Item 133
git_comments:

git_commits:

1. **summary:** [SPARK-8118] [SQL] Mutes noisy Parquet log output reappeared after upgrading Parquet to 1.7.0
   message: [SPARK-8118] [SQL] Mutes noisy Parquet log output reappeared after upgrading Parquet to 1.7.0 Author: Cheng Lian <lian@databricks.com> Closes #6670 from liancheng/spark-8118 and squashes the following commits: b6e85a6 [Cheng Lian] Suppresses unnecesary ParquetRecordReader log message (PARQUET-220) 385603c [Cheng Lian] Mutes noisy Parquet log output reappeared after upgrading Parquet to 1.7.0

github_issues:

github_issues_comments:

github_pulls:

1. **title:** [SPARK-8118] [SQL] Mutes noisy Parquet log output reappeared after upgrading Parquet to 1.7.0
   body:

github_pulls_comments:

1. [Test build #34269 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/34269/console) for PR 6670 at commit [`40ba1fd`] (https://github.com/apache/spark/commit/40ba1fd5bcdb255875877ea75da5754e7d90a135). - This patch **fails Spark unit tests**. - This patch merges cleanly. - This patch adds no public classes.
2. [Test build #34276 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/34276/console) for PR 6670 at commit [`385603c`] (https://github.com/apache/spark/commit/385603c44ad0f503dfbfd6096102efb30981666b). - This patch **passes all tests**. - This patch merges cleanly. - This patch adds no public classes.
3. Looks like this covered most of the logs, but there's still one message that's appearing: ``` 04:34:49.075 WARN org.apache.parquet.hadoop.ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptContextImpl ``` Do we need to pass a different type of context or something to fix this?
4. @JoshRosen, this warning is logged mistakenly by Parquet: https://issues.apache.org/jira/browse/PARQUET-220
5. @JoshRosen Yeah, this log line had been there even before bumping Parquet version. @kostya-sh Since this line is the only log message `ParquetRecordReader` issues, I'm going to turn off `ParquetRecordReader` log output here, until Parquet fixes this issue (probably in 1.8.0).
6. [Test build #34354 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/34354/console) for PR 6670 at commit [`b6e85a6`] (https://github.com/apache/spark/commit/b6e85a6811947d4b759be598d8ede71a20af35a0). - This patch **passes all tests**. - This patch merges cleanly. - This patch adds no public classes.
7. Merging to master. Thanks all for reviewing this!
8. Somehow github is no longer closing pull requests.
9. Closing this manually.

github_pulls_reviews:

jira_issues:

1. **summary:** Unnecessary warning in ParquetRecordReader.initialize

description: When reading a parquet file using spark 1.3.0 lots of warnings are printed in the log: {noformat} WARNING: parquet.hadoop.ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptContextImpl {noformat} I have checked the source of ParquetRecordReader and found that while it checks for context to be TaskInputOutputContext it seems to never actually rely on this fact.

jira_issues_comments:

1. {code} diff --git a/parquet-hadoop/src/main/java/parquet/hadoop/ParquetRecordReader.java b/parquet-hadoop/src/main/java/parquet/hadoop/ParquetRecordReader.java index abf65c1..120eeb6 100644 --- a/parquet-hadoop/src/main/java/parquet/hadoop/ParquetRecordReader.java +++ b/parquet-hadoop/src/main/java/parquet/hadoop/ParquetRecordReader.java @@ -41,7 +41,6 @@ import org.apache.hadoop.mapreduce.TaskAttemptContext; import org.apache.hadoop.mapreduce.TaskInputOutputContext; import org.apache.hadoop.mapreduce.lib.input.FileSplit; -import parquet.Log; import parquet.filter.UnboundRecordFilter; import parquet.filter2.compat.FilterCompat; import parquet.filter2.compat.FilterCompat.Filter; @@ -63,7 +62,6 @@ import parquet.schema.MessageType; */ public class ParquetRecordReader<T> extends RecordReader<Void, T> { - private static final Log LOG = Log.getLog(ParquetRecordReader.class); private final InternalParquetRecordReader<T> internalReader; /** @@ -130,12 +128,7 @@ public class ParquetRecordReader<T> extends RecordReader<Void, T> { @Override public void initialize(InputSplit inputSplit, TaskAttemptContext context) throws IOException, InterruptedException { - if (context instanceof TaskInputOutputContext<?, ?, ?, ?>) { - BenchmarkCounter.initCounterFromContext((TaskInputOutputContext<?, ?, ?, ?>) context); - } else { - LOG.error("Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is " + context.getClass().getCanonicalName()); - } + BenchmarkCounter.initCounterFromContext(context); initializeInternalReader(toParquetSplit(inputSplit), ContextUtil.getConfiguration(context)); } diff --git a/parquet-hadoop/src/main/java/parquet/hadoop/util/ContextUtil.java b/parquet-hadoop/src/main/java/parquet/hadoop/util/ContextUtil.java index bb344ad..2e25093 100644 --- a/parquet-hadoop/src/main/java/parquet/hadoop/util/ContextUtil.java +++ b/parquet-hadoop/src/main/java/parquet/hadoop/util/ContextUtil.java @@ -35,7 +35,6 @@ import org.apache.hadoop.mapreduce.RecordWriter; import org.apache.hadoop.mapreduce.StatusReporter; import org.apache.hadoop.mapreduce.TaskAttemptContext; import org.apache.hadoop.mapreduce.TaskAttemptID; -import org.apache.hadoop.mapreduce.TaskInputOutputContext; /* * This is based on ContextFactory.java from hadoop-2.0.x sources. @@ -251,7 +250,7 @@ public class ContextUtil { } } - public static Counter getCounter(TaskInputOutputContext context, + public static Counter getCounter(TaskAttemptContext context, String groupName, String counterName) { return (Counter) invoke(GET_COUNTER_METHOD, context, groupName, counterName); } diff --git a/parquet-hadoop/src/main/java/parquet/hadoop/util/counters/BenchmarkCounter.java b/parquet-hadoop/src/main/java/parquet/hadoop/util/counters/BenchmarkCounter.java index c388d5b..c550c16 100644 --- a/parquet-hadoop/src/main/java/parquet/hadoop/util/counters/BenchmarkCounter.java +++ b/parquet-hadoop/src/main/java/parquet/hadoop/util/counters/BenchmarkCounter.java @@ -20,7 +20,7 @@ package parquet.hadoop.util.counters; import org.apache.hadoop.conf.Configuration; import org.apache.hadoop.mapred.Reporter; -import org.apache.hadoop.mapreduce.TaskInputOutputContext; +import org.apache.hadoop.mapreduce.TaskAttemptContext; import parquet.hadoop.util.counters.mapred.MapRedCounterLoader; import parquet.hadoop.util.counters.mapreduce.MapReduceCounterLoader; @@ -48,7 +48,7 @@ public class BenchmarkCounter { * * @param context */ - public static void initCounterFromContext(TaskInputOutputContext<?, ?, ?, ?> context) { + public static void initCounterFromContext(TaskAttemptContext context) { counterLoader = new MapReduceCounterLoader(context); loadCounters(); } diff --git a/parquet-hadoop/src/main/java/parquet/hadoop/util/counters/mapreduce/MapReduceCounterLoader.java b/parquet-hadoop/src/main/java/parquet/hadoop/util/counters/mapreduce/MapReduceCounterLoader.java index 75ec1a2..c74bfb3 100644 --- a/parquet-hadoop/src/main/java/parquet/hadoop/util/counters/mapreduce/MapReduceCounterLoader.java +++ b/parquet-hadoop/src/main/java/parquet/hadoop/util/counters/mapreduce/MapReduceCounterLoader.java @@ -18,7 +18,7 @@ */ package parquet.hadoop.util.counters.mapreduce; -import org.apache.hadoop.mapreduce.TaskInputOutputContext; +import org.apache.hadoop.mapreduce.TaskAttemptContext; import parquet.hadoop.util.counters.BenchmarkCounter; import parquet.hadoop.util.counters.CounterLoader; @@ -29,7 +29,7 @@ import parquet.hadoop.util.counters.CounterLoader; */ public class MapReduceCounterLoader implements CounterLoader { - private final TaskInputOutputContext context; + private final TaskAttemptContext context; /** * @param context

org.apache.hadoop.mapreduce.TaskInputOutputContext; +import org.apache.hadoop.mapreduce.TaskAttemptContext; import parquet.hadoop.util.ContextUtil; import parquet.hadoop.util.counters.BenchmarkCounter; import parquet.hadoop.util.counters.CounterLoader; @@ -30,9 +30,9 @@ import parquet.hadoop.util.counters.ICounter; * @author Tianshuo Deng */ public class MapReduceCounterLoader implements CounterLoader { - private TaskInputOutputContext<?, ?, ?, ?> context; + private TaskAttemptContext context; - public MapReduceCounterLoader(TaskInputOutputContext<?, ?, ?, ?> context) { + public MapReduceCounterLoader(TaskAttemptContext context) { this.context = context; } {code}

2. Thanks [~k.shaposhnikov@gmail.com], this looks fine to me. Would you mind submitting a pull request for this?
3. https://github.com/apache/incubator-parquet-mr/pull/152
4. Merged in #162.
5. I had to revert the commit because this had test failures (my bad for not running the hadoop-1 tests). Evidently, there are cases in hadoop-1 where the context is not a TaskInputOutputContext and TaskAttemptContext has no getCounter method. I don't think the error message was unnecessary after all, so I'll leave this closed. Feel free to reopen to address it in a different way if you think there is one.
6. Sorry for late response, I was away from my computer on holidays. I've created a new pull request that hopefully addresses the issue with Hadoop 1.x: https://github.com/apache/incubator-parquet-mr/pull/163 The fix moves the check and logging to ContextUtil getCounter() method. Context class doesn't need to be TaskInputOutputContext, it just needs to have getCounter(String, String) method. I've included your changes to the pull request (pom.xml and new deprecated methods) as well.
7. I'm pushing this out of 1.6.0 to get the RC out today. This isn't time-sensitive since it is just avoiding a warning message.
8. Added new pull request for this issue since the old one seems to be dead. Thanks
9. [~sircodesalot] I found the old link: https://github.com/apache/parquet-mr/pull/163
10. I beleive https://github.com/apache/parquet-mr/pull/163/files is still valid though it has to be rebased. I am happy to do it if there is interest to merge it. [~sircodesalot]'s PR looks good too and I will be happy to see it merged as long as the issue is fixed.
11. This PR should have high priority, since this bug creates a mess in our logs.
12. Issue resolved by pull request 280 [https://github.com/apache/parquet-mr/pull/280]