# 1 环境搭建记录

-- 版本

hadoop 3.1.3

hive 3.1.3

spark 2.4.4

-- 主要问题如下：

hadoop使用非root安装，否则启动时会报错,具体步骤参考《03 hadoop集群搭建.docx》

hive里的guava-19.0.jar要改成hadoop里的版本，具体步骤参考《05 hive搭建.docx》

spark使用hive做元数据管理时，解压即可，然后把hdfs-site.xml、yarn-site.xml、hive-site.xml拷贝到spark的conf目录里。hive-site.xml里的数据库连接池要改成dbcp

-- 常用端口

yarn 任务监控8088

hdfs 页面 9870（与2.xx不一样了）

spark-ui 默认8080，如果被占用，依次增加1

spark-master 默认7077

spark-job 默认4040

-- 配置更改

一定要把JAVA\_HOME和HADOOP\_HOME写到spark-env.sh里

一pid路径配置

默认放在/tmp下，带来的问题是万一被删除，就没法通过stop命令来停止集群

编辑spark-env.sh 配置属性SPARK\_PID\_DIR

一日志路径配置

编辑spark-env.sh 配置属性SPARK\_LOG\_DIR

# 2 问题记录

搭建很简单，拷贝几个xml配置文件到spark的conf目录下，启动一下yarn就行了

配置好spark on yarn之后，执行任务总是报错client.TransportClient: Failed to send RPC RPC java.nio.channels.ClosedChannelException

查了一下，好像是内存分配不够的问题，但是内存有200多G呢，怎么会不够呢？

按如下方法解决了，但是需要看究竟怎么回事：

<property>

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

<value>false</value>

</property>

<property>

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

<value>false</value>

</property>

# 3 补充信息

spark集群搭建，生产模式的cluster manager采用yarn还是standalone？

如果用standalong，那么HA应该通过zookeeper实现，在spark这边做配置

如果使用yarn，则yarn自生有HA，或者zookeeper，不需要再spark这边做配置

对于standalong模式：

提交sql文件执行的时候，如果使用命令 spark-sql --master spark://{host}:{port} -f test.sql

报错

Cluster deploy mode is not applicable to Spark SQL shell

因为交互式的关系，deploy-mode不能使用cluster，使用默认的client模式，使用如下即可

spark-sql --master spark://{host}:{port} -f test.sql

此外，如果配置了HA，那么可以这样提交：

spark-sql --master spark://{host1}:{port1},{host2}:{port2}... -hivevar -f test.sql

对于yarn模式：

提交任务时指定 --master yarn就行了

特别注意，yarn模式下是不需要启动worker，提交任务之后，yarn会根据slaves文件自动启动相应的worker

# 4 Spark编程

-- 提示

如在集群外访问，只能访问standalong模式的集群；yarn模式集群的只能打包提交执行。

此外，提交集群的话，应该就不必写master了。

## 4.1 开发环境

-- 地址配置：

在集群中所有机器的hosts文件中配置好window机器的ip hostname，否则会报错Failed to connect to DESKTOP-IAK1UQ8:62298

-- 用户配置：

在集群中所有机器新建window机器的用户

-- 权限配置：

如果有数据写入需求，则要么关闭hadoop的鉴权，要么在各个节点上配置window的用户并添加到hadoop的代理组里

-- 拷贝配置文件：

hdfs-site.xml core-site.xml hive-site.xml拷贝到resource目录下

-- 配置maven依赖：

<dependency>

<groupId>org.apache.spark</groupId>

<artifactId>spark-core\_2.11</artifactId>

<version>2.4.4</version>

</dependency>

<dependency>

<groupId>org.apache.spark</groupId>

<artifactId>spark-sql\_2.11</artifactId>

<version>2.4.4</version>

</dependency>

<dependency>

<groupId>org.apache.spark</groupId>

<artifactId>spark-hive\_2.11</artifactId>

<version>2.4.4</version>

</dependency>

-- 服务启动：

必须要启动worker，否则报错

Initial job has not accepted...

## 4.2 示例

连接standalone集群

SparkSession sparkSession = SparkSession.*builder*().  
 appName(**"test20191114"**).  
 config(**"hive.metastore.uris"**, **"thrift://10.204.145.35:9083"**).  
 enableHiveSupport().  
 getOrCreate();  
Dataset<Row> ds = sparkSession.sql(**"show databases"**);  
ds.show();

-- 示例 提交yarn集群

SparkSession sparkSession = SparkSession.*builder*().  
 appName(**"test20191114"**).  
 *// config("hive.metastore.uris", "thrift://10.204.145.35:9083").* enableHiveSupport().  
 getOrCreate();  
Dataset<Row> ds = sparkSession.sql(**"show databases"**);  
ds.show();

# 5依赖其它的jar文件怎么办

5.1 以HBase为例

至少两种方式：

1. 安装环境层面

把依赖的jar放到${SPARK\_HOME}/jars/目录下即可。

以上是根据spark执行方式发现的。

1. 执行方式层面

添加参数 --jars a.jar,b.jar,c.jar ...多个jar用逗号隔开

具体使用方式可以参考以下文件

