

Hadoop常见问题

Hadoop常见问题

1.安装Hadoop3.x常见问题

问题1.1.VMware报VMware Workstation 不可恢复错误: (vcpu-1) Exception

步骤1: 关闭windows自身的虚拟化Hyper-V

步骤2: 关闭内存完整性

步骤3: 重启电脑即可

问题1.2,安装虚拟机提示“无法安装服务VMAuthdService”, 请确保您有足够的权限安装系统服务

问题1.3,WMware的虚拟网络配置中, 无法成功保存将主机虚拟适配器连接到此网络选项

问题1.4, 在安装WMware过程中提示“Failed to install the hcmon driver”, 并且安装失败

问题1.5, 如何修改HDFS副本数

问题1.6, 克隆了一个虚拟机后不能重启它的网络服务

现象: 使用ip addr 查看, 没有IP地址, 即例ens33设置正确了也没有用

解决办法:

问题1.7, 集群启动后在WEB端未能查看到从机, 可能是多次格式化HDFS原因

在所有主机和从机上操作

在Master主机上操作

问题1.8, 在安装NTP服务时报错

错误提示如: Could't open file /media/repodata/repomd.xml

错误截图:

错误原因: /media目录没有挂载虚拟光驱

处理1:

处理1:

问题1.9Failed to replace a bad datanode

错误信息

出现原因

解决报错

问题1.10NoClassDefFoundError: org/sparkproject/guava/base/Preconditions

问题1.11 ClassNotFoundException:org.apache.spark.kafka010.KafkaConfigUpdater

参考

处理:

问题1.12 NoClassDefFoundError:com/google/common/util/concurrent/ThreadFactoryBuilder

问题1.13 java.io.IOException: All datanodes are bad. Aborting...

问题1.14 如何退出安全模式

查看hdfs safe mode

退出hdfs safe mode

问题1.15 如何降低hdfs上权限

问题1.16 如何进行HDFS块修复

检测缺失块

查看某一个文件的情况

手动修复

删除所有损坏的块

问题1.17 如何进行HDFS重新格式化

在所有主机和从机上操作

在Master主机上操作

问题1.18 如何解决HDFS磁盘空间不足

现象

处理

问题1.19 如何进行HADOOP问题定位

定位从机的问题

定位主机master的问题

问题1.20 如何设置Linux时区设置

2.IDEA运行MR常见问题

问题2.1： EBADF: Bad file descriptor

问题2.2： IDEA Compilation failed internal java compiler error

原因一： 多处的JDK的版本不匹配

查看项目的jdk

查看工程的jdk

查看java编译器版本：

视频观看

原因二： 编译器内部错误

问题2.3： IDEA上运行MR程序日志很少或没有，如何添加日志

问题2.4： cannot be cast to com.google.protobuf.Message

问题2.5： 在IDEA上运行MR程序是报access0错误

3.Hadoop集群上运行MR常见问题

问题3.1： 提示NoClassDefFoundError: com/mongodb/DBObject

4.Hive集群上安装常见问题

问题4.1： needs *MB on the / filesystem

问题4.2： 运行Hive时报错： unable to instantiate

org.apache.hadoop.hive.ql.metadata.SessionHiveMeta

1.安装Hadoop3.x常见问题

问题1.1.VMware报VMware Workstation 不可恢复错误: (vcpu-1) Exception

步骤1： 关闭windows自身的虚拟化Hyper-V

VMware Workstation 的虚拟化与 window10 自带的 Hyper-V 服务发生了冲突，关闭 windows 的 Hyper-V 功能，并稍后重启。

具体步骤：

以管理员身份运行 CMD，输入如下命令：

```
1 | bcdedit /set hypervisorlaunchtype off
```

注意：若需要使用docker时，需要将Hyper-V打开

步骤2： 关闭内存完整性

打开windows安全中心，选择设备安全性，然后选取内核隔离

Windows 安全中心

←

≡

🏠 主页

🛡️ 病毒和威胁防护

👤 帐户保护

🔒 防火墙和网络保护

📁 应用和浏览器控制

📁 设备安全性

💓 设备性能和运行状况

👨‍👩‍👧‍👦 家庭选项

🕒 保护历史记录

🛡️ 内核隔离

使用基于虚拟化的安全性的设备上可用的安全功能。

此更改需要你重启设备。

内存完整性

防止攻击将恶意代码插入到高安全性进程中。

⚠️

内存完整性已关闭。你的设备可能易受 [忽略](#) 攻击。

●

关

[了解更多信息](#)

内存访问保护

恶意的外部设备免受攻击保护你的设备的内存。

[了解更多信息](#)

有什么疑问?
[获取帮助](#)

帮助改进 Winc
[提供反馈](#)

更改你的隐私
查看和更改 Wi
设备的隐私设置
[隐私设置](#)
[隐私仪表板](#)
[隐私声明](#)

关闭内核完整性检验

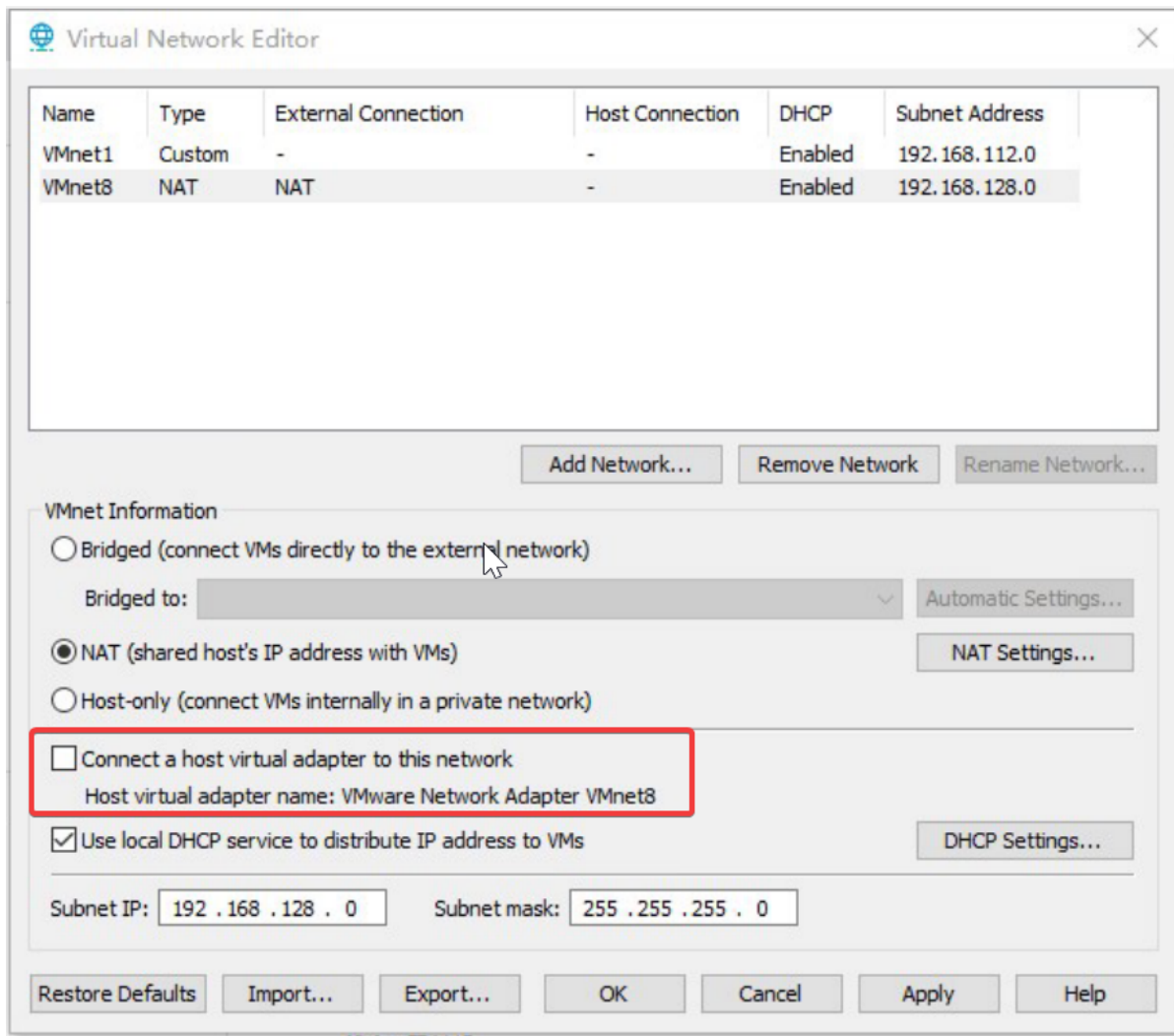
步骤3: 重启电脑即可

问题1.2,安装虚拟机提示“无法安装服务VMAuthdService”，请确保您有足够的权限安装系统服务

以管理员权限启动cmd（cmd在C:/windows/system32目录下 从运行-cmd出的是普通权限）

```
1 | sc delete VMUSBarbService
2 | net stop VMAuthdService
3 | taskkill /F /IM mmc.exe
4 | sc delete VMAuthdService （会报1060）
```

问题1.3,WMware的虚拟网络配置中，无法成功保存将主机虚拟适配器连接到此网络选项

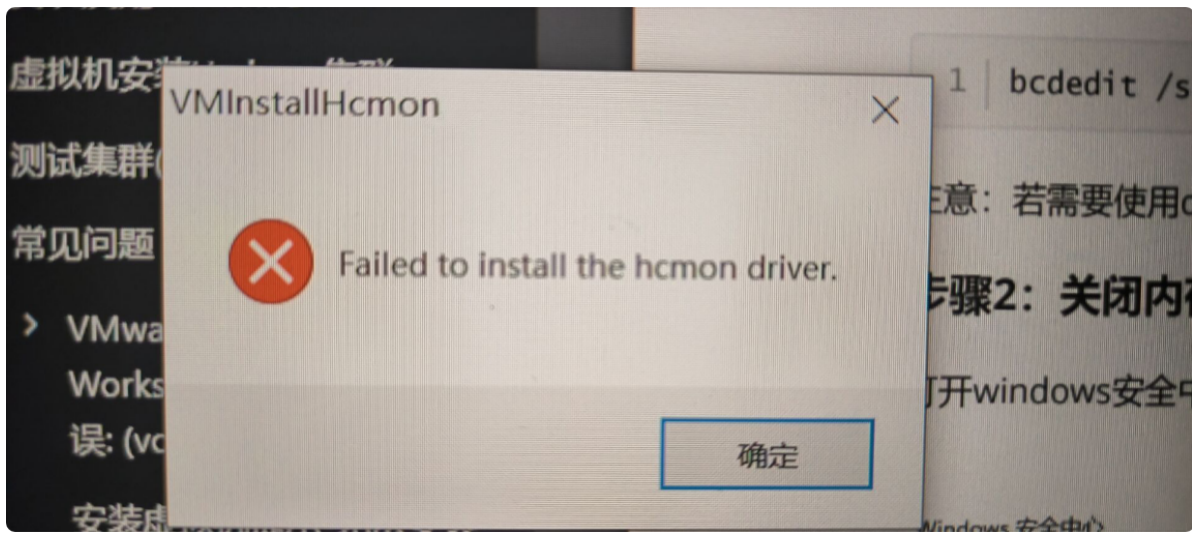


问题处理参考：<https://blog.csdn.net/MerryMaking7946/article/details/120555787>

处理如下：

- 1, 下载[CCleaner](#)
- 2, 打开CCleaner, 进行以下两项操作
 - 1->定制清理
 - 2->无用注册表清理
- 3, 重启虚拟机, 再进行虚拟网络设置, 再选中“Connect a host virtual adapter to this network”, 然后保存观察

问题1.4, 在安装WMware过程中提示“Failed to install the hcmon driver”, 并且安装失败



问题处理参考: <https://blog.csdn.net/eamonwu68/article/details/130356100>

参考2: <https://code84.com/716345.html>

问题5, 启动集群过程报ERROR:Cannot set priority of datanode process

```
Last login: Wed Sep 20 23:48:54 CST 2023 on pts/0
[root@master sbin]#
[root@master sbin]# mapred --daemon stop historyserver
[root@master sbin]#
[root@master sbin]# 1
-bash: 1: command not found
[root@master sbin]# cd $HADOOP_HOME/sbin/
[root@master sbin]# ./start-dfs.sh
Starting namenodes on [master]
Last login: Wed Sep 20 23:48:55 CST 2023 on pts/0
Starting datanodes
Last login: Wed Sep 20 23:50:07 CST 2023 on pts/0
slave1: ERROR: Cannot set priority of datanode process 7319
slave2: ERROR: Cannot set priority of datanode process 7205
slave3: ERROR: Cannot set priority of datanode process 7187
Starting secondary namenodes [master]
Last login: Wed Sep 20 23:50:10 CST 2023 on pts/0
[root@master sbin]#
```

问题1.5, 如何修改HDFS副本数

在Master主机上操作:

```
1 | cd /usr/local/hadoop-3.3.1/etc/hadoop/
2 | vi ./hdfs-site.xml
```

将文件 hdfs-site.xml 的 dfs.replication 的 values 由3改为 2:

```
1 <property>
2     <name>dfs.replication</name>
3     <value>3</value>
4 </property>
```

然后:wq保存退出；再重启集群：

```
1 stop-all.sh
2 start-all.sh
```

问题1.6，克隆了一个虚拟机后不能重启它的网络服务

现象：使用ip addr 查看，没有IP地址，即例ens33设置正确了也没有用

编辑IP配置文件：

vi /etc/sysconfig/network-scripts/ifcfg-ens33

重新修改了ip后，发现还是报错如下：

```
1 systemctl status network.service
```

```
-- Unit network.service has begun starting up.
Aug 04 19:56:25 tyf00 NetworkManager[839]: <warn> [1596542185.6957] ifcfg-rh: missing PREFIX, a
Aug 04 19:56:25 tyf00 network[1649]: [ ]
Aug 04 19:56:25 tyf00 NetworkManager[839]: <warn> [1596542185.8145] ifcfg-rh: missing PREFIX, a
Aug 04 19:56:25 tyf00 NetworkManager[839]: <info> [1596542185.8431] agent-manager: req[0x55e81e636e
Aug 04 19:56:25 tyf00 NetworkManager[839]: <info> [1596542185.8437] audit: op="connection-activate"
Aug 04 19:56:25 tyf00 network[1649]: [ ] ens33 [ ] No suitable device fou
Aug 04 19:56:25 tyf00 network[1649]: [ ]
Aug 04 19:56:25 tyf00 NetworkManager[839]: <info> [1596542185.8940] agent-manager: req[0x55e81e636e
Aug 04 19:56:25 tyf00 NetworkManager[839]: <info> [1596542185.8949] audit: op="connection-activate"
Aug 04 19:56:25 tyf00 network[1649]: [ ] ens34 [ ] No suitable device fou
Aug 04 19:56:25 tyf00 network[1649]: [ ]
Aug 04 19:56:25 tyf00 network[1649]: RTNETLINK answers: File exists
Aug 04 19:56:25 tyf00 network[1649]: RTNETLINK answers: File exists
Aug 04 19:56:25 tyf00 network[1649]: RTNETLINK answers: File exists
Aug 04 19:56:25 tyf00 network[1649]: RTNETLINK answers: File exists
Aug 04 19:56:25 tyf00 network[1649]: RTNETLINK answers: File exists
Aug 04 19:56:25 tyf00 network[1649]: RTNETLINK answers: File exists
Aug 04 19:56:25 tyf00 network[1649]: RTNETLINK answers: File exists
Aug 04 19:56:25 tyf00 network[1649]: RTNETLINK answers: File exists
Aug 04 19:56:25 tyf00 systemd[1]: network.service: control process exited, code=exited status=1
Aug 04 19:56:25 tyf00 systemd[1]: Failed to start LSB: Bring up/down networking.
-- Subject: Unit network.service has failed
```

错误信息：Failed to start LSB 网络服务启动失败

解决办法：

```
1 systemctl stop NetworkManager
2 systemctl disable NetworkManager
3 systemctl restart network
```

查看网络服务：

```
1 systemctl status network.service
```

好啦，又可以启动网络服务了

问题1.7，集群启动后在WEB端未能查看到从机，可能是多次格式化HDFS原因

解决办法：

在所有主机和从机上操作

1. 删除所有从机的/data/hadoop目录
2. 删除hadoop目录下的/tmp目录
3. 删除log目录下的临时目录

```
1 rm -rf /data/hadoop
2 rm -rf /usr/local/hadoop-3.3.1/tmp/*
3 rm -rf /var/log/hadoop/tmp/dfs/data
```

在Master主机上操作

1. 停止集群
2. 格式化hdfs
3. 启动集群

```
1 stop-all.sh
2 hdfs namenode -format
3 start-all.sh
```

问题1.8，在安装NTP服务时报错

错误提示如： Could't open file /media/repodata/repomd.xml

错误截图：

```

try.
file:///media/repodata/repomd.xml: [Errno 256] No more mirrors to
e /media/repodata/repomd.xml"
[root@slave1 ~]# yum install -y ntp
Loaded plugins: fastestmirror
Loading mirror speeds from cached hostfile
* base: mirrors.ustc.edu.cn
* extras: ftp.sjtu.edu.cn
* updates: mirrors.ustc.edu.cn
base
file:///media/repodata/repomd.xml: [Errno 14] curl#37 - "couldn't open fi
e /media/repodata/repomd.xml"
Trying other mirror.

one of the configured repositories failed (CentOS-7 - Media),
and yum doesn't have enough cached data to continue. At this point the on
ly
safe thing yum can do is fail. There are a few ways to work "fix" this:

1. Contact the upstream for the repository and get them to fix the pr
oblem.

2. Reconfigure the baseurl/etc. for the repository, to point to a wor
king
upstream. This is most often useful if you are using a newer
distribution release than is supported by the repository (and the
packages for the previous distribution release still work).

3. Run the command with the repository temporarily disabled
yum --disablerepo=c7-media ...

4. Disable the repository permanently, so yum won't use it by default
. Yum
it
will then just ignore the repository until you permanently enable
again or use --enablerepo for temporary usage:

    yum-config-manager --disable c7-media
or
    subscription-manager repos --disable=c7-media

5. Configure the failing repository to be skipped, if it is unavailab
le.
Note that yum will try to contact the repo. when it runs most comm
ands,
so will have to try and fail each time (and thus, yum will be be m
uch
slower). If it is a very temporary problem though, this is often a
nice
compromise:

```

错误原因：/media目录没有挂载虚拟光驱

处理1：

执行linux命令

```
1 | mount /dev/sr0 /media
```

运行结果如下：

```

1 | [root@slave2 yum.repos.d]# mount /dev/sr0 /media
2 | mount: /dev/sr0 is write-protected, mounting read-only
3 | [root@slave2 yum.repos.d]#

```

但这个命令在虚拟机重启后会失效，为了重启后也能生效，需要修改/etc/fstab文件

可以使用 more /etc/fstab查看文件，如果没有行：“/dev/sr0 /media iso9660 defaults 0 0”这行，则执行下面echo命令：

```
1 | echo "/dev/sr0 /media iso9660 defaults 0 0" >> /etc/fstab
```

处理1:

不使用光驱的镜像源，使用外网的服务:

```
1 | cd /etc/yum.repos.d/  
2 | mv ./CentOS-Media.repo ./CentOS-Media.repo.bak
```

问题1.9Failed to replace a bad datanode

错误信息

java.io.IOException: Failed to replace a bad datanode on the existing pipeline due to no more good datanodes being available to try.

出现原因

因为，我的集群共有3个DataNode节点，而我设置的默认副本数是3个。在执行写入到HDFS的操作时，当我的一台Datanode写入失败时，它要保持副本数为3,它就会去寻找一个可用的DataNode节点来写入，可是我的流水线上就只有3台DataNode节点，所以导致会报错Failed to replace a bad datanode

查看自己已经存在的副本数:

```
1 | more /usr/local/hadoop/etc/hadoop/hdfs-site.xml
```

包含内容如:

```
1 | <property>  
2 | <name>dfs.replication</name>  
3 | <value>3</value>  
4 | </property>
```

解决报错

在hdfs-site.xml 文件中添加下面几行代码

```
1 | <property>  
2 | <name>dfs.client.block.write.replace-datanode-on-failure.policy</name>  
3 | <value>NEVER</value>  
4 | </property>
```

问题1.10NoClassDefFoundError: org/sparkproject/guava/base/Preconditions

- ❑原因,hadoop的common目录下包版本冲突
- ❑处理,将common新加的包移到../bak目录下, 然后一个个加回来测试
- ❑经测试, 发现是由spark-streaming-kafka-0-10-assembly_2.12引起的hadoop启不来
- ❑处理: 改使用spark-streaming-kafka-0-10_2.12

问题1.11 ClassNotFoundException:org.apache.spark.kafka010.Kafka ConfigUpdater

参考

<https://blog.csdn.net/dkl12/article/details/118637084#%E6%9C%80%E7%BB%88%E6%B7%BB%E5%8A%A0%E7%9A%84jar%E5%8C%85>

处理:

添加:

commons-pool2-2.11.1.jar

spark-token-provider-kafka-0-10_2.12-3.1.3.jar

jedis-3.0.1.jar

问题1.12 NoClassDefFoundError:com/google/common/util/concurrent/ ThreadFactoryBuilder

处理: 添加

guava-27.0-jre.jar

find / -name guava*

cp /opt/hadoop-3.2.1/share/hadoop/common/lib/guava-27.0-jre.jar /usr/local/spark/jars/

问题1.13 java.io.IOException: All datanodes are bad. Aborting...

在执行spark streaming程序时, 日志显示:

java.io.IOException: All datanodes [DatanodeInfoWithStorage[10.0.0.5:9866,DS-19fdef37-cfbe-4217-ae33-1291982e197d,DISK]] are bad. Aborting...

问题原因

- ❑是linux机器打开了过多的文件导致。 
- ❑用命令ulimit -n可以发现linux默认的文件打开数目为1024

❑查看docker主机:

```
[root@server4 ~]# ulimit -n  
1024  
[root@server4 ~]#
```

❑问题处理

❑增加可打开文件数

```
echo "* soft nofile 128000" >> /etc/security/limits.conf  
echo "* hard nofile 128000" >> /etc/security/limits.conf  
echo "* soft nproc 128000" >> /etc/security/limits.conf  
echo "* hard nproc 128000" >> /etc/security/limits.conf
```

❑重启后查看

```
[root@server4 ~]# ulimit -n  
128000  
[root@server4 ~]
```

问题1.14 如何退出安全模式

查看hdfs safe mode

```
1 | hadoop dfsadmin -safemode get
```

退出hdfs safe mode

```
1 | hadoop dfsadmin -safemode leave
```

问题1.15 如何降低hdfs上权限

给根目录加权限，在master或slave上执行都可以。

```
1 | hdfs dfs -chmod -R 777 /
```

问题1.16 如何进行HDFS块修复

检测缺失块

```
1 | hdfs fsck -list-corruptfileblocks  
2 | hdfs fsck /
```

查看某一个文件的情况

```
1 | hdfs fsck /路径/文件名 -locations -blocks -files
```

手动修复

```
1 | hdfs debug recoverLease -path 文件位置 -retries 重试次数
```

删除所有损坏的块

```
1 | hdfs fsck / -delete
```

问题1.17 如何进行HDFS重新格式化

在所有主机和从机上操作

```
1 | rm -rf /data/hadoop
2 | rm -rf /usr/local/hadoop-3.3.1/tmp/*
3 | rm -rf /var/log/hadoop/tmp/dfs/data
```

在Master主机上操作

```
1 | stop-all.sh
2 | hdfs namenode -format
3 | start-all.sh
```

问题1.18 如何解决HDFS磁盘空间不足

现象

在多次上传到Hdfs和下载文件到本地时出现no space情况

处理

处理:当前Linux至少分配50G空间,但实际使用为10G,可以扩大

步骤1: 创建分区vda3

```
1 | fdisk /dev/vda
2 | 进入后输入:n
3 | 然后选择主分区:p
4 | 使用默认分区号:3
5 | 回车默认
6 | 回车默认
7 | 使用wq退出
8 | 运行partprobe使用分区更改立即生效:
9 | partprobe
10
```

步骤2: 创建物理卷

```
1 | pvcreate /dev/vda3
```

步骤3: 扩展卷组

```
1 | vgextend centos /dev/vda3
```

步骤4: 扩展逻辑卷

```
1 | lvextend -L +10G /dev/centos/root
2 | xfs_growfs /dev/mapper/centos-root
```

问题1.19 如何进行HADOOP问题定位

定位从机的问题

以下在从机slave1或slave2上执行, 如果发现从机的jps检查时没有DataNode,NodeManager进程的话

```
1 | cd /usr/local/hadoop-3.1.4/logs/
2 | rm -rf hadoop*
```

然后在master主机上重启集群

```
1 | stop-all.sh
2 | start-all.sh
```

重启集群后, 查看slave1或slave2的日志

```
1 | cd /usr/local/hadoop-3.1.4/logs/
2 | more ./hadoop-root-datanode-slave1.log
3 | more ./hadoop-root-nodemanager-slave1.log
```

观察日志文件中是否有Error等关键信息, 再针对错误信息进行处理

定位主机master的问题

以下在master主机上进行, 如果使用jps检查master,发现少了NameNode等进程的话

```
1 | cd /usr/local/hadoop-3.1.4/logs/
2 | rm -rf hadoop*
3 | stop-all.sh
4 | start-all.sh
5 | more ./hadoop-root-namenode-master.log
6 | more ./hadoop-root-resourcemanager-master.log
```

观察日志文件中是否有Error等关键信息, 再针对错误信息进行处理

问题1.20 如何设置Linux时区设置

使用cp,替换当前时区为上海时区, 如有提示, 输入yes进行确认

```
1 cp /usr/share/zoneinfo/Asia/Shanghai /etc/localtime
```

2.IDEA运行MR常见问题

问题2.1: EBADF: Bad file descriptor

WARN io.ReadaheadPool: Failed readahead on ifile

EBADF: Bad file descriptor

如图：

```
mester_128 x @ slave_128 |@ slave0_128
```

```
Merged Map outputs=0  
GC time elapsed(ms)=275  
Total allocated heap usage (bytes)-126808064  
  
File Input Format Counter:  
Bytes Read-190923121  
2023-11-02 22:03:42 INFO mapred.local.jobrunner: Finishing task: attempt_local1144739933_0001_m_000000_0  
2023-11-02 22:03:42 INFO mapred.local.jobrunner: map task executor complete.  
2023-11-02 22:03:42 INFO mapred.local.jobrunner: Waiting for reduce tasks  
2023-11-02 22:03:43 INFO mapred.local.jobrunner: Starting task: attempt_local1144739933_0001_r_000000_0  
2023-11-02 22:03:43 INFO mapred.local.jobrunner: File Output Committer: File output commiter will cleanup temporary folders under output directory:false, ignore cleanup failures: false  
2023-11-02 22:03:44 INFO output.FileOutputCommitter: FileOutputCommitter skipCleanUpTemporayFolders under output directory:false, ignore cleanup failures: false  
2023-11-02 22:03:44 INFO TaskRunner: Using HadoopLocalFileSystem as local filesystem  
2023-11-02 22:03:44 INFO mapred.SystemTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.ShuffleB2D23d5234  
2023-11-02 22:03:44 WARN Impl.MetricSystemImpl: jobTracker metrics system already initialized.  
2023-11-02 22:03:45 INFO org.apache.hadoop.metrics.impl.JobMetricsReporter: Metrics are disabled by default due to excessive logging - disable via hadoop-metrics-core.properties  
2023-11-02 22:03:46 INFO EventFetcher: atorg.apache.hadoop.mapreduce.TaskInputOutputContext$ShuffleFlushLimit[1144739933_0001_r_000000_0] Thread started: EventFetcher for fetching Map Completion Events  
2023-11-02 22:03:46 INFO EventFetcher: about to shuffle file output of map attempt_local1144739933_0001_m_000000_0 decom: 410 len: 414 to MEMORY  
2023-11-02 22:03:46 INFO reduce.InMemoryMapOutput: Read 410 bytes from map-output for attempt_local1144739933_0001_m_000000_0  
2023-11-02 22:03:46 INFO reduce.MemoryManagerImpl: Write 410 bytes from memory-to-disk for attempt-local1144739933_0001_r_000000_0 write size: 410, InMemoryMapOutputs.size()=> 1, commfMemory-> 0, useMemory->410  
2023-11-02 22:03:50 ERROR java.io.FileNotFoundException: failed readhead on file /  
CAOR: java.lang.IllegalArgumentException: Invalid native method signature:  
at org.apache.hadoop.io.native.NativeIOIOSocketPosix.<jdk.internal.reflect.GeneratedMethodAccessor1>  
at org.apache.hadoop.io.native.NativeIOIOSocketPosix.<jdk.internal.reflect.GeneratedMethodAccessor1>  
at org.apache.hadoop.io.native.NativeIOIOSocketPosix.access$1$0(NativeIOIOSocketPosix.java:99)  
at org.apache.hadoop.io.ReadAheadHeaderRequest.run(ReadAheadHeaderRequest.java:50)  
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1132)  
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)  
at java.lang.Thread.run(Thread.java:833)  
  
2023-11-02 22:03:50 INFO reduce.EventFetcher: EventFetcher is interrupted.. returning  
2023-11-02 22:03:50 INFO mapred.local.jobrunner: r / i copied.  
2023-11-02 22:03:50 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs  
2023-11-02 22:03:50 INFO MergeManagerImpl: Merging 1 sorted segments  
2023-11-02 22:03:50 INFO merge.Reduce.Merge: Down to the last merge-pass, with 1 segments left of total size: 401 bytes  
2023-11-02 22:03:50 INFO reduce.MergeManagerImpl: Merged 1 segments, 410 bytes to disk to satisfy reduce memory limit  
2023-11-02 22:03:50 INFO reduce.MergeManagerImpl: Merging 1 files, 414 bytes from disk  
2023-11-02 22:03:50 INFO merge.Reduce.Merge: Merging 0 segments, 0 bytes from memory into reduce  
2023-11-02 22:03:50 INFO MergeManagerImpl: Merging 1 sorted segments  
2023-11-02 22:03:50 INFO merge.Reduce.Merge: Down to the last merge-pass, with 1 segments left of total size: 401 bytes  
2023-11-02 22:03:50 INFO mapred.local.jobrunner: r / i'ed.  
2023-11-02 22:03:50 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiptrecords  
2023-11-02 22:03:50 INFO mapreduce.TaskAttemptMapper: Saved output of task attempt_local1144739933_0001_r_000000_0 to hdfs://master:8020/user/mymame/output/tmpcount  
2023-11-02 22:03:50 INFO mapred.local.jobrunner: 1 / 1 copied.  
2023-11-02 22:03:50 INFO mapred.Task: Task attempt_local1144739933_0001_r_000000_0 is allowed to commit non-zero amount of data to hdfs://master:8020/user/mymame/output/tmpcount  
2023-11-02 22:03:50 INFO mapred.local.jobrunner: reduce * reduce  
2023-11-02 22:03:50 INFO mapred.Task: Task attempt_local1144739933_0001_r_000000_0 done  
2023-11-02 22:03:50 INFO mapred.Task: Final counters for Attempt_local1144739933_0001_r_000000_0: Counters: 29  
File System Summary:  
FILE: Number of bytes read=82473  
FILE: Number of bytes written=597467  
HDFS: Number of read operations=0  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=0  
HDFS: Number of bytes read=190923121  
HDFS: Number of bytes written=230  
HDFS: Number of read operations=0  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=0  
Map-Reduce Framework  
Combine Input records=0  
Combine output records=0  
Reduce Input groups=12
```

查阅信息后，说由于在快速读取文件的时候，文件被关闭引起，也可能是其他bug导致，此处忽略。也可以 `mapreduce.file.readahead = false` 临时禁掉

```
1 cd /root/hadoop
2 hadoop jar /root/hadoop/hadoop.jar chap5_tempcount.MaxTAndMinT \
3 -D mapreduce.ifile.readahead=false \
4 /user/myname/temp/temp2021.txt \
5 /user/myname/output_tempcount
```

问题2.2: IDEA Compilation failed internal java compiler error

IDEA在编译项目时报错

如图:

问题原因分两种：

原因一：多处的JDK的版本不匹配

导致这个错误的原因主要是因为jdk版本问题，此处有两个因素，一个是编译版本不匹配，一个是当前项目jdk版本不支持。

查看项目的jdk

File -> Project Structure -> Project Settings -> Project 或使用快捷键 Ctrl+Alt+shift+S 打开项目的jdk配置
要求：1， project jdk版本要求：1.8；2， project language level要求:8

查看工程的jdk

点击上例中 Modules（File -> Project Structure -> Project Settings -> Modules）查看对应jdk版本，其中 Language level要求:8

查看java编译器版本：

File -> Settings -> Build, Execution, Deployment -> Compiler -> Java Compiler

要求：1， Project bytecode version是：8；2， module的target bytecode version也是：8

视频观看

[如何检查JDK的版本是否是1.8?](#)

原因二：编译器内部错误

真的就是编译器内部错误，此时就得去查看 错误日志，

我之前遇到的错误：java.lang.OutOfMemoryError: GC overhead limit exceeded：

加载太多资源到内存，导致GC耗时较多

GC overhead limit exceeded检查是Hotspot VM 1.6定义的一个策略，通过统计GC时间来预测是否要OOM了，提前抛出异常，防止OOM发生。Sun 官方对此的定义是：“并行/并发回收器在GC回收时间过长时会抛出OutOfMemoryError。过长的定义是，超过98%的时间用来做GC并且回收了不到2%的堆内存。用来避免内存过小造成应用不能正常工作。”

听起来没啥用...预测OOM有啥用？起初开来这玩意只能用来Catch住释放内存资源，避免应用挂掉。后来发现一般情况下这个策略不能拯救你的应用，但是可以在应用挂掉之前做最后的挣扎，比如数据保存或者保存现场（Heap Dump）。

解决办法

1.增加参数-XX:-UseGCOverheadLimit，关闭这个特性，同时增加heap大小-Xmx1024m -Xms512m，系统环境变量 新增两行

```
1 _JAVA_OPTIONS
2 -Xms512m Xmx1024m
```

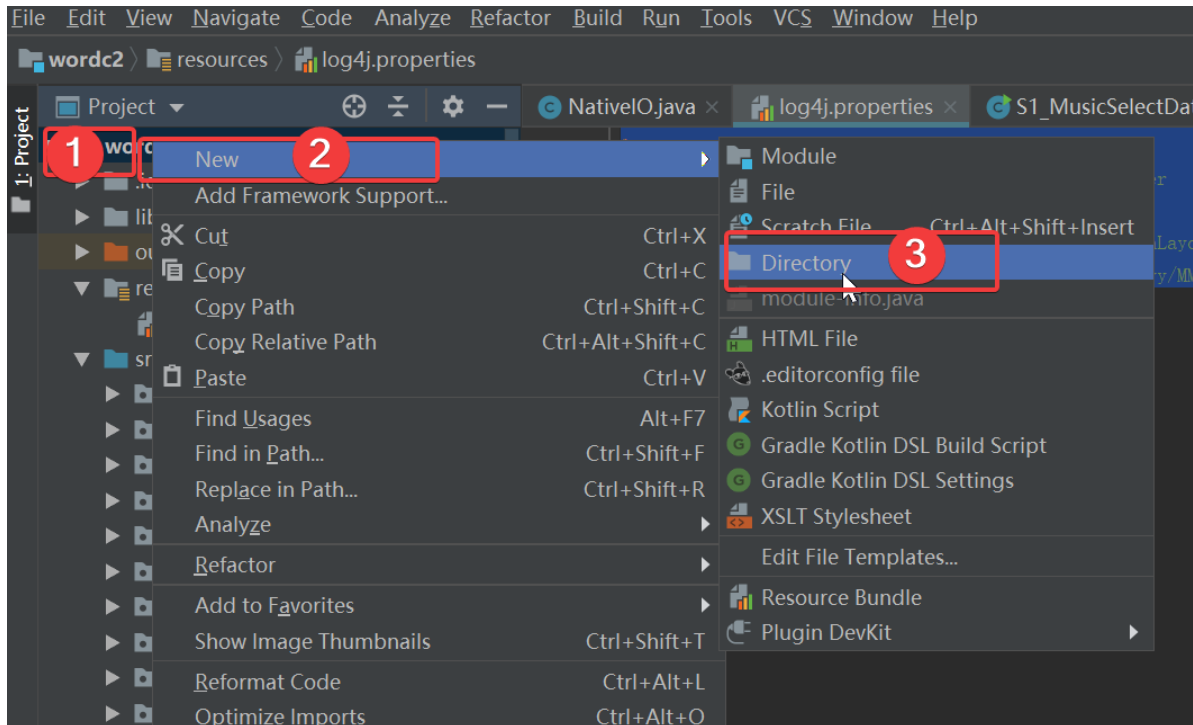
每个人的电脑配置不一样，上面那个适合电脑配置低的，配置完点击ok关闭环境变量窗口，重启idea！
编译完成后，将新加的环境变量参数删除、再重启，点击debug 或者run ok

将jdk升级到jdk1.8及以上版本，就能完全解决 java.lang.OutOfMemoryError: GC overhead limit exceeded 的问题

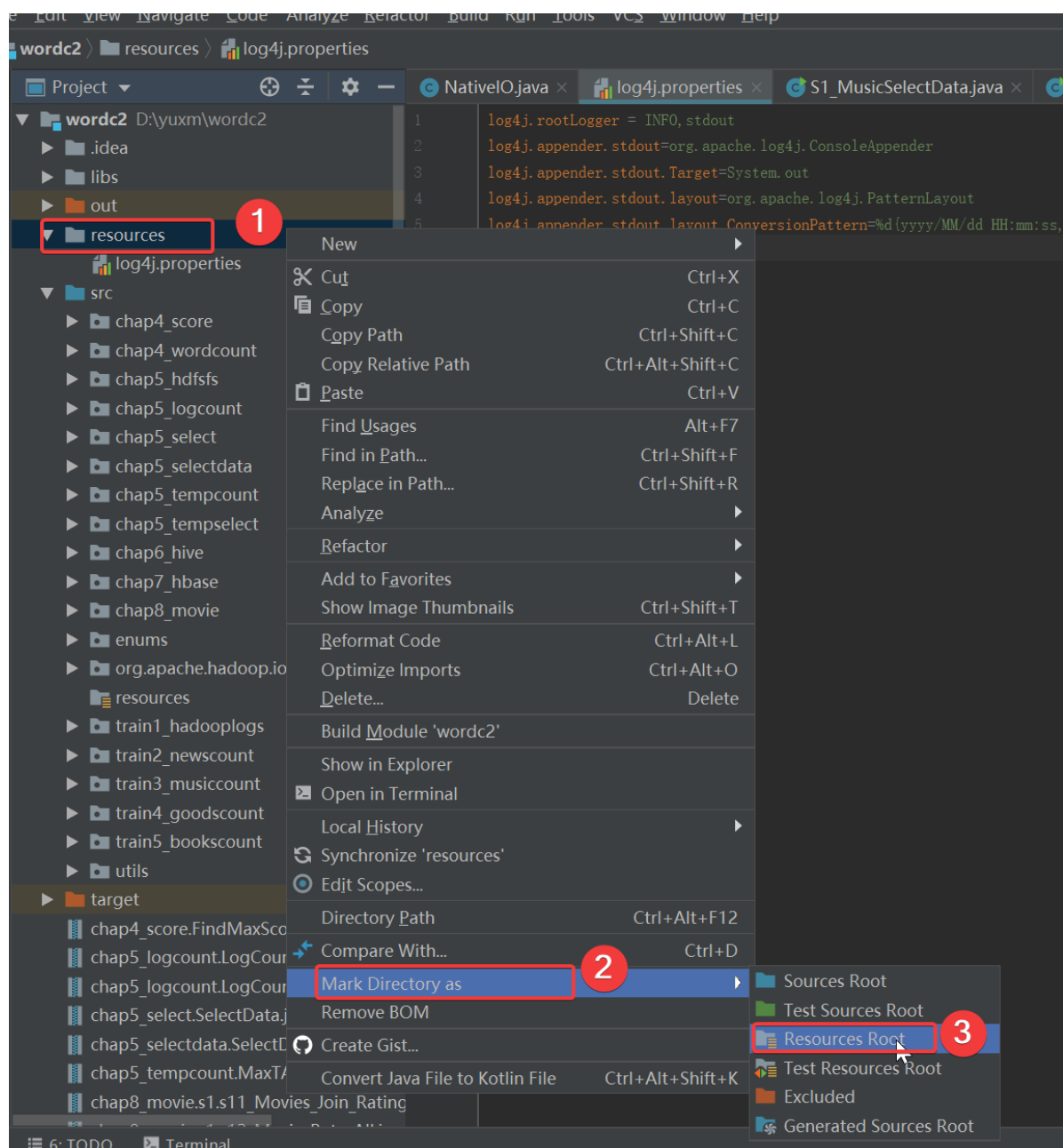
问题2.3: IDEA上运行MR程序日志很少或没有, 如何添加日志

IDEA上运行MR程序日志很少或没有, 如何添加日志?

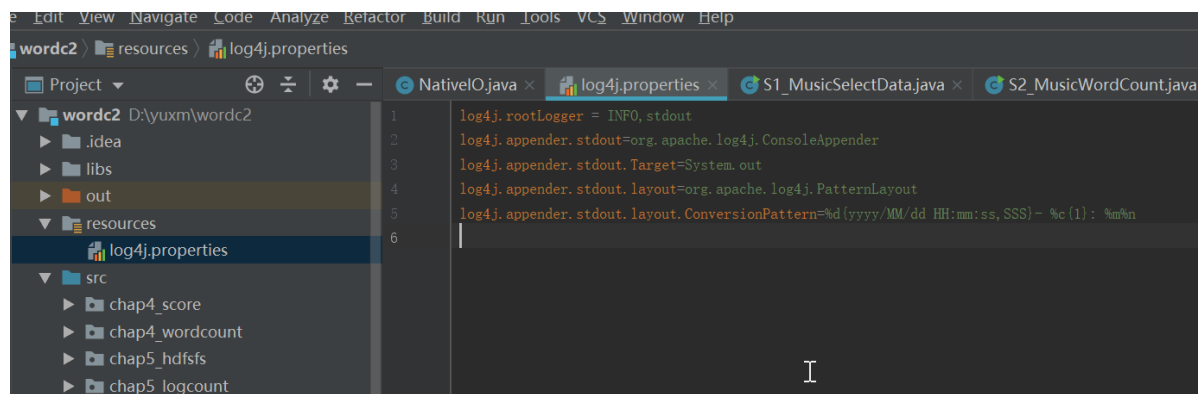
首先, 在IDEA项目目录下, 右击项目名称, 选择New->Directory, 添加resources目录



然后, 右击resources, 选择菜单: Make Directory as ->Resources Root



最后，在resources目录下，添加或从别处拷贝Log4j.properties文件，截图和内容如下：



```

1 log4j.rootLogger = INFO,stdout
2 log4j.appender.stdout=org.apache.log4j.ConsoleAppender
3 log4j.appender.stdout.Target=System.out
4 log4j.appender.stdout.layout=org.apache.log4j.PatternLayout
5 log4j.appender.stdout.layout.ConversionPattern=%d{yyyy/MM/dd HH:mm:ss,SSS}- %c{1}: %m%n

```

(注意：上面第5行，被转换成pdf格式会换行，同学们注意合并行)

问题2.4: cannot be cast to com.google.protobuf.Message

错误信息：Exception in thread "main" java.lang.ClassCastException:

org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos\$GetFileInfoRequestProto
cannot be cast to com.google.protobuf.Message

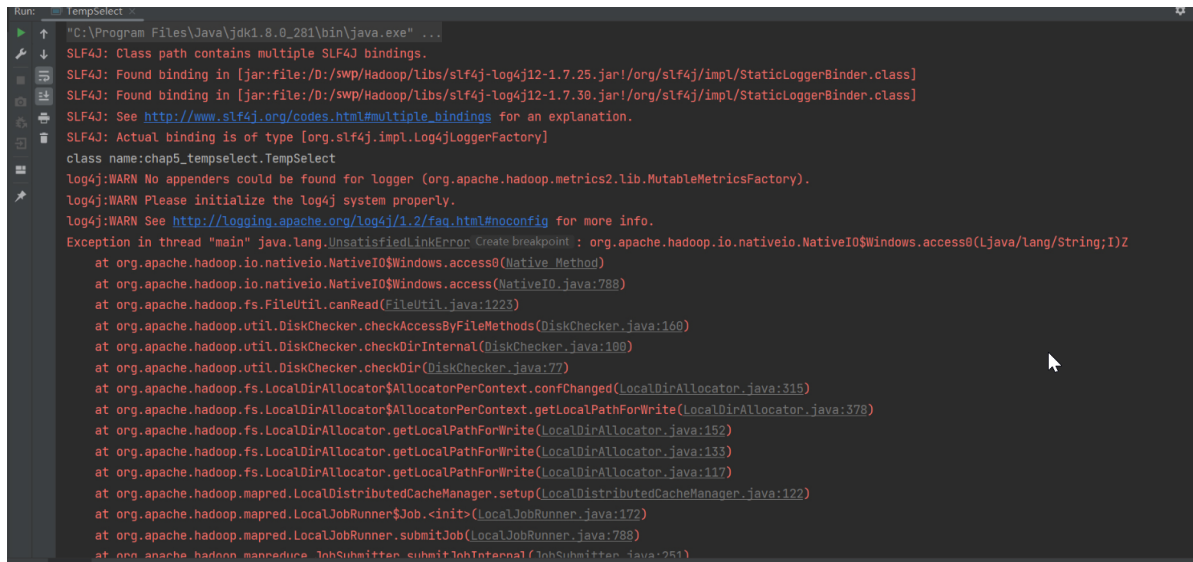
原因分析：缺少依赖包protobuf-java, protobuf-java-util, 或者是依赖包相冲突

问题处理：

1. 先删除项目根目录下的libs目录下的所有jar文件
2. 然后重新git clone 下来hadoop.git, 从hadoop目录下找到\hadoop\b57562\wordcount\libs,将libs的所有jar文件copy到项目根目录下的libs目录下
3. 重新右击libs目录, 在菜单中找到Add as Library;

问题2.5: 在IDEA上运行MR程序是报access0错误

错误信息, 提示access0.java/lang/string, Z错误, 如下截图:



错误原因, 某个java文件问题, 需要重写该java。

处理, 从gitclone下来包中将: hadoop\b57562\wordcount\src目录下的org目录, 复制到本项的源码根目录下, 比如src目录, 然后重新运行

3.Hadoop集群上运行MR常见问题

问题3.1: 提示NoClassDefFoundError: com/mongodb/DBObject

原因: 缺少Mongodb需要的依赖包

处理: 通过git clone下载hadoop.git的目录下: \hadoop\b57562\wordcount\libs下, 将mongo-hadoop-1.0.0.jar、mongo-hadoop-core-2.0.2.jar、mongo-java-driver-3.9.1.jar上传到集群master,slave1,slave2等主从机的目录:

```
1 | /usr/local/hadoop-3.1.4/share/hadoop/common
```

问题处理视频: [学习视频](#)

4.Hive集群上安装常见问题

问题4.1: needs *MB on the / filesystem

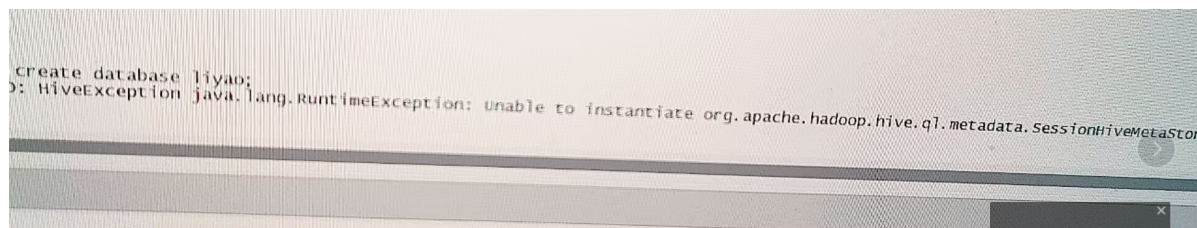
现象: 报错[root@master ~]# rpm -ivh mysql-community-* --force --nodeps
warning: mysql-community-client-5.7.18-1.el6.x86_64.rpm: Header V3 DSA/SHA1 Signature, key ID 5072e1f5: NOKEY
Preparing... ##### [100%]
installing package mysql-community-server-5.7.18-1.el6.x86_64 needs 621MB on the / filesystem

原因: 磁盘空间不足, 显示需要621M的系统空间, 但目前不足

解决: 释放大文件, 可用使用 du -sh ./ * 命令查询大文件, 然后删除大文件。

问题4.2: 运行Hive时报错: unable to instantiate org.apache.hadoop.hive.ql.metadata.SessionHiveMeta

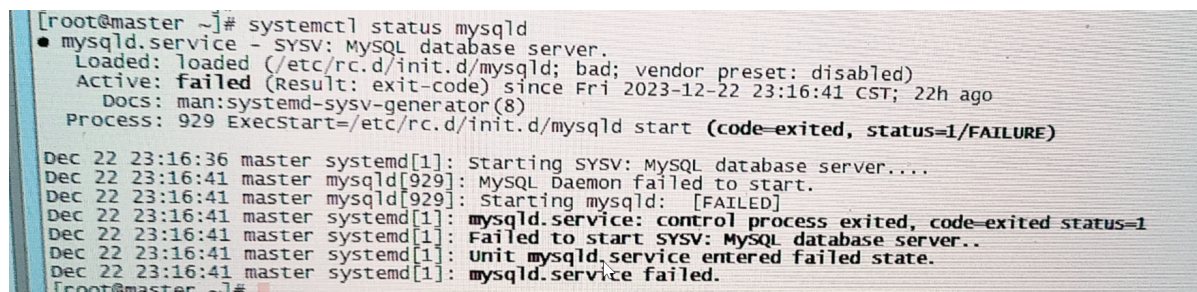
现象, 如图:



原因分析: 连接不上元数据, 这里我们一般使用的是mysql数据库, 所以需要检查一下mysql服务是否有起来

```
1 | systemctl status mysqld
```

结果如:



问题处理: 分析mysql的日志, 找寻问题的原因:

```
1 | tail -100 /var/log/mysqld.log
```

结果如

```

1 2023-12-11T04:31:09.470727Z 0 [Note] InnoDB: Loading buffer pool(s) from
  /var/lib/mysql/ib_buffer_pool
2 2023-12-11T04:31:09.471983Z 0 [Note] Plugin 'FEDERATED' is disabled.
3 2023-12-11T04:31:09.477237Z 0 [Note] InnoDB: Buffer pool(s) load completed
  at 231211 12:31:09
4 2023-12-11T04:31:09.506344Z 0 [Note] Found ca.pem, server-cert.pem and
  server-key.pem in data directory. Trying to enable SSL support using them.
5 2023-12-11T04:31:09.508537Z 0 [Warning] CA certificate ca.pem is self
  signed.
6 2023-12-11T04:31:09.509970Z 0 [Note] Server hostname (bind-address): '*';
  port: 3306
7 2023-12-11T04:31:09.510218Z 0 [Note] IPV6 is available.
8 2023-12-11T04:31:09.510228Z 0 [Note] - '::' resolves to '::';
9 2023-12-11T04:31:09.510249Z 0 [Note] Server socket created on IP: '::'.
10 2023-12-11T04:31:09.511892Z 0 [ERROR] /usr/sbin/mysqld: Can't create/write
  to file '/var/run/mysqld/mysqld.pid' (Errcode: 2 - No such file or
  directory)
11 2023-12-11T04:31:09.511912Z 0 [ERROR] Can't start server: can't create PID
  file: No such file or directory

```

错误原因是不能创建目录和文件： /usr/sbin/mysqld: Can't create/write to file
/var/run/mysqld/mysqld.pid

根据问题解决方法如下：

```

1 mkdir -p /var/run/mysqld/
2 chmod 777 -R /var/run/mysqld/
3 systemctl start mysqld
4 systemctl status mysqld

```

完成后，正常的结果如下，如第4行，active(running) 表示mysql启动正常，这样就可以再次进入hive了

```

1 [root@master ~]# systemctl status mysqld
2 • mysqld.service - SYSV: MySQL database server.
3   Loaded: loaded (/etc/rc.d/init.d/mysqld; bad; vendor preset: disabled)
4   Active: active (running) since Sat 2023-12-23 22:16:32 CST; 2s ago
5     Docs: man:systemd-sysv-generator(8)
6   Process: 81634 ExecStart=/etc/rc.d/init.d/mysqld start (code=exited,
  status=0/SUCCESS)
7   Main PID: 81868 (mysqld)
8     CGroup: /system.slice/mysqld.service
9             └─81662 /bin/sh /usr/bin/mysqld_safe --datadir=/var/lib/mysql --
  socket=/var/lib/mysql/mysql.sock --pid-file=/var/run/mysqld/mysqld.pid --
  basedir=/usr --user=mysql
10             └─81868 /usr/sbin/mysqld --basedir=/usr --datadir=/var/lib/mysql
  --plugin-dir=/usr/lib64/mysql/plugin --user=mysql --log-
  error=/var/log/mysqld.log --pid-file=/var/run/mysqld/mysqld.pid --
  socket=/var/lib/mys...
11
12 Dec 23 22:16:30 master systemd[1]: Starting SYSV: MySQL database server....
13 Dec 23 22:16:32 master mysqld[81634]: Starting mysqld: [ OK ]
14 Dec 23 22:16:32 master systemd[1]: mysqld.service: Supervising process 81868
  which is not our child. We'll most likely not notice when it exits.
15 Dec 23 22:16:32 master systemd[1]: Started SYSV: MySQL database server..

```

