Item 115
git_comments:

git_commits:

1. **summary:** [SPARK-15313][SQL] EmbedSerializerInFilter rule should keep exprIds of output of surrounded SerializeFromObject.
   message: [SPARK-15313][SQL] EmbedSerializerInFilter rule should keep exprIds of output of surrounded SerializeFromObject. ## What changes were proposed in this pull request? The following code: ``` val ds = Seq(("a", 1), ("b", 2), ("c", 3)).toDS() ds.filter(_._1 == "b").select(expr("_1").as[String]).foreach(println(_)) ``` throws an Exception: ``` org.apache.spark.sql.catalyst.errors.package$TreeNodeException: Binding attribute, tree: _1#420 at org.apache.spark.sql.catalyst.errors.package$.attachTree(package.scala:50) at org.apache.spark.sql.catalyst.expressions.BindReferences$$anonfun$bindReference$1.applyOrElse(BoundAttribute.scala:88) at org.apache.spark.sql.catalyst.expressions.BindReferences$$anonfun$bindReference$1.applyOrElse(BoundAttribute.scala:87) ... Cause: java.lang.RuntimeException: Couldn't find _1#420 in [_1#416,_2#417] at scala.sys.package$.error(package.scala:27) at org.apache.spark.sql.catalyst.expressions.BindReferences$$anonfun$bindReference$1$$anonfun$applyOrElse$1.apply(BoundAttribute.scala:94) at org.apache.spark.sql.catalyst.expressions.BindReferences$$anonfun$bindReference$1$$anonfun$applyOrElse$1.apply(BoundAttribute.scala:88) at org.apache.spark.sql.catalyst.errors.package$.attachTree(package.scala:49) at org.apache.spark.sql.catalyst.expressions.BindReferences$$anonfun$bindReference$1.applyOrElse(BoundAttribute.scala:88) at org.apache.spark.sql.catalyst.expressions.BindReferences$$anonfun$bindReference$1.applyOrElse(BoundAttribute.scala:87) ... ``` This is because `EmbedSerializerInFilter` rule drops the `exprId`s of output of surrounded `SerializeFromObject`. The analyzed and optimized plans of the above example are as follows: ``` == Analyzed Logical Plan == _1: string Project [_1#420] +- SerializeFromObject [staticinvoke(class org.apache.spark.unsafe.types.UTF8String, StringType, fromString, input[0, scala.Tuple2]._1, true) AS _1#420,input[0, scala.Tuple2]._2 AS _2#421] +- Filter <function1>.apply +- DeserializeToObject newInstance(class scala.Tuple2), obj#419: scala.Tuple2 +- LocalRelation [_1#416,_2#417], [[0,1800000001,1,61],[0,1800000001,2,62],[0,1800000001,3,63]] == Optimized Logical Plan == !Project [_1#420] +- Filter <function1>.apply +- LocalRelation [_1#416,_2#417], [[0,1800000001,1,61],[0,1800000001,2,62],[0,1800000001,3,63]] ``` This PR fixes `EmbedSerializerInFilter` rule to keep `exprId`s of output of surrounded `SerializeFromObject`. The plans after this patch are as follows: ``` == Analyzed Logical Plan == _1: string Project [_1#420] +- SerializeFromObject [staticinvoke(class org.apache.spark.unsafe.types.UTF8String, StringType, fromString, input[0, scala.Tuple2]._1, true) AS _1#420,input[0, scala.Tuple2]._2 AS _2#421] +- Filter <function1>.apply +- DeserializeToObject newInstance(class scala.Tuple2), obj#419: scala.Tuple2 +- LocalRelation [_1#416,_2#417], [[0,1800000001,1,61],[0,1800000001,2,62],[0,1800000001,3,63]] == Optimized Logical Plan == Project [_1#416] +- Filter <function1>.apply +- LocalRelation [_1#416,_2#417], [[0,1800000001,1,61],[0,1800000001,2,62],[0,1800000001,3,63]] ``` ## How was this patch tested? Existing tests and I added a test to check if `filter and then select` works. Author: Takuya UESHIN <ueshin@happy-camper.st> Closes #13096 from ueshin/issues/SPARK-15313.

github_issues:

github_issues_comments:

github_pulls:

1. **title:** SPARK-1455: Better isolation for unit tests.
   body: This is a simple first step towards avoiding running the Hive tests whenever possible.

github_pulls_comments:

1. Merged build triggered.
2. Merged build started.
3. Merged build finished.
4. Refer to this link for build results: https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/14160/
5. I killed this test on purpose to trigger a re-test. Jenkins, test this please.
6. Jenkins, retest this please.
7. Merged build triggered.
8. Merged build started.
9. Merged build finished. All automated tests passed.
10. All automated tests passed. Refer to this link for build results: https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/14161/

github_pulls_reviews:

jira_issues:

jira_issues_comments: