HBASE-22769

nbanholzer

2019-07-30

Contents

Chapter 1

Root issue HBASE-22769

1.1 Summary

Runtime Error on join (with filter) when using hbase-spark connector

1.2 Description

I am attempting to do a left outer join (though any join with a push down filter causes this issue) between a Spark Structured Streaming DataFrame and a DataFrame read from HBase. I get the following stack trace when running a simple spark app that reads from a streaming source and attempts to left outer join with a dataframe read from HBase:

{{19/07/30 18:30:25 INFO DAGScheduler: ShuffleMapStage 1 (start at SparkAppTest.scala:88) failed in 3.575 s due to Job aborted due to stage failure: Task 0 in stage 1.0 failed 4 times, most recent failure: Lost task 0.3 in stage 1.0 (TID 10, wn5-edpspa.hnyo2upsdeau1bffc34wwrkgwc.ex.internal.cloudapp.net, executor 2): org.apache.hadoop.hbase.DoNotRetryIOException: org.apache.hadoop.hbase.DoNotRetryIOException: java.lang.reflect.InvocationTargetException at org.apache.hadoop.hbase.shaded.protobuf.ProtobufUtil.toFilter(ProtobufU
at org.apache.hadoop.hbase.shaded.protobuf.ProtobufUtil.toScan(ProtobufUtil.java:1154) at org.apache.hadoop.hbase.reg
at org.apache.hadoop.hbase.regionserver.RSRpcServices.scan(RSRpcServices.java:3301) at org.apache.hadoop.hbase.shad
at org.apache.hadoop.hbase.ipc.RpcServer.call(RpcServer.java:413) at org.apache.hadoop.hbase.ipc.CallRunner.run(CallR
at org.apache.hadoop.hbase.ipc.RpcExecutor$Handler.run(RpcExecutor.java:324) at org.apache.hadoop.hbase.ipc.RpcEx
Caused by: java.lang.reflect.InvocationTargetException at sun.reflect.GeneratedMethodAccessor15461.invoke(Unknown Source) at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498) at org.apache.hadoop.hbase.shaded.protobuf.ProtobufUtil.toFilter(P
}}

{{... 8 more }}

{{Caused by: java.lang.NoClassDefFoundError: org/apache/hadoop/hbase/spark/datasources/-JavaBytesEncoder$ at org.apache.hadoop.hbase.spark.datasources.JavaBytesEncoder.create(JavaBytesEncoder.scala)
at org.apache.hadoop.hbase.spark.SparkSQLPushDownFilter.parseFrom(SparkSQLPushDownFilter.java:196)
}}

{{... 12 more }}

{{at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method) at sun.reflect.NativeConstructorAccessorI
at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
at java.lang.reflect.Constructor.newInstance(Constructor.java:423) at org.apache.hadoop.hbase.ipc.RemoteWithExtrasEx
at org.apache.hadoop.hbase.ipc.RemoteWithExtrasException.unwrapRemoteException(RemoteWithExtrasException.jav

at org.apache.hadoop.hbase.shaded.protobuf.ProtobufUtil.makeIOExceptionOfException(ProtobufUtil.java:359)
at org.apache.hadoop.hbase.shaded.protobuf.ProtobufUtil.handleRemoteException(ProtobufUtil.java:347)
at org.apache.hadoop.hbase.client.ScannerCallable.openScanner(ScannerCallable.java:344) at org.apache.hadoop.hbase.cl
at org.apache.hadoop.hbase.client.ScannerCallable.rpcCall(ScannerCallable.java:58) at org.apache.hadoop.hbase.client.Re
at org.apache.hadoop.hbase.client.RpcRetryingCallerImpl.callWithoutRetries(RpcRetryingCallerImpl.java:192)
at org.apache.hadoop.hbase.client.ScannerCallableWithReplicas$RetryingRPC.call(ScannerCallableWithReplicas.java:38
at org.apache.hadoop.hbase.client.ScannerCallableWithReplicas$RetryingRPC.call(ScannerCallableWithReplicas.java:36
at org.apache.hadoop.hbase.client.RpcRetryingCallerImpl.callWithRetries(RpcRetryingCallerImpl.java:107)
at org.apache.hadoop.hbase.client.ResultBoundedCompletionService$QueueingFuture.run(ResultBoundedCompletionSer
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149) at java.util.concurrent.ThreadPo
at java.lang.Thread.run(Thread.java:748)}}

It appears to be attempting to reference a file called "JavaBytesEncoder$.class" resulting in a NoClassDefFoundError. Interestingly, when I unzipped the jar I found that both "JavaBytesEncoder.class" and "JavaBytesEncoder$.class" exist, but the latter is simply an empty file. This might just be a case of me misunderstanding how Java links classes upon build however.

1.3 Attachments

No attachments

1.4 Comments

1. **nbanholzer:** Possibly related to [HBASE-17989|https://issues.apache.org/jira/browse/HBASE-17989]
2. **ujunko:** Hello, I'm hitting this issue too. I can read data from HBase table fine as long as I don't use where() or filter(). Once I put where/filter I get the same exception. Below is the very simple code I used for testing. I didn't do anything fancy.

```
1
2  val hbaseDF = spark.read
3
4                  .options(Map(HBaseTableCatalog.tableCatalog -> catalog))
5
6  Âă Âă Âă Âă Âă Âă Âă Âă Âă.format("org.apache.hadoop.hbase.spark").load()
7
8  val lookedup = hbaseDF.filter($"colName" === "some value")
```

The class in question is JavaBytesEncoder, which is in hbase-spark connector jar. This jar exists in my spark job(executor). It looks like the cause is InvocationTargetException due to Scala Reflection, and it is happening NOT in spark executor, but probably somewhere in between Zookeeper and HBase.

CDH Cluster ver. 6.1.1

Hbase-spark library version 2.1.0-cdh6.1.1

Scala version 2.11.8

Spark version 2.4.3

3. **jan101:** Same problem. I can load and show data in a {{hbaseDF}} as created in the comment above. When I try to filter that DataFrame I get

```
1
2  org.apache.hadoop.hbase.DoNotRetryIOException: org.apache.hadoop.hbase.DoN
3
4  otRetryIOException: java.lang.ClassNotFoundException: org.apache.hadoop.hbase.spark
       .SparkSQLPushDownFilter
5
6          at org.apache.hadoop.hbase.protobuf.ProtobufUtil.toFilter(ProtobufUtil.java
               :1679)
7
8          at org.apache.hadoop.hbase.protobuf.ProtobufUtil.toScan(ProtobufUtil.java
               :1163)
9
10         at org.apache.hadoop.hbase.regionserver.RSRpcServices.newRegionScanner(
               RSRpcServices.java:2682)
11
12         at org.apache.hadoop.hbase.regionserver.RSRpcServices.scan(RSRpcServices.
               java:3013)
13
14         at org.apache.hadoop.hbase.protobuf.generated.ClientProtos$ClientService$2.
               callBlockingMethod(ClientProtos.java:36613)
15
16         at org.apache.hadoop.hbase.ipc.RpcServer.call(RpcServer.java:2380)
17
18         at org.apache.hadoop.hbase.ipc.CallRunner.run(CallRunner.java:124)
19
20         at org.apache.hadoop.hbase.ipc.RpcExecutor$Handler.run(RpcExecutor.java
               :297)
21
22         at org.apache.hadoop.hbase.ipc.RpcExecutor$Handler.run(RpcExecutor.java
               :277)
23
24 Caused by: java.lang.ClassNotFoundException: org.apache.hadoop.hbase.spark.
       SparkSQLPushDownFilter
25
26         at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
27
28         at org.apache.hadoop.hbase.util.DynamicClassLoader.tryRefreshClass(
               DynamicClassLoader.java:173)
29
30         at org.apache.hadoop.hbase.util.DynamicClassLoader.loadClass(
               DynamicClassLoader.java:140)
31
32         at java.lang.Class.forName0(Native Method)
33
34         at java.lang.Class.forName(Class.java:348)
35
36         at org.apache.hadoop.hbase.protobuf.ProtobufUtil.toFilter(ProtobufUtil.java
               :1670)
37
38         ... 8 more
```

This is on AWS emr-5.27.0 with Spark 2.4.4, HBase 1.4.10, Scala 2.11.12

And spark-shell with the following packages

```
1
2 spark-shell --packages org.apache.hbase.connectors.spark:hbase-spark:1.0.0,org.
      apache.hbase:hbase-client:2.1.0,org.apache.hbase:hbase-common:2.1.0,org.apache.
      hbase:hbase-server:2.1.0,org.apache.hbase:hbase:2.1.0
```

4. **aniruddh02:** Any update on this issue?
5. **RISHI__23:** Hi can I work on this issue?

6. **lucacanali:** The error can be worked around with ‘.option("hbase.spark.pushdown.columnfilter", false)‘ but this bypasses SQL pushdown, which is a useful feature.
7. **lucacanali:** ‘.option("hbase.spark.pushdown.columnfilter", true)‘ is the default. This requires additional configuration on the HBase server side, in particular one needs to have a few jars in the HBase region servers CLASSPATH: scala-library, hbase-spark and hbase-spark-protocol-shaded. For exmaple:

   - scala-library-2.11.12.jar

   - hbase-spark-1.0.0.jar

   - hbase-spark-protocol-shaded-1.0.0.jar