使用Spark处理存储于Hive中的Twitter数据的一些技巧

本文将介绍使用Spark batch作业处理存储于Hive中Twitter数据的一些实用技巧。

首先我们需要引入一些依赖包，参考如下：

```
name := "Sentiment"

version := "1.0"

scalaVersion := "2.10.6"

assemblyJarName in assembly := "sentiment.jar"

libraryDependencies += "org.apache.spark" % "spark-core_2.10" % "1.6.0" % "provided"
libraryDependencies += "org.apache.spark" % "spark-sql_2.10" % "1.6.0" % "provided"
libraryDependencies += "org.apache.spark" %% "spark-hive" % "1.6.0" % "provided"
libraryDependencies += "edu.stanford.nlp" % "stanford-corenlp" % "3.5.1"
libraryDependencies += "edu.stanford.nlp" % "stanford-corenlp" % "3.5.1" classifier "models"

resolvers += "Akka Repository" at "http://repo.akka.io/releases/"

assemblyMergeStrategy in assembly := {
  case PathList("META-INF", xs @ _*) => MergeStrategy.discard
  case x => MergeStrategy.first}
```

编写一个Scala case class用于存储解析好的Twitter Json数据：

```
case class Tweet(coordinates: String, geo:String, handle: String,
                 hashtags: String, language: String,
                 location: String, msg: String, time: String,
                 tweet_id: String, unixtime: String,
                 user_name: String, tag: String,
                 profile_image_url: String,
                 source: String, place: String, friends_count: String,
                 followers_count: String, retweet_count: String,
                 time_zone: String, sentiment: String,
                 stanfordSentiment: String)
```

引入以下的包：

```
import java.util.Properties
import com.vader.SentimentAnalyzer
import edu.stanford.nlp.ling.CoreAnnotations
import edu.stanford.nlp.neural.rnn.RNNCoreAnnotations
import edu.stanford.nlp.pipeline.StanfordCoreNLP
import edu.stanford.nlp.sentiment.SentimentCoreAnnotations
import org.apache.log4j.{Level, Logger}
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.serializer.KryoSerializer
import org.apache.spark.sql._
```

用Scala编写的用于从Hive中读取数据的Spark代码片段：

```
def main(args: Array[String]) {
  Logger.getLogger("org.apache.spark").setLevel(Level.ERROR)
  Logger.getLogger("org.apache.spark.storage.BlockManager").setLevel(Level.ERROR)
  val logger: Logger = Logger.getLogger("com.iteblog.sentiment.TwitterSentimentAnalysis")
  val sparkConf = new SparkConf().setAppName("TwitterSentimentAnalysis")
  sparkConf.set("spark.streaming.backpressure.enabled", "true")
  sparkConf.set("spark.cores.max", "32")
  sparkConf.set("spark.serializer", classOf[KryoSerializer].getName)
  sparkConf.set("spark.sql.tungsten.enabled", "true")
  sparkConf.set("spark.eventLog.enabled", "true")
  sparkConf.set("spark.app.id", "Sentiment")
  sparkConf.set("spark.io.compression.codec", "snappy")
  sparkConf.set("spark.rdd.compress", "true")
  val sc = new SparkContext(sparkConf)
  val sqlContext = new org.apache.spark.sql.hive.HiveContext(sc)
  import sqlContext.implicits._
  val tweets = sqlContext.read.json("hdfs://www.iteblog.com:8020/social/twitter")
  sqlContext.setConf("spark.sql.orc.filterPushdown", "true")
  tweets.printSchema()
  tweets.count
  tweets.take(5).foreach(println)
```

其中我们需要注意的是我们需要创建Hive context而不是标准的SQL context

在运行我们的代码之前，先确认Hive中存储Twitter Json数据的表，以及用于存放结果数据的表格

是否存在，本文用于存储结果数据的表格使用了ORC 格式

```
beeline
!connect jdbc:hive2://localhost:10000/default;
!set showHeader true;
set hive.vectorized.execution.enabled=true;
set hive.execution.engine=tez;
set hive.vectorized.execution.enabled =true;
set hive.vectorized.execution.reduce.enabled =true;
set hive.compute.query.using.stats=true;
set hive.cbo.enable=true;
set hive.stats.fetch.column.stats=true;
set hive.stats.fetch.partition.stats=true;
show tables;
describe sparktwitterorc;
describe twitterraw;
describe sparktwitterorc;
analyze table sparktwitterorc compute statistics;
analyze table sparktwitterorc compute statistics for columns;
```

上面名为twitterraw的表格是用于存放Twitter
Json数据的表；而名为sparktwitterorc的表格是用于存放Spark处理结果的表。

如何将RDD或者DataFrame中的数据写入到Hive ORC表呢？操作如下：

```
outputTweets.toDF().write.format("orc").mode(SaveMode.Overwrite).saveAsTable("default.spa
rktwitterorc")
```

在编译的程序时候设置JVM相关参数

```
export SBT_OPTS="-Xmx2G -XX:+UseConcMarkSweepGC -XX:+CMSClassUnloadingEnabled -XX:
MaxPermSize=2G -Xss2M  -Duser.timezone=GMT"
sbt -J-Xmx4G -J-Xms4G assembly
```

将Spark作业提交到YARN集群：

```
spark-submit --class com.iteblog.sentiment.TwitterSentimentAnalysis --master yarn-
```

client sentiment.jar --verbose

这里附上我们的rawtwitter表建表语句：

```
CREATE TABLE rawtwitter
(
  handle            STRING,
  hashtags          STRING,
  msg               STRING,
  language          STRING,
  time              STRING,
  tweet_id          STRING,
  unixtime          STRING,
  user_name         STRING,
  geo               STRING,
  coordinates       STRING,
  `location`        STRING,
  time_zone         STRING,
  retweet_count     STRING,
  followers_count   STRING,
  friends_count     STRING,
  place             STRING,
  source            STRING,
  profile_image_url STRING,
  tag               STRING,
  sentiment         STRING,
  stanfordsentiment STRING
)
ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe'
LOCATION 'hdfs://www.iteblog.com:8020/social/twitter'
```

本文翻译自：https://dzone.com/articles/spark-tips-must-have-for-twitter-batch-processing