Spark Cluster Setup（Standalone模式）

# tar -zxvf spark-2.4.0-bin-hadoop2.7.tgz

# /Spark/spark-2.4.0-bin-hadoop2.7/conf/spark-env.sh

SPARK\_MASTER\_HOST = Host1

SPARK\_MASTER\_PORT=7077

SPARK\_MASTER\_WEBUI\_POR=8080

# To launch a Spark standalone cluster with the launch scripts, you should create a file called *conf/slaves* in your Spark directory, which must contain the hostnames of all the machines where you intend to start Spark workers, one per line.

# 复制文件到其他机器

将/Spark复制到其他机器。scp -r /Spark Host3:/Spark

# 启动Spark

/Spark/spark-2.4.0-bin-hadoop2.7/sbin/start-all.sh

# History-server启动设置：

# 修改 log4j.properties把INFO级别设置成WARN级别

log4j.rootCategory=WARN, console

# 解决：WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...

* 环境变量$JAVA\_LIBRARY\_PATH并未定义
* 编辑：/etc/profile
  + export JAVA\_LIBRARY\_PATH=/Hadoop/hadoop-2.8.0/lib/native

# 解决：WARN metastore.ObjectStore: Failed to get database global\_temp, returning NoSuchObjectException

# 运行“pyspark --master yarn”出现警告：

Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploadi ng libraries under SPARK\_HOME.

* List of libraries containing Spark code to distribute to YARN containers. By default, Spark on YARN will use Spark jars installed locally, but the Spark jars can also be in a world-readable location on HDFS. This allows YARN to cache it on nodes so that it doesn't need to be distributed each time an application runs. To point to jars on HDFS, for example, set this configuration to hdfs:///some/path. Globs are allowed.
* 在hdfs上创建目录：

hdfs dfs -mkdir /home/hadoop/spark\_jars

上传spark的jars（spark1.6 只需要上传spark-assembly-1.6.0-SNAPSHOT-hadoop2.6.0.jar）

hdfs dfs –put /Spark/spark-2.1.0-bin-hadoop2.7/jars/\* /home/hadoop/spark\_jars/

* 在spark的conf的spark-default.conf
* 添加配置 spark.yarn.jars hdfs://V1:9000/Spark/spark-2.1.0-bin-hadoop2.7/jars/\*.jar即可解决