**赵强老师简介**

15年以上的IT行业从业经历，清华大学计算机软件工程专业毕业，现任Oracle中国有限公司高级技术顾问；曾在BEA、甲骨文、摩托罗拉等世界500强公司担任高级软件架构师或咨询顾问等要职，精通大数据、数据库、中间件技术和Java技术。

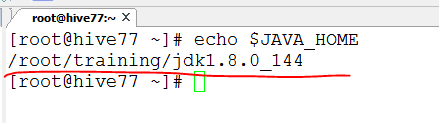


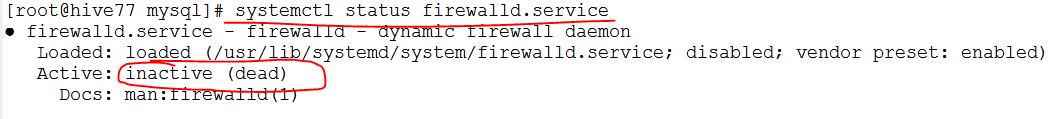
# 版本如下

**注意：Hive on Spark对版本有着严格的要求，下面的版本是经过验证的版本**

* 1. apache-hive-2.3.2-bin.tar.gz
  2. hadoop-2.7.2.tar.gz
  3. jdk-8u144-linux-x64.tar.gz
  4. mysql-5.7.19-1.el7.x86\_64.rpm-bundle.tar
  5. mysql-connector-java-5.1.43-bin.jar
  6. spark-2.0.0.tgz（spark源码包，需要从源码编译）
  7. Redhat Linux 7.4 64位

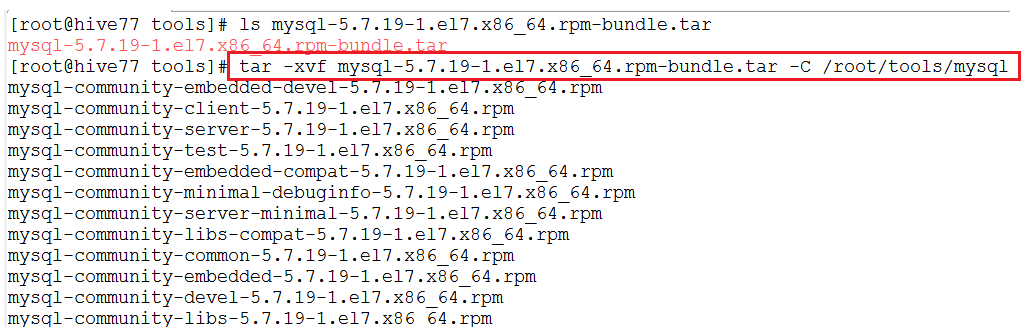
# 安装Linux和JDK、关闭防火墙





# 安装和配置MySQL数据库

# 解压MySQL安装包



# **安装MySQL**

yum remove mysql-libs

rpm -ivh mysql-community-common-5.7.19-1.el7.x86\_64.rpm

rpm -ivh mysql-community-libs-5.7.19-1.el7.x86\_64.rpm

rpm -ivh mysql-community-client-5.7.19-1.el7.x86\_64.rpm

rpm -ivh mysql-community-server-5.7.19-1.el7.x86\_64.rpm

rpm -ivh mysql-community-devel-5.7.19-1.el7.x86\_64.rpm （可选）

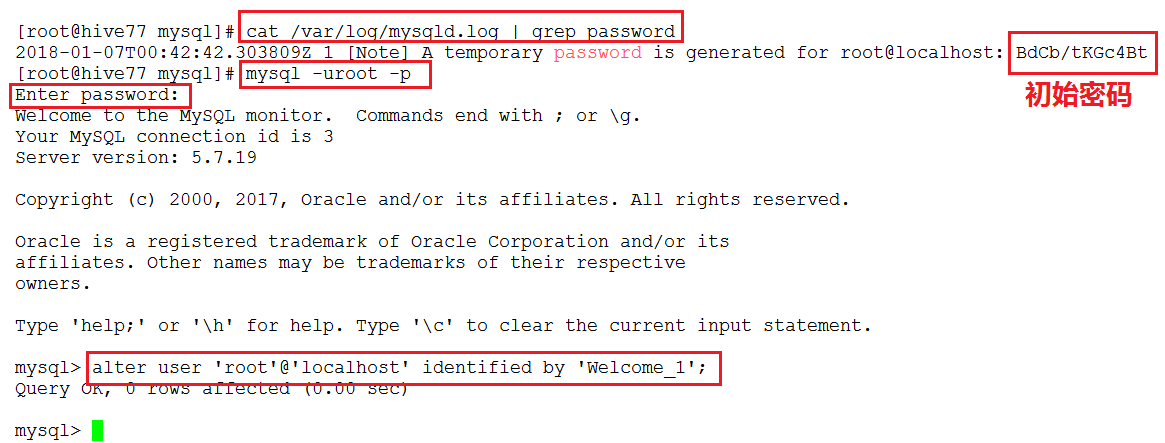
# **启动MySQL**

systemctl start mysqld.service

# **查看并修改root用户的密码**

查看root用户的密码：cat /var/log/mysqld.log | grep password

登录后修改密码：alter user 'root'@'localhost' identified by 'Welcome\_1';



# **创建hive的数据库和hiveowner用户：**

* 创建一个新的数据库：create database hive;
* 创建一个新的用户：
  + create user 'hiveowner'@'%' identified by ‘Welcome\_1’;
* 给该用户授权
  + grant all on hive.\* TO 'hiveowner'@'%';
  + grant all on hive.\* TO 'hiveowner'@'localhost' identified by 'Welcome\_1';

# 安装Hadoop（以伪分布式为例）

**由于Hive on Spark默认支持Spark on Yarn的方式，所以需要配置Hadoop。**

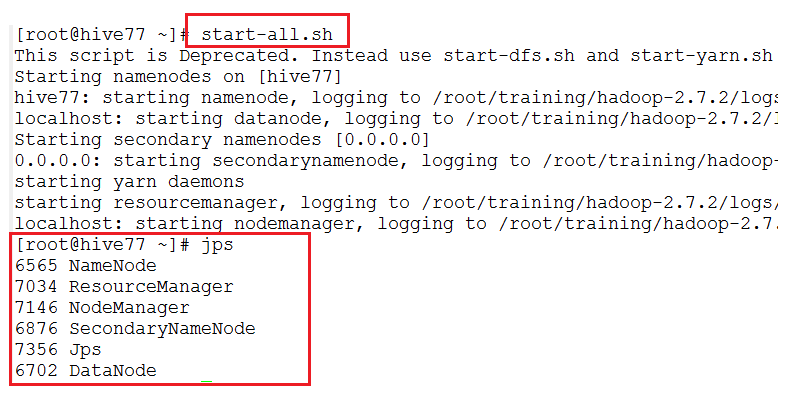
## 准备工作：

* + 1. 配置主机名（编辑/etc/hosts文件）
    2. 配置免密码登录

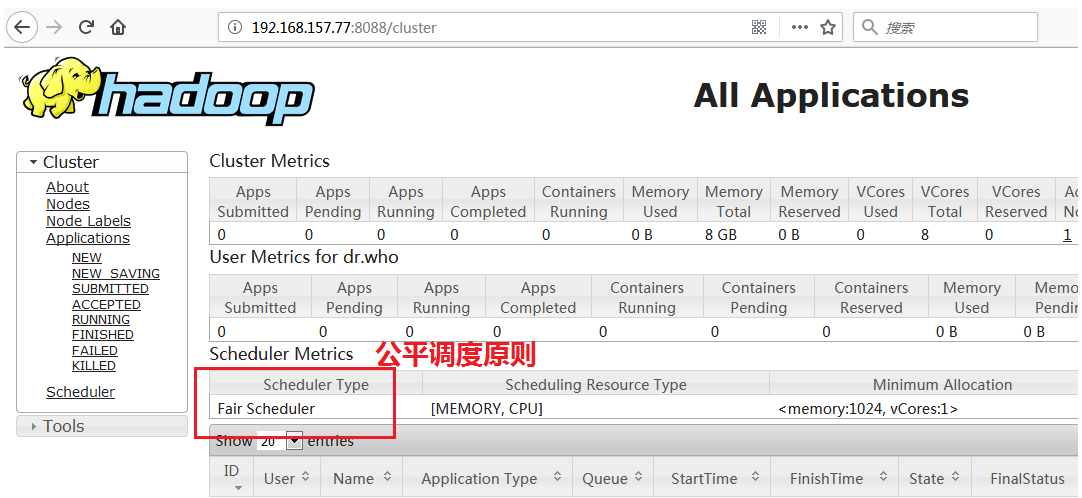
## **Hadoop的配置文件如下：**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **hadoop-env.sh** | | | | | |
| JAVA\_HOME | /root/training/jdk1.8.0\_144 | | |  | |
| **hdfs-site.xml** | | | | | |
| dfs.replication | 1 | | | 数据块的冗余度，默认是3 | |
| dfs.permissions | false | | | 是否开启HDFS的权限检查 | |
| **core-site.xml** | | | | | |
| fs.defaultFS | hdfs://hive77:9000 | | | NameNode的地址 | |
| hadoop.tmp.dir | /root/training/hadoop-2.7.2/tmp/ | | | HDFS数据保存的目录 | |
| **mapred-site.xml** | | | | | |
| mapreduce.framework.name | | | yarn |  | |
| **yarn-site.xml** | | | | | |
| yarn.resourcemanager.hostname | | hive77 | | |  |
| yarn.nodemanager.aux-services | | mapreduce\_shuffle | | |  |
| yarn.resourcemanager.scheduler.class | | org.apache.hadoop.yarn.server.resourcemanager.scheduler.fair.FairScheduler | | | **Spark on Yarn的方式，需要使用公平调度原则来保证Yarn集群中的任务都能获取到相等的资源运行。** |

## **启动Hadoop**



## **通过Yarn Web Console检查是否为公平调度原则**



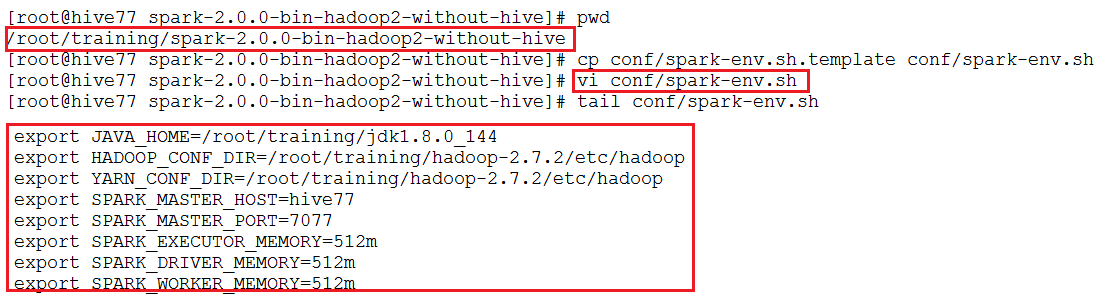
# 编译Spark源码

**（需要使用Maven，Spark源码包中自带Maven）**

* 1. 执行下面的语句进行编译（**执行时间很长，耐心等待**）

./dev/make-distribution.sh --name "hadoop2-without-hive" --tgz "-Pyarn,hadoop-provided,hadoop-2.7,parquet-provided"

* 1. 编译成功后，会生成：**spark-2.0.0-bin-hadoop2-without-hive.tgz**
  2. 安装和配置Spark
     1. 目录结构如下：



* + 1. 将下面的配置加入spark-env.sh

export JAVA\_HOME=/root/training/jdk1.8.0\_144

export HADOOP\_CONF\_DIR=/root/training/hadoop-2.7.2/etc/hadoop

export YARN\_CONF\_DIR=/root/training/hadoop-2.7.2/etc/hadoop

export SPARK\_MASTER\_HOST=hive77

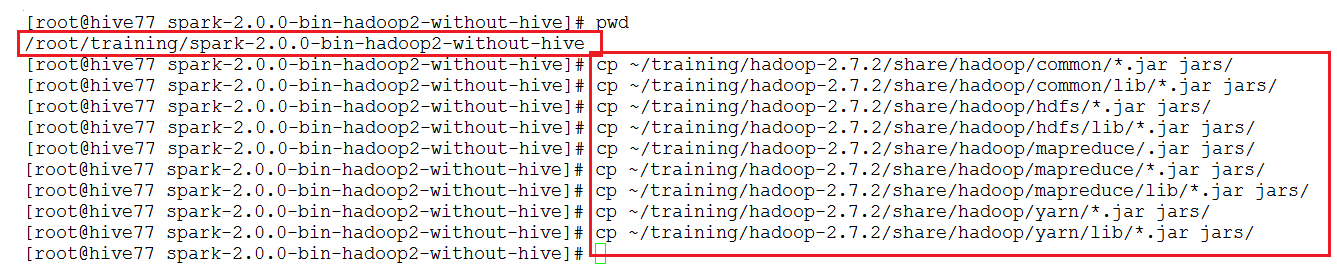
export SPARK\_MASTER\_PORT=7077

export SPARK\_EXECUTOR\_MEMORY=512m

export SPARK\_DRIVER\_MEMORY=512m

export SPARK\_WORKER\_MEMORY=512m

* + 1. 将hadoop的相关jar包放入spark的lib目录下，如下：



cp ~/training/hadoop-2.7.2/share/hadoop/common/\*.jar jars/

cp ~/training/hadoop-2.7.2/share/hadoop/common/lib/\*.jar jars/

cp ~/training/hadoop-2.7.2/share/hadoop/hdfs/\*.jar jars/

cp ~/training/hadoop-2.7.2/share/hadoop/hdfs/lib/\*.jar jars/

cp ~/training/hadoop-2.7.2/share/hadoop/mapreduce/.jar jars/

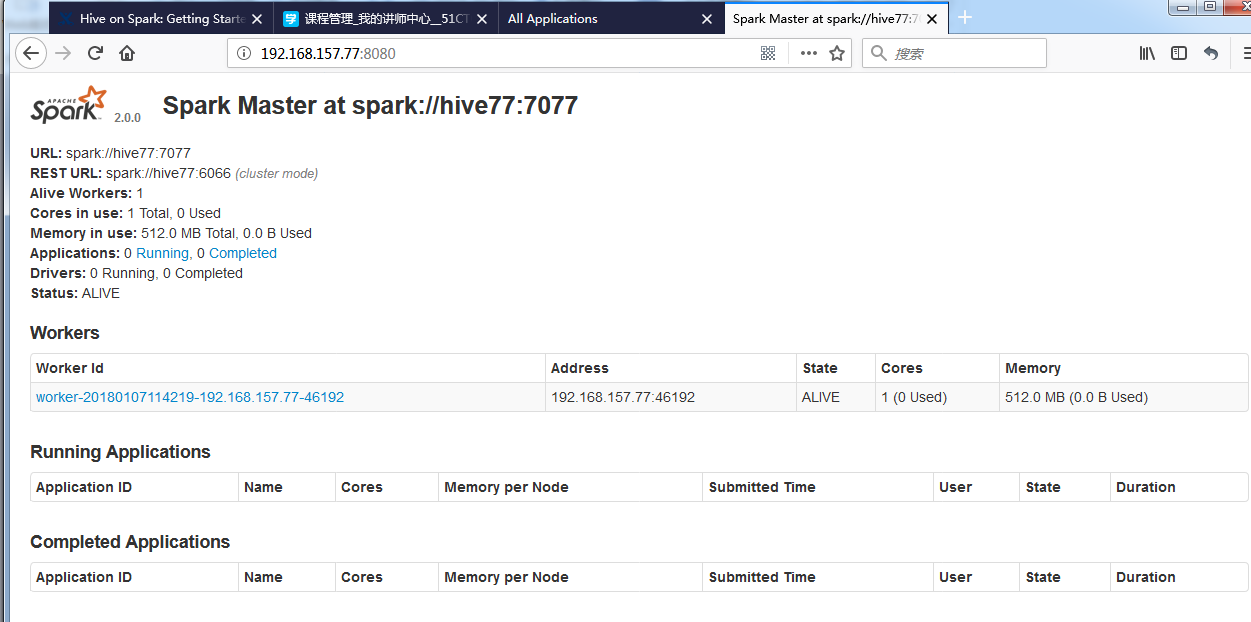
cp ~/training/hadoop-2.7.2/share/hadoop/mapreduce/\*.jar jars/

cp ~/training/hadoop-2.7.2/share/hadoop/mapreduce/lib/\*.jar jars/

cp ~/training/hadoop-2.7.2/share/hadoop/yarn/\*.jar jars/

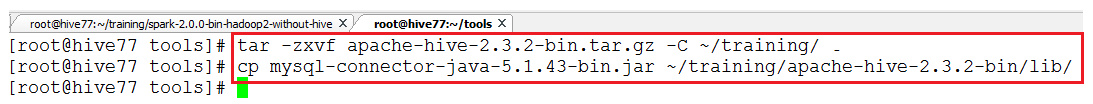
cp ~/training/hadoop-2.7.2/share/hadoop/yarn/lib/\*.jar jars/

* + 1. 在HDFS上创建目录：spark-jars，并将spark的jars上传至该目录。**这样在运行Application的时候，就无需每次都分发这些jar包。**
* hdfs dfs -mkdir /spark-jars
* hdfs dfs -put jars/\*.jar /spark-jars
  1. 启动Spark：sbin/start-all.sh，验证Spark是否配置成功



# 安装配置Hive

* 1. 解压Hive安装包，并把mysql的JDBC驱动放到HIve的lib目录下，如下图：



* 1. 设置Hive的环境变量

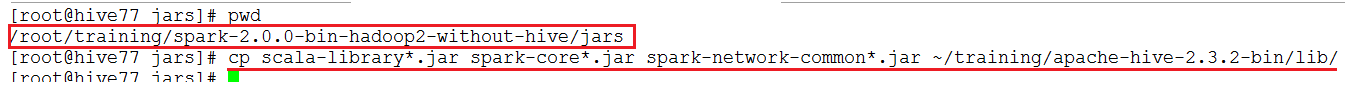
HIVE\_HOME=/root/training/apache-hive-2.3.2-bin

export HIVE\_HOME

PATH=$HIVE\_HOME/bin:$PATH

export PATH

* 1. 拷贝下面spark的jar包到Hive的lib目录
     1. scala-library
     2. spark-core
     3. spark-network-common



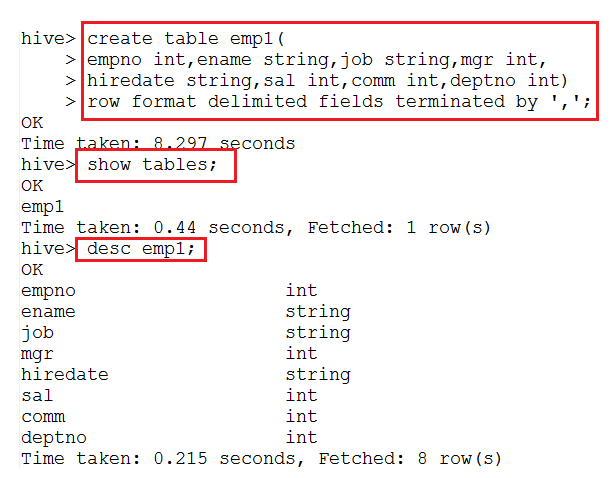
* 1. 在HDFS上创建目录：/sparkeventlog用于保存log信息

hdfs dfs -mkdir /sparkeventlog

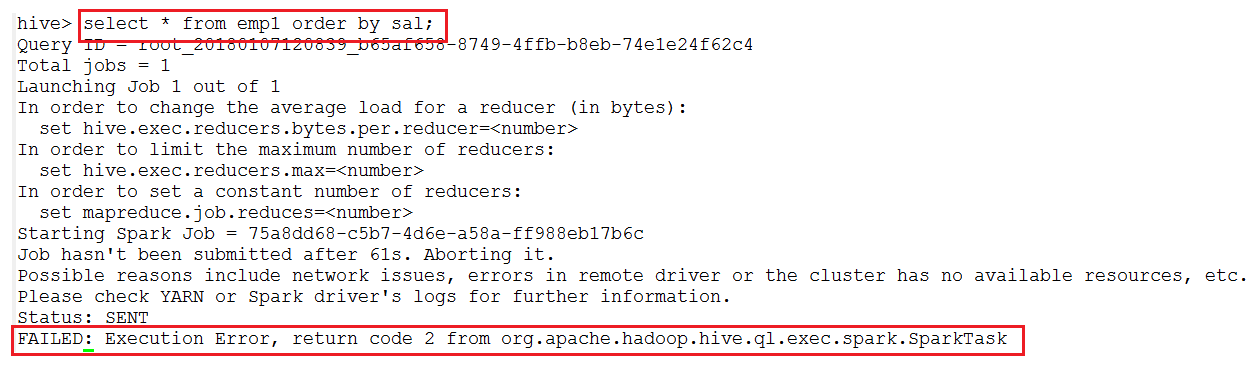
* 1. 配置hive-site.xml，如下：

|  |  |
| --- | --- |
| **参数** | **参考值** |
| javax.jdo.option.ConnectionURL | jdbc:mysql://localhost:3306/hive?useSSL=false |
| javax.jdo.option.ConnectionDriverName | com.mysql.jdbc.Driver |
| javax.jdo.option.ConnectionUserName | hiveowner |
| javax.jdo.option.ConnectionPassword | Welcome\_1 |
| hive.execution.engine | spark |
| hive.enable.spark.execution.engine | true |
| spark.home | /root/training/spark-2.0.0-bin-hadoop2-without-hive |
| spark.master | yarn-client |
| spark.eventLog.enabled | true |
| spark.eventLog.dir | hdfs://hive77:9000/sparkeventlog |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |
| spark.executor.memeory | 512m |
| spark.driver.memeory | 512m |

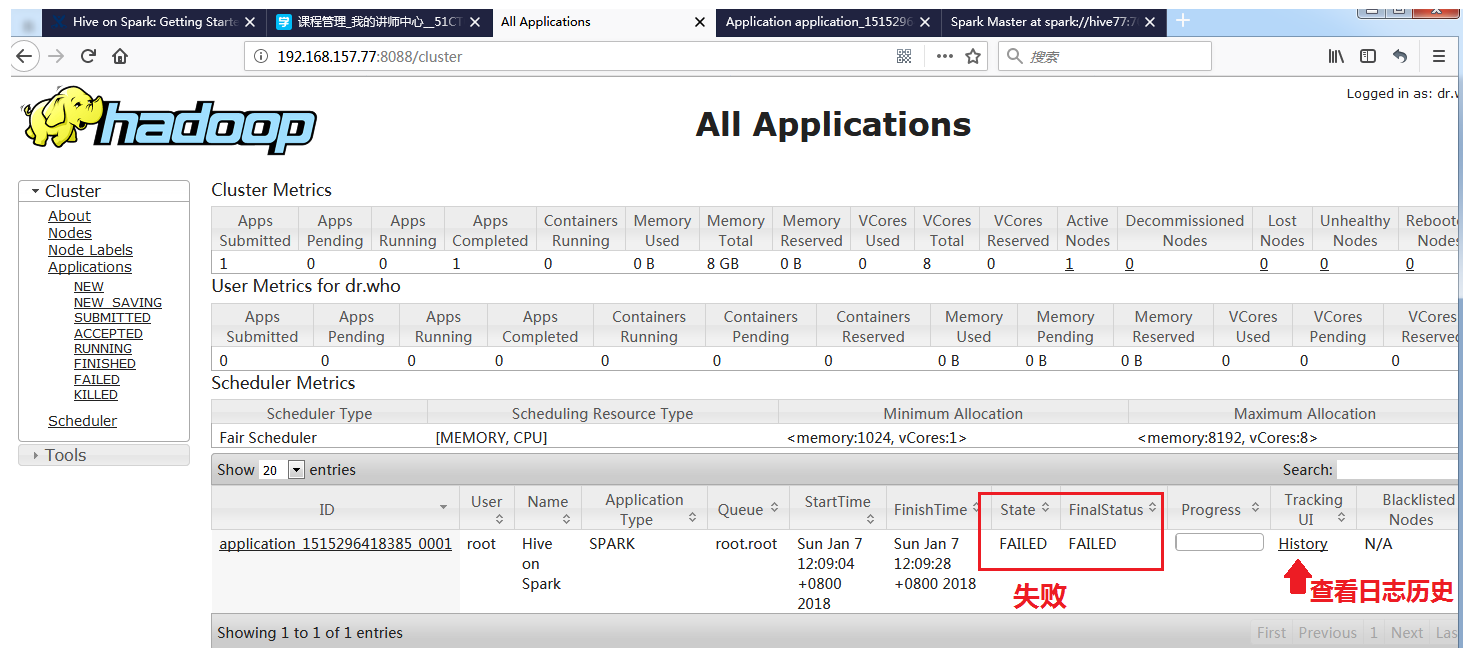
* 1. 初始化MySQL数据库：**schematool -dbType mysql -initSchema**
  2. 启动hive shell，并创建员工表，用于保存员工数据

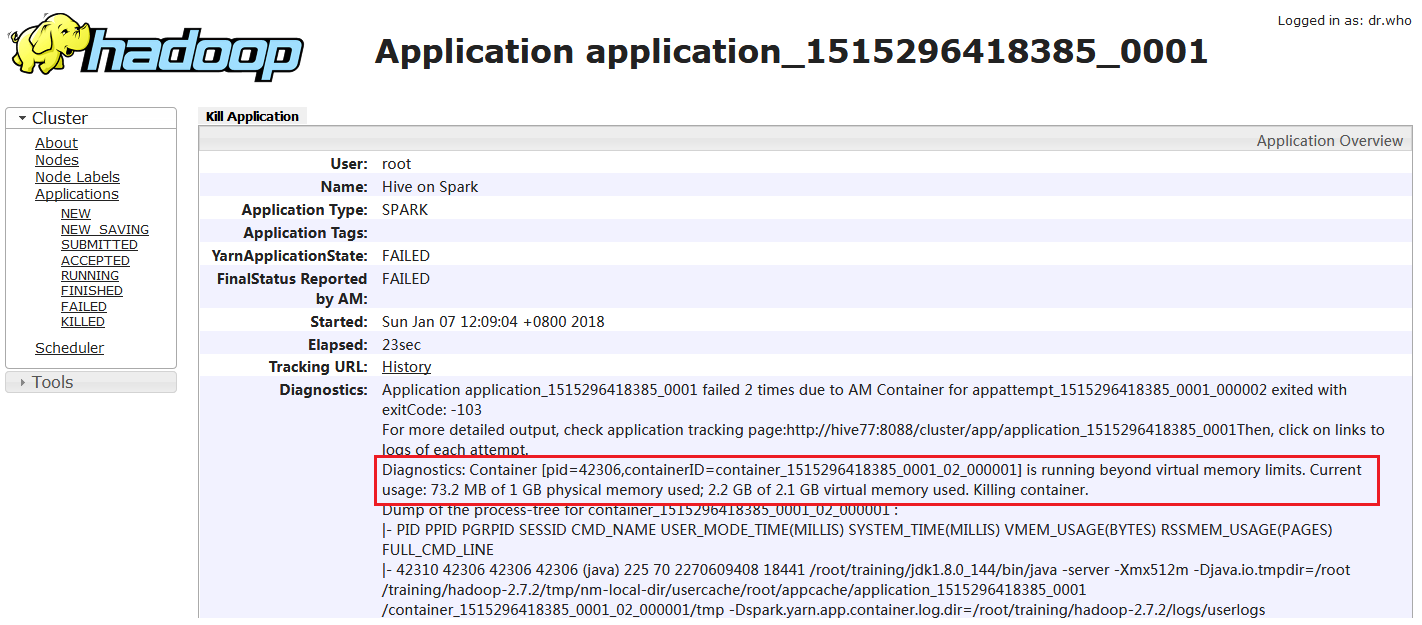


* 1. 导入emp.csv文件：
* load data local inpath '/root/temp/emp.csv' into table emp1;
  1. 执行查询，按照员工薪水排序：（**执行失败**）
* select \* from emp1 order by sal;



* 1. 检查Yarn Web Console





**该错误是由于是Yarn的虚拟内存计算方式导致，可在yarn-site.xml文件中，将yarn.nodemanager.vmem-check-enabled设置为false，禁用虚拟内存检查。**

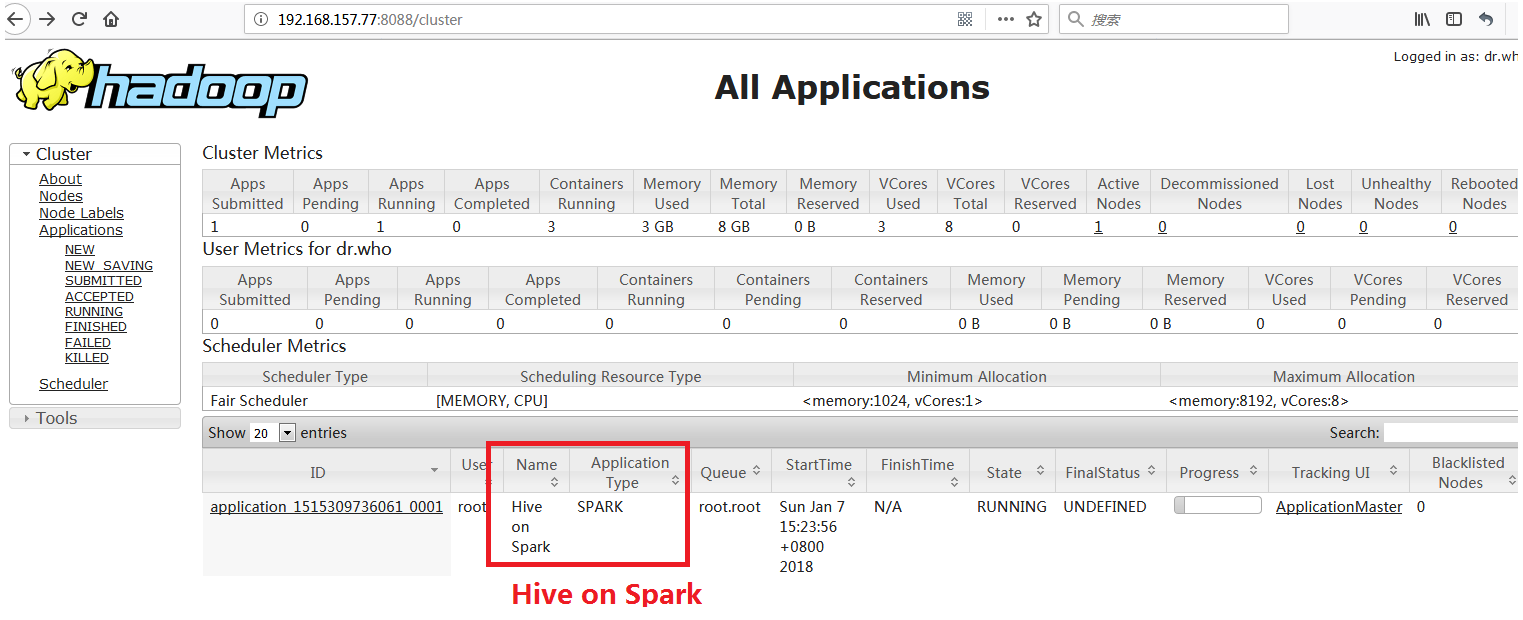
**<property>**

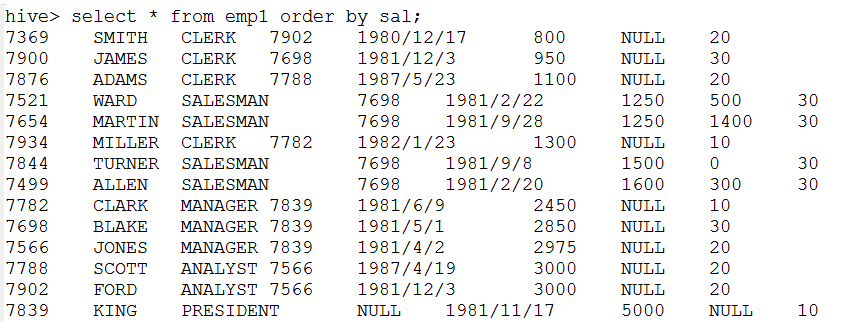
**<name>yarn.nodemanager.vmem-check-enabled</name>**

**<value>false</value>**

**</property>**

* 1. 重启：Hadoop、Spark、Hive，并执行查询





**最后说明一下：由于配置好了Spark on Yarn，我们在执行Hive的时候，可以不用启动Spark集群，因为此时都有Yarn进行管理。**