ssh免密码登录

简单操作，也就是将本机的公钥拷贝到想要免登陆的机器上。

ssh-copy-id -i worker1

ssh-copy-id -i worker2

参考：

<https://blog.csdn.net/fanrenxiang/article/details/69212647>

hadoop配置文件

1. hadoop\_env.sh

配置java 环境变量 export JAVA\_HOME=/opt/java

2. core-site.xml

3. hdfs-site.xml

启动hdfs

先要格式化文件系统

./bin/hdfs namenode –format

然后启动hdfs

./sbin/start-dfs.sh

设置Yarn服务

设置Yarn服务、启动相关组件、运行和操作Yarn集群。

1. 启动ResourceManager守护进程和NodeManager守护进程

$sbin/start-yarn.sh

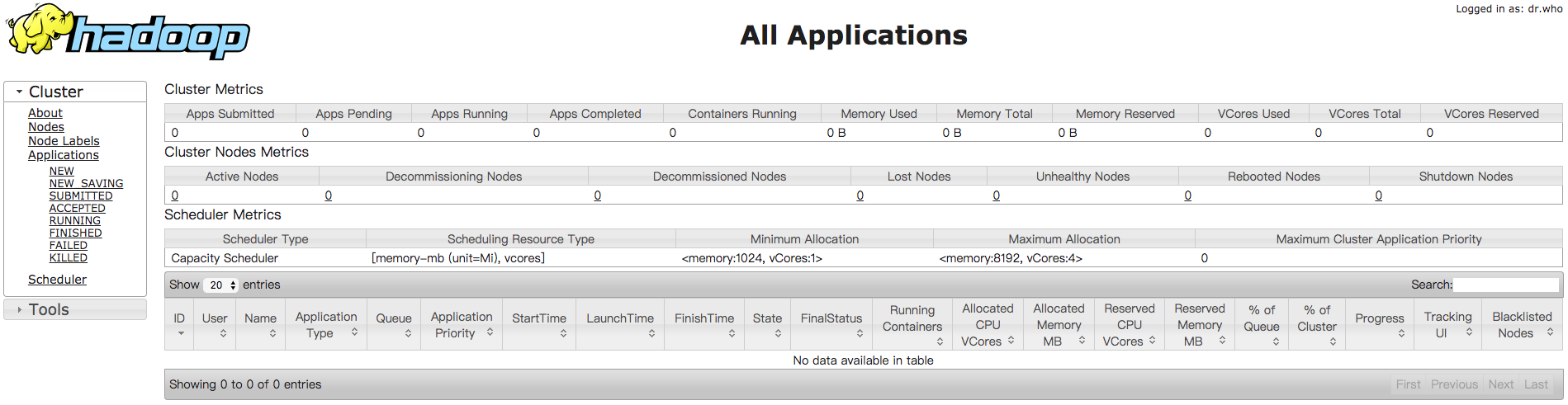
2、浏览ResourceManager的web界面，默认状态下地址为<http://localhost:8088>

3、运行MapReduce任务

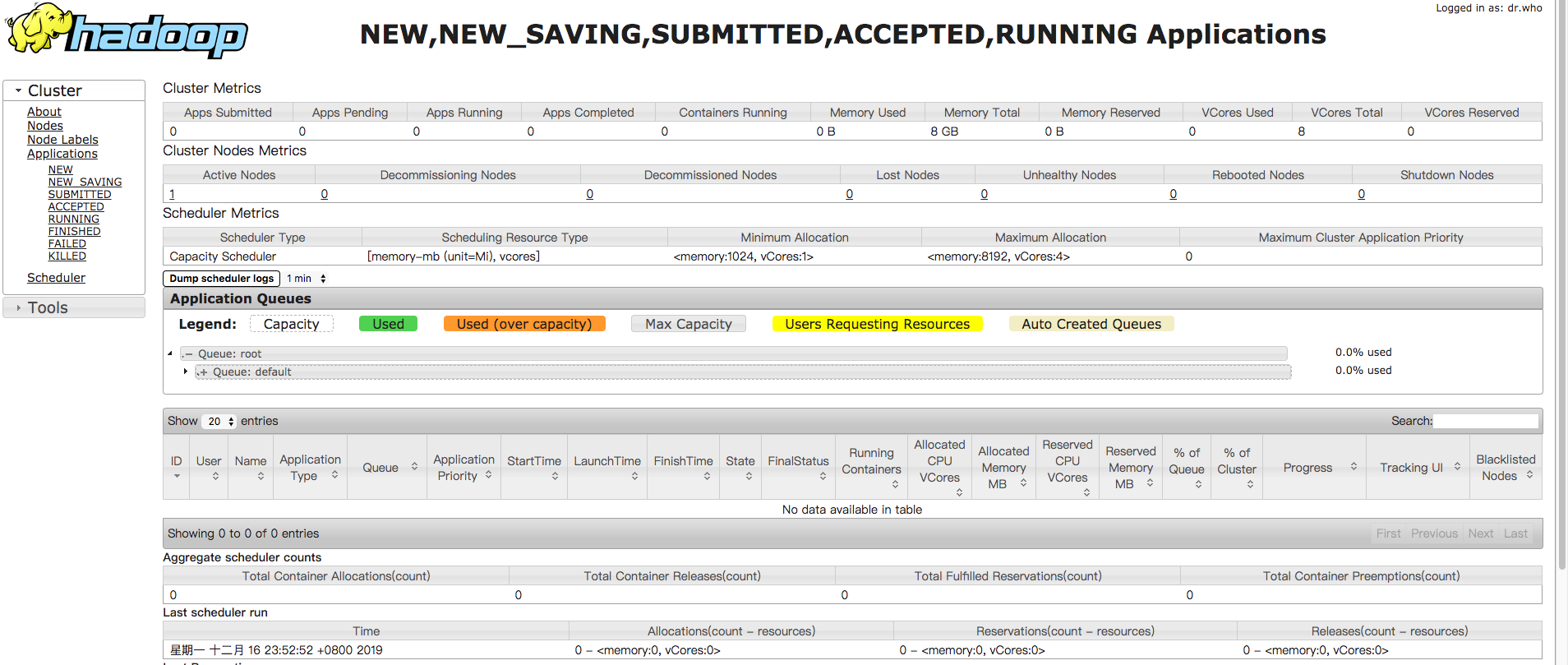
4、结束后，终止守护进程，运行：

$sbin/stop-yarn.sh

下图显示Yarn ResourceManager。



下图显示了集群中的资源队列，以及处于运行状态下的应用程序。除此之外，还可以在此处查看和监视运行中的各项任务。



Yarn

图片包含 文字, 地图

描述已自动生成

hive

下载安装hive-3.1.2

<https://hive.apache.org/downloads.html>

hive-3.1.2版本与Hadoop3适配。

cd /usr/local/Cellar/hive/3.1.1/libexec/conf

cp hive-default.xml.template hive-site.xml #原始配置文件备份



为什么使用 Hive

直接使用 MapReduce 所面临的问题：

　　1、人员学习成本太高

　　2、项目周期要求太短

　　3、MapReduce实现复杂查询逻辑开发难度太大

为什么要使用 Hive：

　　1、更友好的接口：操作接口采用类 SQL 的语法，提供快速开发的能力

　　2、更低的学习成本：避免了写 MapReduce，减少开发人员的学习成本

　　3、更好的扩展性：可自由扩展集群规模而无需重启服务，还支持用户自定义函数

Hive 特点

优点：

　　1、可扩展性,横向扩展，Hive 可以自由的扩展集群的规模，一般情况下不需要重启服务 横向扩展：通过分担压力的方式扩展集群的规模 纵向扩展：一台服务器cpu i7-6700k 4核心8线程，8核心16线程，内存64G => 128G

　　2、延展性，Hive 支持自定义函数，用户可以根据自己的需求来实现自己的函数

　　3、良好的容错性，可以保障即使有节点出现问题，SQL 语句仍可完成执行

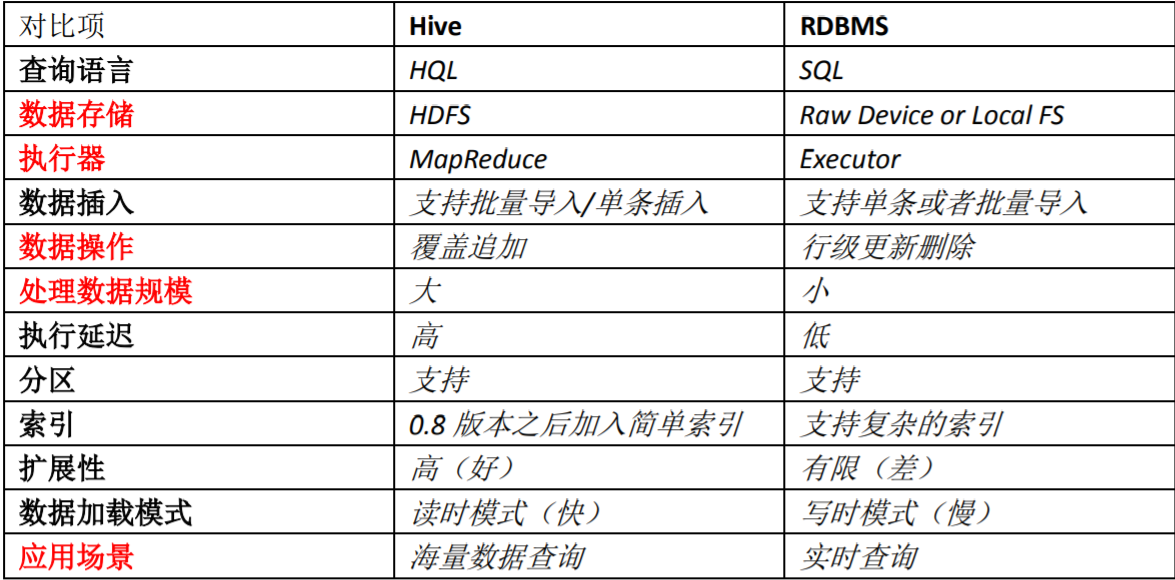
缺点：

　　1、Hive 不支持记录级别的增删改操作，但是用户可以通过查询生成新表或者将查询结 果导入到文件中（当前选择的 hive-2.3.2 的版本支持记录级别的插入操作）

　　2、Hive 的查询延时很严重，因为 MapReduce Job 的启动过程消耗很长时间，所以不能 用在交互查询系统中。

　　3、Hive 不支持事务（因为不没有增删改，所以主要用来做 OLAP（联机分析处理），而 不是 OLTP（联机事务处理），这就是数据处理的两大级别）。

Hive 和 RDBMS 的对比



总结：

　　Hive 具有 SQL 数据库的外表，但应用场景完全不同，Hive 只适合用来做海量离线数 据统计分析，也就是数据仓库。

Hive的架构

图片包含 文字

描述已自动生成

Pip

下载安装Pig-0.17.0

<https://pig.apache.org/releases.html#Download>