Apache Spark 历史服务器（HistoryServer）日志过大解决

最近突然收到线上服务器发出来的磁盘满了的报警，然后到服务器上发现 Apache Spark 的历史服务器（HistoryServer）日志居然占了近 500GB，如下所示：

```
[root@iteblog.com spark]# ll -h
total 328
-rw-rw-r-- 1 spark spark 15.4G Jul 11 13:09 spark-spark-org.apache.spark.deploy.history.HistoryServer-1-iteblog.com.out
-rw-rw-r-- 1 spark spark  369M May 30 09:07 spark-spark-org.apache.spark.deploy.history.HistoryServer-1-iteblog.com.out.1
-rw-rw-r-- 1 spark spark 13.6G May 11 21:36 spark-spark-org.apache.spark.deploy.history.HistoryServer-1-iteblog.com.out.2
-rw-rw-r-- 1 spark spark  4.4G Apr 04 11:04 spark-spark-org.apache.spark.deploy.history.HistoryServer-1-iteblog.com.out.3
-rw-rw-r-- 1 spark spark 11.8G Mar 24 12:47 spark-spark-org.apache.spark.deploy.history.HistoryServer-1-iteblog.com.out.4
-rw-rw-r-- 1 spark spark  7.5G Feb 12 23:43 spark-spark-org.apache.spark.deploy.history.HistoryServer-1-iteblog.com.out.5
...
```

原来是因为 HistoryServer 启动的时候使用了默认的 log4j.properties 文件里面的配置，导致每次只有重启 HistoryServer 的时候日志才会切割，如果不重启 HistoryServer，日志文件会无限制的增长下去；而且切割出去的日志文件也不会被 Spark 系统清理，久而久之导致存放 HistoryServer 日志的文件夹越来越大，这就导致了今天的磁盘报警。

如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：iteblog_hadoop

问题已经找到，解决这个问题的方法也很直接，就是重新定义 Spark 自带的 log4j.properties 文件，将默认的 org.apache.log4j.ConsoleAppender 修改成 org.apache.log4j.RollingFileAppender，并设置每个日志文件最大为1G，最多保持7个这样的文件，具体如下：

```
log4j.logger.org.apache.spark.deploy.history=info,historyserver
log4j.appender.historyserver=org.apache.log4j.RollingFileAppender
log4j.appender.historyserver.layout=org.apache.log4j.PatternLayout
log4j.appender.historyserver.layout.ConversionPattern=%d{yy/MM/dd HH:mm:ss} %p %c{1}: %m%n
log4j.appender.historyserver.MaxFileSize=1GB
log4j.appender.historyserver.MaxBackupIndex=7
log4j.appender.historyserver.File=/data/logs/spark/historyserver.log
```

经过这样的配置，服务器的日志如下

```
[root@iteblog.com spark]# ll -h
-rw-rw-r-- 1 spark spark  235M Jul 11 13:09 historyserver.log
-rw-rw-r-- 1 spark spark    1G Jul 11 13:09 historyserver.log.1
-rw-rw-r-- 1 spark spark    1G Jul 11 13:07 historyserver.log.2
```

可以看到，HistoryServer
的日志超过1GB就会自动切割，而且最多只保存7个这样的文件，占用的磁盘空间可控。

本文虽然介绍的是 Spark HistoryServer 日志磁盘占用问题，我们可以举一反三，对于 Spark Streaming 的输出日志也可以使用类似的方法解决；同理，Spark 的 ThriftServer
也可以采用这种方法解决。