# spark-yarn 模式提交优化调研

调研人: 李振超

在使用 yarn 模式提交 spark 程序的时候, spark 会自动上传一些用到的 jar 包和配置到 hdfs .有些 jar包在每次提交的时候都需要上传, 这在一定程度上延长了代码执行的时间.

下面给出一些优化措施可以减少代码上传时间.

## 第 1 章 应该如何打包spark应用程序

由于项目需要用到很多额外 jar 包, 所以打包的时候应该使用 assembly 插件打包成一个 fat jar.

Spark 和 Hadoop 相关的依赖建议不要打包, 因为spark 环境和集群可以提供相应的 jar, 从而可以减少 fat jar 的大小.

* pom.xml 示例:

<build>  
 <plugins>  
 <!-- 打包插件, 否则 scala 类不会编译并打包进去 -->  
 <plugin>  
 <groupId>net.alchim31.maven</groupId>  
 <artifactId>scala-maven-plugin</artifactId>  
 <version>3.4.6</version>  
 <executions>  
 <execution>  
 <goals>  
 <goal>compile</goal>  
 <goal>testCompile</goal>  
 </goals>  
 </execution>  
 </executions>  
 </plugin>  
  
 <plugin>  
 <artifactId>maven-compiler-plugin</artifactId>  
 <version>2.3.2</version>  
 <configuration>  
 <source>1.8</source>  
 <target>1.8</target>  
 </configuration>  
 </plugin>  
 <plugin>  
 <artifactId>maven-assembly-plugin</artifactId>  
 <configuration>  
 <descriptorRefs>  
 <descriptorRef>jar-with-dependencies</descriptorRef>  
 </descriptorRefs>  
 </configuration>  
 <executions>  
 <execution>  
 <id>make-assembly</id>  
 <phase>package</phase>  
 <goals>  
 <goal>single</goal>  
 </goals>  
 </execution>  
 </executions>  
 </plugin>  
 </plugins>  
</build>

## 第 2 章 提交前的准备

不做任何优化前提交需要上传到 hdfs 的文件一般有这些:

1. 用户的 jar 包

Added JAR file:/opt/module/spark-yarn/./examples/jars/spark-examples\_2.11-2.1.1.jar at spark://192.168.2.102:38931/jars/spark-examples\_2.11-2.1.1.jar

1. 配置文件

Uploading resource file:/tmp/spark-49d38ea2-a808-48e6-91a9-5bf93e015cc6/\_\_spark\_conf\_\_8892838950200912185.zip -> hdfs://hadoop102:9000/user/atguigu/.sparkStaging/application\_1583029596886\_0001/\_\_spark\_conf\_\_.zip

配置文件主要包含 $SPARK\_HOME/conf 和 $HADOOP\_HOME/etc/hadoop/ 下所有的配置文件. 这些配置文件会提前压缩成 zip 文件, 放在 /tmp/spark-….. 临时目录下, 然后统一上传. 等程序执行完毕后, 临时目录会自动删除



1. SPARK\_HOME 下所有的 lib

会提前压缩, 然后统一上传

Uploading resource file:/tmp/spark-49d38ea2-a808-48e6-91a9-5bf93e015cc6/\_\_spark\_libs\_\_8938131311537532107.zip -> hdfs://hadoop102:9000/user/atguigu/.sparkStaging/application\_1583029596886\_0001/\_\_spark\_libs\_\_8938131311537532107.zip

1. spark-default.xml中 spark.jars 配置的 jar 包

spark.jars=/opt/module/hadoop-2.7.2/share/hadoop/common/hadoop-lzo-0.4.20.jar

Uploading resource file:/opt/module/hadoop-2.7.2/share/hadoop/common/hadoop-lzo-0.4.20.jar -> hdfs://hadoop102:9000/user/atguigu/.sparkStaging/application\_1583029596886\_0001/hadoop-lzo-0.4.20.jar

## 第 3 章 需要上传的文件的优化

我们可以把需要上传的比较大的文件提前缓存到 hdfs, 然后就不用每次上传了.

1. 在 spark-default.xml内通过属性 spark.yarn.jars (或者 spark.yarn.archive)配置 spark 的 lib 路径

# 在 hdfs 创建目录并上传

hadoop fs -mkdir /spark\_jars

hadoop fs -put lib/\* /spark\_jars

# 在 spark-default.xml 上配置 spark.yarn.jars

**spark.yarn.jars**=hdfs://hadoop102:9000/spark\_jars/\*

1. 把用户的 jar 上传到 hdfs

注意: 如果是 yarn-client 只能使用本地的用户类 jar 包. yarn-cluster才可以使用hdfs 上的 jar

hadoop fs -put Spark-test-1.0-SNAPSHOT-jar-with-dependencies.jar /

## 第 4 章 提交应用

bin/spark-submit \

--master yarn \

--class com.atguigu.yarn.WordCount \

--deploy-mode cluster \

hdfs://hadoop102:9000/Spark-test-1.0-SNAPSHOT-jar-with-dependencies.jar

