使用Spark SQL读取Hive上的数据

Spark SQL主要目的是使得用户可以在Spark上使用SQL，其数据源既可以是RDD，也可以是外部的数据源（比如Parquet、Hive、Json等）。Spark SQL的其中一个分支就是Spark on Hive，也就是使用Hive中HQL的解析、逻辑执行计划翻译、执行计划优化等逻辑，可以近似认为仅将物理执行计划从MR作业替换成了Spark作业。本文就是来介绍如何通过Spark SQL来读取现有Hive中的数据。

不过，预先编译好的Spark assembly包是不支持Hive的，如果你需要在Spark中使用Hive，必须重新编译，加上-Phive选项既可，具体如下：

```
[iteblog@www.iteblog.com spark]$ ./make-distribution.sh --tgz -Phadoop-2.2 -Pyarn -DskipTests -Dhadoop.version=2.2.0  -Phive
```

编译完成之后，会在SPARK_HOME的lib目录下多产生三个jar包，分别是datanucleus-api-jdo-3.2.6.jar、datanucleus-core-3.2.10.jar、datanucleus-rdbms-3.2.9.jar，这些包都是Hive所需要的。下面就开始介绍步骤。

一、环境准备

为了让Spark能够连接到Hive的原有数据仓库，我们需要将Hive中的hive-site.xml文件拷贝到Spark的conf目录下，这样就可以通过这个配置文件找到Hive的元数据以及数据存放。

如果Hive的元数据存放在Mysql中，我们还需要准备好Mysql相关驱动，比如：mysql-connector-java-5.1.22-bin.jar。

二、启动spark-shell

环境准备好之后，为了方便起见，我们使用spark-shell来进行说明如何通过Spark SQL读取Hive中的数据。我们可以通过下面的命令来启动spark-shell：

```
[iteblog@www.iteblog.com spark]$  bin/spark-shell --master yarn-client  --jars lib/mysql-connector-java-5.1.22-bin.jar
....
15/08/27 18:21:25 INFO repl.SparkILoop: Created spark context..
Spark context available as sc.
....
15/08/27 18:21:30 INFO repl.SparkILoop: Created sql context (with Hive support)..
SQL context available as sqlContext.
```

启动spark-shell的时候会先向ResourceManager申请资源，而且还会初始化SparkContext和SQLContext实例。sqlContext对象其实是HiveContext的实例，sqlContext是进入Spark SQL的切入点。接下来我们来读取Hive中的数据。

```
scala> sqlContext.sql("CREATE EXTERNAL  TABLE IF NOT EXISTS ewaplog (key STRING, value ST
RING)
STORED AS INPUTFORMAT 'com.hadoop.mapred.DeprecatedLzoTextInputFormat' OUTPUTFOR
MAT
'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat' LOCATION '/user/iteblog/ewa
plog' ")

res0: org.apache.spark.sql.DataFrame = [result: string]

scala> sqlContext.sql("LOAD DATA LOCAL INPATH '/data/test.lzo' INTO TABLE ewaplog")
res1: org.apache.spark.sql.DataFrame = [result: string]

scala> sqlContext.sql("FROM ewaplog SELECT key, value").collect().foreach(println)
[12,wyp]
[23,ry]
[12,wyp]
[23,ry]
```

我们先创建了ewaplog表，然后导入数据，最后查询。我们可以看出所有的SQL和在Hive中是一样的，只是在Spark上运行而已。在执行SQL的时候，默认是调用hiveql解析器来解析SQL的。当然，你完全可以调用Spark SQL内置的SQL解析器sql，可以通过spark.sql.dialect参数来设置。但是建议还是使用hivesql解析器，因为它支持的语法更多，而且还支持Hive的UDF函数，在多数情况下推荐使用hivesql解析器。

如果你在创建HiveContext的时候出现了类似以下的错误：

```
15/11/20 16:20:07 WARN metadata.Hive: Failed to access metastore. This class should not acc
essed in runtime.
org.apache.hadoop.hive.ql.metadata.HiveException: java.lang.RuntimeException: Unable to in
stantiate org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient
 at org.apache.hadoop.hive.ql.metadata.Hive.getAllDatabases(Hive.java:1236)
 at org.apache.hadoop.hive.ql.metadata.Hive.reloadFunctions(Hive.java:174)
 at org.apache.hadoop.hive.ql.metadata.Hive.<clinit>(Hive.java:166)
 at org.apache.hadoop.hive.ql.session.SessionState.start(SessionState.java:503)
 at org.apache.spark.sql.hive.client.ClientWrapper.<init>(ClientWrapper.scala:171)
 at org.apache.spark.sql.hive.HiveContext.executionHive$lzycompute(HiveContext.scala:162)
 at org.apache.spark.sql.hive.HiveContext.executionHive(HiveContext.scala:160)
```

```
 at org.apache.spark.sql.hive.HiveContext.<init>(HiveContext.scala:167)
 at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
 at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:57)
 at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
 at java.lang.reflect.Constructor.newInstance(Constructor.java:526)
 at org.apache.spark.repl.SparkILoop.createSQLContext(SparkILoop.scala:1028)
 at $line4.$read$$iwC$$iwC.<init>(<console>:9)
 at $line4.$read$$iwC.<init>(<console>:18)
 at $line4.$read.<init>(<console>:20)
 at $line4.$read$.<init>(<console>:24)
 at $line4.$read$.<clinit>(<console>)
 at $line4.$eval$.<init>(<console>:7)
 at $line4.$eval$.<clinit>(<console>)
 at $line4.$eval.$print(<console>)
 at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
 at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
 at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
 at java.lang.reflect.Method.invoke(Method.java:606)
 at org.apache.spark.repl.SparkIMain$ReadEvalPrint.call(SparkIMain.scala:1065)
 at org.apache.spark.repl.SparkIMain$Request.loadAndRun(SparkIMain.scala:1340)
 at org.apache.spark.repl.SparkIMain.loadAndRunReq$1(SparkIMain.scala:840)
 at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:871)
 at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:819)
 at org.apache.spark.repl.SparkILoop.reallyInterpret$1(SparkILoop.scala:857)
 at org.apache.spark.repl.SparkILoop.interpretStartingWith(SparkILoop.scala:902)
 at org.apache.spark.repl.SparkILoop.command(SparkILoop.scala:814)
 at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:132)
 at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:124)
 at org.apache.spark.repl.SparkIMain.beQuietDuring(SparkIMain.scala:324)
 at org.apache.spark.repl.SparkILoopInit$class.initializeSpark(SparkILoopInit.scala:124)
 at org.apache.spark.repl.SparkILoop.initializeSpark(SparkILoop.scala:64)
 at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1$$anonfun$apply$mcZ$sp$5.apply$mcV$sp(SparkILoop.scala:974)
 at org.apache.spark.repl.SparkILoopInit$class.runThunks(SparkILoopInit.scala:159)
 at org.apache.spark.repl.SparkILoop.runThunks(SparkILoop.scala:64)
 at org.apache.spark.repl.SparkILoopInit$class.postInitialization(SparkILoopInit.scala:108)
 at org.apache.spark.repl.SparkILoop.postInitialization(SparkILoop.scala:64)
 at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply$mcZ$sp(SparkILoop.scala:991)
 at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply(SparkILoop.scala:945)
 at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$proces
```

```
s$1.apply(SparkILoop.scala:945)
 at scala.tools.nsc.util.ScalaClassLoader$.savingContextLoader(ScalaClassLoader.scala:135)
 at org.apache.spark.repl.SparkILoop.org$apache$spark$repl$SparkILoop$$process(SparkILoop.scala:945)
 at org.apache.spark.repl.SparkILoop.process(SparkILoop.scala:1059)
 at org.apache.spark.repl.Main$.main(Main.scala:31)
 at org.apache.spark.repl.Main.main(Main.scala)
 at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
 at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
 at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
 at java.lang.reflect.Method.invoke(Method.java:606)
 at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:674)
 at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:180)
 at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:205)
 at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:120)
 at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Caused by: java.lang.RuntimeException: Unable to instantiate org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient
 at org.apache.hadoop.hive.metastore.MetaStoreUtils.newInstance(MetaStoreUtils.java:1523)
 at org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.<init>(RetryingMetaStoreClient.java:86)
 at org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.getProxy(RetryingMetaStoreClient.java:132)
 at org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.getProxy(RetryingMetaStoreClient.java:104)
 at org.apache.hadoop.hive.ql.metadata.Hive.createMetaStoreClient(Hive.java:3005)
 at org.apache.hadoop.hive.ql.metadata.Hive.getMSC(Hive.java:3024)
 at org.apache.hadoop.hive.ql.metadata.Hive.getAllDatabases(Hive.java:1234)
 ... 59 more
Caused by: java.lang.reflect.InvocationTargetException
 at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
 at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:57)
 at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
 at java.lang.reflect.Constructor.newInstance(Constructor.java:526)
 at org.apache.hadoop.hive.metastore.MetaStoreUtils.newInstance(MetaStoreUtils.java:1521)
 ... 65 more
Caused by: MetaException(message:Version information not found in metastore. )
 at org.apache.hadoop.hive.metastore.ObjectStore.checkSchema(ObjectStore.java:6664)
 at org.apache.hadoop.hive.metastore.ObjectStore.verifySchema(ObjectStore.java:6645)
 at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
 at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
 at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
 at java.lang.reflect.Method.invoke(Method.java:606)
```

```
 at org.apache.hadoop.hive.metastore.RawStoreProxy.invoke(RawStoreProxy.java:114)
 at com.sun.proxy.$Proxy15.verifySchema(Unknown Source)
 at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.getMS(HiveMetaStore.java
:572)
 at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.createDefaultDB(HiveMeta
Store.java:620)
 at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.init(HiveMetaStore.java:46
1)
 at org.apache.hadoop.hive.metastore.RetryingHMSHandler.<init>(RetryingHMSHandler.java:6
6)
 at org.apache.hadoop.hive.metastore.RetryingHMSHandler.getProxy(RetryingHMSHandler.jav
a:72)
 at org.apache.hadoop.hive.metastore.HiveMetaStore.newRetryingHMSHandler(HiveMetaStore
.java:5762)
 at org.apache.hadoop.hive.metastore.HiveMetaStoreClient.<init>(HiveMetaStoreClient.java:19
9)
 at org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient.<init>(SessionHiveMetaSt
oreClient.java:74)
 ... 70 more
15/11/20 16:20:07 INFO metastore.HiveMetaStore: 0: Opening raw store with implemenation c
lass:org.apache.hadoop.hive.metastore.ObjectStore
15/11/20 16:20:07 INFO metastore.ObjectStore: ObjectStore, initialize called
15/11/20 16:20:07 INFO metastore.MetaStoreDirectSql: Using direct SQL, underlying DB is DER
BY
15/11/20 16:20:07 INFO metastore.ObjectStore: Initialized ObjectStore
java.lang.RuntimeException: java.lang.RuntimeException: Unable to instantiate org.apache.ha
doop.hive.ql.metadata.SessionHiveMetaStoreClient
 at org.apache.hadoop.hive.ql.session.SessionState.start(SessionState.java:522)
 at org.apache.spark.sql.hive.client.ClientWrapper.<init>(ClientWrapper.scala:171)
 at org.apache.spark.sql.hive.HiveContext.executionHive$lzycompute(HiveContext.scala:162)
 at org.apache.spark.sql.hive.HiveContext.executionHive(HiveContext.scala:160)
 at org.apache.spark.sql.hive.HiveContext.<init>(HiveContext.scala:167)
 at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
 at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.ja
va:57)
 at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccesso
rImpl.java:45)
 at java.lang.reflect.Constructor.newInstance(Constructor.java:526)
 at org.apache.spark.repl.SparkILoop.createSQLContext(SparkILoop.scala:1028)
 at $iwC$$iwC.<init>(<console>:9)
 at $iwC.<init>(<console>:18)
 at <init>(<console>:20)
 at .<init>(<console>:24)
 at .<clinit>(<console>)
 at .<init>(<console>:7)
 at .<clinit>(<console>)
```

```
 at $print(<console>)
 at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
 at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
 at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
 at java.lang.reflect.Method.invoke(Method.java:606)
 at org.apache.spark.repl.SparkIMain$ReadEvalPrint.call(SparkIMain.scala:1065)
 at org.apache.spark.repl.SparkIMain$Request.loadAndRun(SparkIMain.scala:1340)
 at org.apache.spark.repl.SparkIMain.loadAndRunReq$1(SparkIMain.scala:840)
 at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:871)
 at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:819)
 at org.apache.spark.repl.SparkILoop.reallyInterpret$1(SparkILoop.scala:857)
 at org.apache.spark.repl.SparkILoop.interpretStartingWith(SparkILoop.scala:902)
 at org.apache.spark.repl.SparkILoop.command(SparkILoop.scala:814)
 at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.sca
la:132)
 at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.sca
la:124)
 at org.apache.spark.repl.SparkIMain.beQuietDuring(SparkIMain.scala:324)
 at org.apache.spark.repl.SparkILoopInit$class.initializeSpark(SparkILoopInit.scala:124)
 at org.apache.spark.repl.SparkILoop.initializeSpark(SparkILoop.scala:64)
 at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$proces
s$1$$anonfun$apply$mcZ$sp$5.apply$mcV$sp(SparkILoop.scala:974)
 at org.apache.spark.repl.SparkILoopInit$class.runThunks(SparkILoopInit.scala:159)
 at org.apache.spark.repl.SparkILoop.runThunks(SparkILoop.scala:64)
 at org.apache.spark.repl.SparkILoopInit$class.postInitialization(SparkILoopInit.scala:108)
 at org.apache.spark.repl.SparkILoop.postInitialization(SparkILoop.scala:64)
 at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$proces
s$1.apply$mcZ$sp(SparkILoop.scala:991)
 at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$proces
s$1.apply(SparkILoop.scala:945)
 at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$proces
s$1.apply(SparkILoop.scala:945)
 at scala.tools.nsc.util.ScalaClassLoader$.savingContextLoader(ScalaClassLoader.scala:135)
 at org.apache.spark.repl.SparkILoop.org$apache$spark$repl$SparkILoop$$process(SparkILoo
p.scala:945)
 at org.apache.spark.repl.SparkILoop.process(SparkILoop.scala:1059)
 at org.apache.spark.repl.Main$.main(Main.scala:31)
 at org.apache.spark.repl.Main.main(Main.scala)
 at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
 at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
 at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
 at java.lang.reflect.Method.invoke(Method.java:606)
 at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$$runMain
(SparkSubmit.scala:674)
 at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:180)
 at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:205)
```

```
 at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:120)
 at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Caused by: java.lang.RuntimeException: Unable to instantiate org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient
 at org.apache.hadoop.hive.metastore.MetaStoreUtils.newInstance(MetaStoreUtils.java:1523)
 at org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.<init>(RetryingMetaStoreClient.java:86)
 at org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.getProxy(RetryingMetaStoreClient.java:132)
 at org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.getProxy(RetryingMetaStoreClient.java:104)
 at org.apache.hadoop.hive.ql.metadata.Hive.createMetaStoreClient(Hive.java:3005)
 at org.apache.hadoop.hive.ql.metadata.Hive.getMSC(Hive.java:3024)
 at org.apache.hadoop.hive.ql.session.SessionState.start(SessionState.java:503)
 ... 56 more
Caused by: java.lang.reflect.InvocationTargetException
 at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
 at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:57)
 at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
 at java.lang.reflect.Constructor.newInstance(Constructor.java:526)
 at org.apache.hadoop.hive.metastore.MetaStoreUtils.newInstance(MetaStoreUtils.java:1521)
 ... 62 more
Caused by: MetaException(message:Version information not found in metastore. )
 at org.apache.hadoop.hive.metastore.ObjectStore.checkSchema(ObjectStore.java:6664)
 at org.apache.hadoop.hive.metastore.ObjectStore.verifySchema(ObjectStore.java:6645)
 at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
 at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
 at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
 at java.lang.reflect.Method.invoke(Method.java:606)
 at org.apache.hadoop.hive.metastore.RawStoreProxy.invoke(RawStoreProxy.java:114)
 at com.sun.proxy.$Proxy15.verifySchema(Unknown Source)
 at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.getMS(HiveMetaStore.java:572)
 at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.createDefaultDB(HiveMetaStore.java:620)
 at org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.init(HiveMetaStore.java:461)
 at org.apache.hadoop.hive.metastore.RetryingHMSHandler.<init>(RetryingHMSHandler.java:66)
 at org.apache.hadoop.hive.metastore.RetryingHMSHandler.getProxy(RetryingHMSHandler.java:72)
 at org.apache.hadoop.hive.metastore.HiveMetaStore.newRetryingHMSHandler(HiveMetaStore.java:5762)
 at org.apache.hadoop.hive.metastore.HiveMetaStoreClient.<init>(HiveMetaStoreClient.java:19
```

9)
 at org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient.<init>(SessionHiveMetaStoreClient.java:74)
 ... 67 more

看下你的Hadoop集群是否可以连接Mysql元数据。