:** Nice finding! +1

7. **jhkim:** I've just committed this.
   Thank you.

8. **hudson:** SUCCESS: Integrated in Tajo-trunk-postcommit #332 (See [https://builds.apache.org/job/Tajo-trunk-postcommit/332/])
   TAJO-118: Refactor and Improve text file Scanner. (jinho) (jinossy: https://git-wip-us.apache.org/repos/asf?p=incubator-tajo.git&a=commit&h=cf6bd4b361b57ffebf1c99ed3e4179f4a155c159)
   * tajo-core/tajo-core-storage/src/main/java/org/apache/tajo/storage/CSVFile.java
   * tajo-core/tajo-core-backend/src/test/java/org/apache/tajo/TajoTestingCluster.java
   * tajo-core/tajo-core-storage/src/main/java/org/apache/tajo/storage/FileScanner.java
   * tajo-core/tajo-core-backend/src/test/java/org/apache/tajo/worker/TestRangeRetrieverHandler.java
   * tajo-core/tajo-core-backend/src/main/java/org/apache/tajo/engine/planner/physical/BSTIndexScanExec.java
   * tajo-core/tajo-core-backend/src/main/java/org/apache/tajo/engine/planner/physical/SeqScanExec.java
   * tajo-core/tajo-core-backend/src/main/java/org/apache/tajo/engine/planner/physical/PhysicalExec.java
   * tajo-core/tajo-core-storage/src/test/java/org/apache/tajo/storage/TestStorages.java
   * tajo-core/tajo-core-backend/src/main/java/org/apache/tajo/engine/query/ResultSetImpl.java
   * tajo-core/tajo-core-backend/src/test/java/org/apache/tajo/engine/planner/physical/TestPhysicalPlanner.java
   * tajo-core/tajo-core-storage/src/test/java/org/apache/tajo/storage/index/TestSingleCSVFileBSTIndex.java
   * tajo-core/tajo-core-storage/src/test/java/org/apache/tajo/storage/TestStorageManager.java
   * tajo-core/tajo-core-storage/src/test/java/org/apache/tajo/storage/index/TestBSTIndex.java
   * tajo-core/tajo-core-backend/src/main/java/org/apache/tajo/master/ClientService.java
   * CHANGES.txt

9. **hudson:** SUCCESS: Integrated in Tajo-trunk-postcommit #334 (See [https://builds.apache.org/job/Tajo-trunk-postcommit/334/])
   TAJO-118: Refactor and Improve text file Scanner. (fixed wrong index) (jinossy: https://git-wip-us.apache.org/repos/asf?p=incubator-tajo.git&a=commit&h=301da59dd8448ef68317008556f3bedadcbf6a83)
   * tajo-core/tajo-core-storage/src/main/java/org/apache/tajo/storage/CSVFile.java

1.6 Pull requests

No pull requests