# 配置

## spark-env.sh

HADOOP\_CONF\_DIR=/data/1/service/hadoop-2.6.0-cdh5.8.2/etc/hadoop

## spark-defaults.conf

spark.history.ui.port 18080

spark.eventLog.enabled true

spark.eventLog.dir hdfs://node01.csdn.net:9000/user/spark/eventLog

spark.history.fs.logDirectory hdfs://node01.csdn.net:9000/user/spark/eventLog

spark.yarn.archive hdfs://node01.csdn.net:9000/user/spark/jars/jars.zip

spark.yarn.historyServer.address 192.168.25.23:18080

#spark.yarn.principal spark/client01.csdn.net@CSDN.NET

#spark.yarn.keytab /etc/security/keytabs/spark.service.keytab

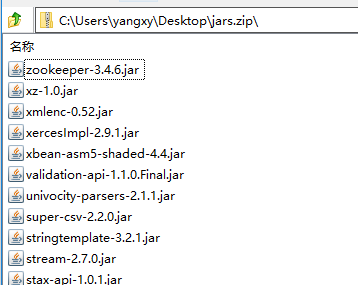
#spark.history.kerberos.enabled true

#spark.history.kerberos.keytab /etc/security/keytabs/spark.service.keytab

#spark.history.kerberos.principal [spark/client01.csdn.net@CSDN.NET](mailto:spark/client01.csdn.net@CSDN.NET)

注意：spark.yarn.archive的zip文件是spark解压后jars目录的压缩文件(#zip ../jars.zip \*)

The archive should contain jar files in its root directory



**启动${SPARK\_HOME}sbin/start-history-server.sh**

## yarn-site.xml

启用日志聚合

<property>

<description>Whether to enable log aggregation</description>

<name>yarn.log-aggregation-enable</name>

<value>true</value>

</property>

<property>

<description>Where to aggregate logs to.</description>

<name>yarn.nodemanager.remote-app-log-dir</name>

<value>/tmp/logs</value>

</property>

<property>

<name>yarn.log.server.url</name>

<value>http://node03.csdn.net:19888/jobhistory/logs</value>

</property>

## mapred-site.xml

启动${HADOOP\_HOME}**sbin/mr-jobhistory-daemon.sh start historyserver**

<property>

<name>mapreduce.jobhistory.address</name>

<value>node03.csdn.net:10020</value>

<description>MapReduce JobHistory Server IPC host:port</description>

</property>

<property>

<name>mapreduce.jobhistory.webapp.address</name>

<value>node03.csdn.net:19888</value>

<description>MapReduce JobHistory Server Web UI host:port</description>

</property>

# Spark Streaming 日志滚动



# 常见问题

## 关闭虚拟内存检测

Diagnostics: Container [pid=12648,containerID=container\_1478584727672\_0004\_02\_000001] is running beyond virtual memory

limits. Current usage: 284.2 MB of 1 GB physical memory used; 2.3 GB of 2.1 GB virtual memory used. Killing container.

<https://issues.apache.org/jira/browse/YARN-4714>

<http://blog.cloudera.com/blog/2014/04/apache-hadoop-yarn-avoiding-6-time-consuming-gotchas/>

**Add following property in yarn-site.xml**

<property>

<name>yarn.nodemanager.vmem-check-enabled</name>

<value>false</value>

<description>Whether virtual memory limits will be enforced for containers</description>

</property>

<property>

<name>yarn.nodemanager.vmem-pmem-ratio</name>

<value>4</value>

<description>Ratio between virtual memory to physical memory when setting memory limits for containers</description>

</property>

## ResourceManager HA

方法一：

在spark-env.sh中指定的hadoop客户端配置**yarn-site.xml**中添加

<property>

<name>yarn.web-proxy.address</name>

<value> web-proxy:8089</value>

</property>

yarn-daemon.sh start proxyserver

方法二：

<http://www.jianshu.com/p/ea85d074a494>

在spark-env.sh中指定的hadoop客户端配置**yarn-site.xml**中添加

yarn.resourcemanager.webapp.address属性即可（配置其中的一个ResourceManager的地址即可），不需要重启yarn

<property>

<description>The http address of the RM web application.</description>

<name>yarn.resourcemanager.webapp.address</name>

<value>${yarn.resourcemanager.hostname}:8088</value>

</property>

## java.lang.NoClassDefFoundError: com/sun/jersey/api/client/config/ClientConfig

Solution

Download jersey-bundle-1.17.1.jar, and copy to $SPARK\_HOME/jars, or use --jars when calling spark-shell

or spark-submit:

$ spark-shell --jars extra\_jars/jersey-bundle-1.17.1.jar

or set spark.hadoop.yarn.timeline-service.enabled to false

$bin/spark-shell --master yarn --conf spark.hadoop.yarn.timeline-service.enabled=false

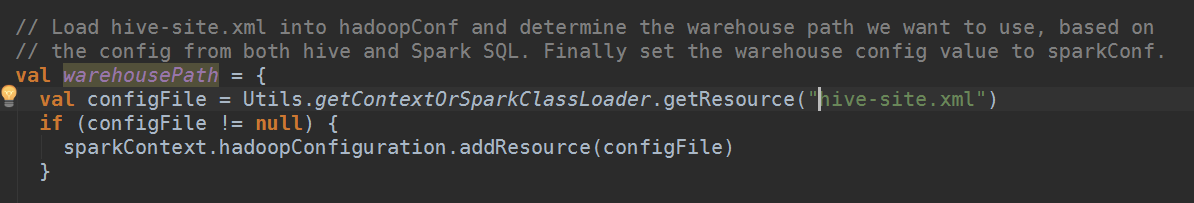
## Error: org.apache.spark.sql.AnalysisException: org.apache.hadoop.hive.ql.metadata.HiveException: Unable to move source



<http://blog.csdn.net/nysyxxg/article/details/61196279>

<https://stackoverflow.com/questions/44233523/spark-sql-2-1-1-thrift-server-unable-to-move-source-hdfs-to-target>

Try setting fs.hdfs.impl.disable.cache in your hive-site.xml like this:



<?xml version="1.0"?>

<configuration>

<property>

<name>hive.metastore.uris</name>

<value>thrift://CSDN-HDP-VCG-01:9083</value>

</property>

<property>

<name>fs.hdfs.impl.disable.cache</name>

<value>true</value>

</property>

<property>

<name>hive.exec.stagingdir</name>

<value>/tmp/hive/spark-${user.name}</value>

</property>

</configuration>

## NullPointerException on empty ORC file

<https://issues.apache.org/jira/browse/SPARK-19809>

先删除分区下的空文件

