

Windows配置Hadoop

目录

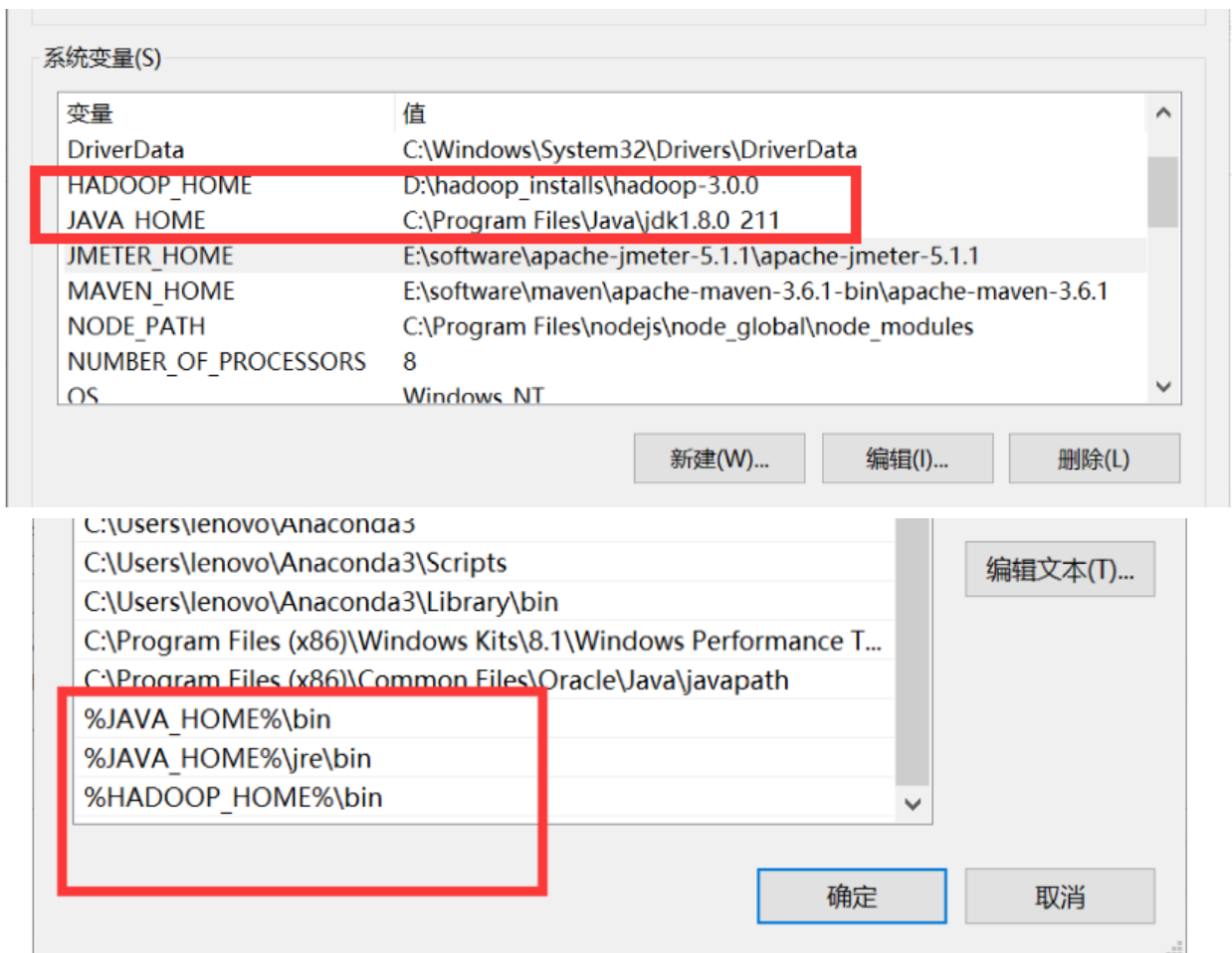
1. Windows下Hadoop-3.0.0安装运行
2. IntelliJ IDEA开发Hadoop伪分布式应用
3. IntelliJ IDEA配置HDFS可视化插件

Windows下Hadoop-3.0.0安装运行

1. 建立hadoop_installs文件夹，官网下载 hadoop-3.0.0 binary到该目录
2. 管理员权限打开shell，解压后出现hadoop-3.0.0目录

```
tar -zxvf hadoop-3.0.0.tar.gz
```

3. 环境变量配置



4. 添加winutils <https://github.com/stveloughran/winutils> clone到本地解压对应版本的替换bin目录（或者将winutils.exe直接放bin里）

查看

电脑 > 新加卷 (D:) > hadoop_installs > winutils >

| 名称 | 修改日期 | 类型 | 大小 |
|------------------|------------------|-------|----|
| .git | 2019/10/14 22:00 | 文件夹 | |
| hadoop-2.6.0 | 2019/10/14 22:00 | 文件夹 | |
| hadoop-2.6.3 | 2019/10/14 22:00 | 文件夹 | |
| hadoop-2.6.4 | 2019/10/14 22:00 | 文件夹 | |
| hadoop-2.7.1 | 2019/10/14 22:00 | 文件夹 | |
| hadoop-2.8.0-RC3 | 2019/10/14 22:00 | 文件夹 | |
| hadoop-2.8.1 | 2019/10/14 22:00 | 文件夹 | |
| hadoop-2.8.3 | 2019/10/14 22:00 | 文件夹 | |
| hadoop-3.0.0 | 2019/10/14 22:00 | 文件夹 | |
| .gitattributes | 2019/10/14 22:00 | 文本文档 | |
| .gitignore | 2019/10/14 22:00 | 文本文档 | |
| KEYS | 2019/10/14 22:00 | 文件 | 2 |
| LICENSE | 2019/10/14 22:00 | 文件 | 1 |
| README.md | 2019/10/14 22:00 | MD 文件 | |

5. 修改hadoop配置文件

- hadoop-3.0.0/etc/hadoop/core-site.xml配置

```
<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

- hadoop-3.0.0/etc/hadoop/mapred-site.xml配置

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

- hadoop-3.0.0目录下创建tmp目录，hadoop-3.0.0/tmp下再创建datanode目录和namenode目录
- hadoop-3.0.0/etc/hadoop/hdfs-site.xml配置

```

<configuration>
  <!-- 这个参数设置为1, 因为是单机版hadoop -->
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.permissions</name>
    <value>false</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/D:/hadoop_installs/hadoop-3.0.0/tmp/namenode</value>
  </property>
  <property>
    <name>fs.checkpoint.dir</name>
    <value>/D:/hadoop_installs/hadoop-3.0.0/tmp/snn</value>
  </property>
  <property>
    <name>fs.checkpoint.edits.dir</name>
    <value>/D:/hadoop_installs/hadoop-3.0.0/tmp/snn</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/D:/hadoop_installs/hadoop-3.0.0/tmp/datanode</value>
  </property>
</configuration>

```

◦ hadoop-3.0.0/etc/hadoop/yarn-site.xml配置

```

<configuration>
  <!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
</configuration>

```

6. hadoop-3.0.0/etc/hadoop/hadoop-env.cmd配置

set JAVA_HOME=%JAVA_HOME%"替换为"set JAVA_HOME=set
 JAVA_HOME=C:\PROGRA~1\Java\jdk1.8.0_211", 路径不能有空格否则会报错,program files用
 PROGRA~1替换

7. 启动服务

```
bin\hdfs namenode-format
sbin\start-all.cmd
```

Apache Hadoop Distribution - hadoop namenode

5a451c3c0, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-5842b66c-d9fc-41c8-8d7e-1c58039989c8;nsid=445785223;c=1571071625229) storage f4af76ff-d9c4-4696-8d79-0dd5a451c3c0 2019-10-15 00:57:25,854 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866 2019-10-15 00:57:25,856 INFO blockmanagement.BlockReportLeaseManager: Registered DN f4af76ff-d9c4-4696-8d79-0dd5a451c3c0 (127.0.0.1:9866). 2019-10-15 00:57:26,024 INFO blockmanagement.DataNodeDescriptor: Adding new storage ID DS-74c435f7-d3f7-4ba5-8631-d241c6ef4848 for DN 127.0.0.1:9866 2019-10-15 00:57:26,101 INFO BlockStateChange: BLOCK* processReport 0x2a3dfec8b48c3dlc: Processing first storage report for DS-74c435f7-d3f7-4ba5-8631-d241c6ef4848 from datanode f4af76ff-d9c4-4696-8d79-0dd5a451c3c0 2019-10-15 00:57:26,104 INFO BlockStateChange: BLOCK* processReport 0x2a3dfec8b48c3dlc: from storage DS-74c435f7-d3f7-4ba5-8631-d241c6ef4848 node DatanodeRegistration(127.0.0.1:9866, datanodeUid=f4af76ff-d9c4-4696-8d79-0dd5a451c3c0, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-5842b66c-d9fc-41c8-8d7e-1c58039989c8;nsid=445785223;c=1571071625229), blocks: 0, hasStaleStorage: false, processing time: 3 msec, invalidatedBlocks: 0

Apache Hadoop Distribution - hadoop datanode

2019-10-15 00:57:26,158 INFO datanode.DataNode: Successfully sent block report 0x2a3dfec8b48c3dlc, containing 1 storage report(s), of which we sent 1. The reports had 0 total blocks and used 1 RPC(s). This took 7 msec to generate, and 81 msec for RPC and NN processing. Got back one command: FinalizeCommand /5. 2019-10-15 00:57:26,159 INFO datanode.DataNode: Got finalize command for block pool BP-1838055007-192.168.164.1-1571071625229

Apache Hadoop Distribution - yarn nodemanager

2019-10-15 00:57:28,631 INFO webapp.WebApps: Web app node started at 8042 2019-10-15 00:57:28,635 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : windows10.microdone.cn:63617 2019-10-15 00:57:28,640 INFO util.JvmPauseMonitor: Starting JVM pause monitor 2019-10-15 00:57:28,654 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8031 2019-10-15 00:57:28,761 INFO nodemanager.NodeStatusUpdaterImpl: Sending out 0 NM container statuses: [] 2019-10-15 00:57:28,776 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM using containers: [] 2019-10-15 00:57:29,317 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key with id 790453404 2019-10-15 00:57:29,319 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with id -1424443187 2019-10-15 00:57:29,322 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as windows10.microdone.cn:63617 with total resource of <memory:8192, vCores:8>

Apache Hadoop Distribution - yarn resourcemanager

a.util.concurrent.LinkedBlockingQueue queueCapacity: 100 scheduler: class org.apache.hadoop.ipc.DefaultRpcScheduler 2019-10-15 00:57:26,032 INFO ipc.Server: Starting Socket Reader #1 for port 8033 2019-10-15 00:57:26,039 INFO pb.RpcServerFactoryPBImpl: Adding protocol org.apache.hadoop.yarn.server.api.ResourceManagerAdministrationProtocolPB to the server 2019-10-15 00:57:26,042 INFO ipc.Server: IPC Server listener on 8033: starting 2019-10-15 00:57:26,044 INFO ipc.Server: IPC Server Responder: starting 2019-10-15 00:57:29,274 INFO resourcemanager.ResourceTrackerService: NodeManager from node windows10.microdone.cn(cmlPort: 63617 httpPort: 8042) registered with capability: <memory:8192, vCores:8>, assigned nodeId windows10.microdone.cn:63617 2019-10-15 00:57:29,280 INFO rmnode.RMNodeImpl: windows10.microdone.cn:63617 Node Transitioned from NEW to RUNNING 2019-10-15 00:57:29,299 INFO capacity.CapacityScheduler: Added node windows10.microdone.cn:63617 clusterResource: <memory:8192, vCores:8>

Namenode information

All Applications

localhost:9870/dfshealth.html#tab-overview

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview 'localhost:9000' (active)

| | |
|----------------|--|
| Started: | Tue Oct 15 00:57:22 +0800 2019 |
| Version: | 3.0.0, rc25427ceca461ee979d30edd7a4b0f50718e6533 |
| Compiled: | Sat Dec 09 03:16:00 +0800 2017 by andrew from branch-3.0.0 |
| Cluster ID: | CID-5842b66c-d9fc-41c8-8d7e-1c58039989c8 |
| Block Pool ID: | BP-1838055007-192.168.164.1-1571071625229 |

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks = 1 total filesystem object(s).

Heap Memory used 41.01 MB of 282 MB Heap Memory. Max Heap Memory is 889 MB.

The screenshot displays the Hadoop All Applications web interface. The top navigation bar includes 'Namenode information' and 'All Applications'. The main content area is titled 'All Applications' and features a sidebar with a 'Cluster' menu. The cluster metrics section shows a table with columns for various metrics, all of which are currently at zero. The cluster nodes metrics section shows a table with columns for node states, also all at zero. The scheduler metrics section shows a table with columns for scheduler types, all at zero. The main table is empty, displaying 'No data available in table'.

8. 运行grep示例

- 操作hdfs和linux下一样，不再赘述，似乎要用绝对路径...
- grep卡在map 0% reduce 100%，报错内容：path is not a file，解决方案：hdfs的input改成只放进去文件/*.*xml而不是文件夹，成功运行

```

at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invoke(RetryInvocationHandler.java:157)
at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invokeOnce(RetryInvocationHandler.java:157)
at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:359)
at com.sun.proxy.$Proxy18.getBlockLocations(Unknown Source)
at org.apache.hadoop.hdfs.DFSClient.callGetBlockLocations(DFSClient.java:858)
... 17 more

2019-10-15 01:21:35,428 INFO mapreduce.Job: map 100% reduce 100%
2019-10-15 01:21:35,448 INFO mapreduce.Job: Job job_1571072241532_0002 failed with state FAILED due to: Task failed as tasks failed. failedMaps:1 failedReduces:0 killedMaps:0 killedReduces: 0

2019-10-15 01:21:35,663 INFO mapreduce.Job: Counters: 9
Job Counters
  Failed map tasks=4
  Killed reduce tasks=1
  Launched map tasks=4
  Other local map tasks=4
  Total time spent by all maps in occupied slots (ms)=12951
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=12951

```

```

at com.sun.proxy.$Proxy18.getBlockLocations(Unknown Source)
at org.apache.hadoop.hdfs.DFSClient.callGetBlockLocations(DFSClient.java:858)
... 17 more

2019-10-15 01:21:28,371 INFO mapreduce.Job: task Id : attempt_1571072241532_0002_m_000000_2, Status : FAILED
Error: java.io.FileNotFoundException: Path is not a file: /user/lenovo/input/hadoop
at org.apache.hadoop.hdfs.server.namenode.INodeFile.valueOf(INodeFile.java:89)
at org.apache.hadoop.hdfs.server.namenode.INodeFile.valueOf(INodeFile.java:75)
at org.apache.hadoop.hdfs.server.namenode.FSDirStatAndListingOp.getBlockLocations(FSDirStatAndListingOp.java:152)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getBlockLocations(FSNamesystem.java:1849)
at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.getBlockLocations(NameNodeRpcServer.java:727)
at org.apache.hadoop.hdfs.protocolPB.ClientNameNodeProtocolServerSideTranslatorPB.getBlockLocations(ClientNameNodeProtocolServerSideTranslatorPB.java:414)

```

修改后

Browse Directory

/user/lenovo/input

Show 25 entries Search:

| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|------------|--------|------------|---------|---------------|-------------|------------|------------------------|
| -rw-r--r-- | lenovo | supergroup | 7.68 KB | Oct 15 01:25 | 1 | 128 MB | capacity-scheduler.xml |
| -rw-r--r-- | lenovo | supergroup | 881 B | Oct 15 01:25 | 1 | 128 MB | core-site.xml |
| -rw-r--r-- | lenovo | supergroup | 9.97 KB | Oct 15 01:25 | 1 | 128 MB | hadoop-policy.xml |
| -rw-r--r-- | lenovo | supergroup | 1.5 KB | Oct 15 01:25 | 1 | 128 MB | hdfs-site.xml |
| -rw-r--r-- | lenovo | supergroup | 620 B | Oct 15 01:25 | 1 | 128 MB | https-site.xml |
| -rw-r--r-- | lenovo | supergroup | 3.44 KB | Oct 15 01:25 | 1 | 128 MB | kms-acls.xml |
| -rw-r--r-- | lenovo | supergroup | 682 B | Oct 15 01:25 | 1 | 128 MB | kms-site.xml |
| -rw-r--r-- | lenovo | supergroup | 868 B | Oct 15 01:25 | 1 | 128 MB | mapred-site.xml |
| -rw-r--r-- | lenovo | supergroup | 963 B | Oct 15 01:25 | 1 | 128 MB | yarn-site.xml |

Showing 1 to 9 of 9 entries

Previous 1 Next

```
-publisher.enabled
2019-10-15 01:26:46,659 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1571072241532_0004
2019-10-15 01:26:46,661 INFO mapreduce.JobSubmitter: Executing with tokens: []
2019-10-15 01:26:46,840 INFO conf.Configuration: resource-types.xml not found
2019-10-15 01:26:46,841 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2019-10-15 01:26:46,917 INFO impl.YarnClientImpl: Submitted application application_1571072241532_0004
2019-10-15 01:26:46,958 INFO mapreduce.Job: The url to track the job: http://LAPTOP-S490M1TL:8088/proxy/application_1571072241532_0004/
2019-10-15 01:26:46,960 INFO mapreduce.Job: Running job: job_1571072241532_0004
2019-10-15 01:26:55,056 INFO mapreduce.Job: Job job_1571072241532_0004 running in uber mode : false
2019-10-15 01:26:55,059 INFO mapreduce.Job: map 0% reduce 0%
2019-10-15 01:27:05,382 INFO mapreduce.Job: map 67% reduce 0%
2019-10-15 01:27:12,477 INFO mapreduce.Job: map 100% reduce 0%
2019-10-15 01:27:14,498 INFO mapreduce.Job: map 100% reduce 100%
2019-10-15 01:27:14,516 INFO mapreduce.Job: Job job_1571072241532_0004 completed successfully
2019-10-15 01:27:14,646 INFO mapreduce.Job: Counters: 54
File System Counters
  FILE: Number of bytes read=6
  FILE: Number of bytes written=2067675
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
```

IntelliJ IDEA开发Hadoop伪分布式应用

坑：先关闭命令行的java进程再打开IDEA否则IDEA可能会打不开

1. 新建IDEA的Maven项目

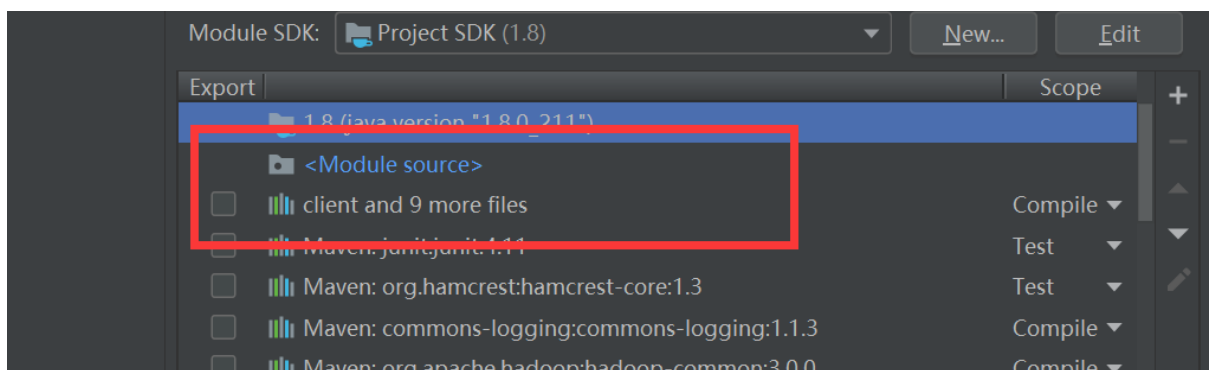
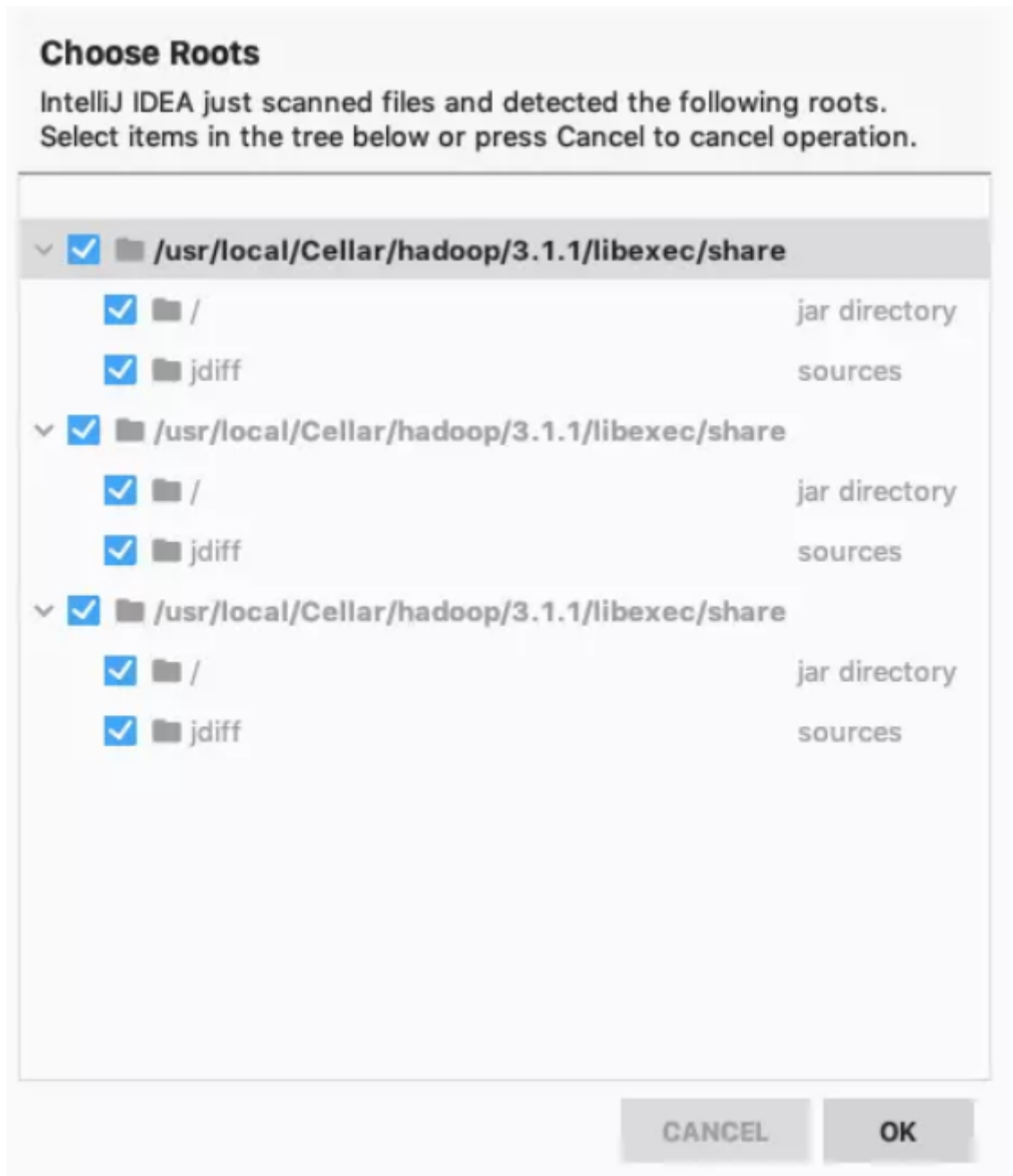
- File-New-Project-Maven, project SDK 1.8 再点击next
- 填写groupid, artifactid, next, finish

2. 修改 Target bytecode version

- 打开 Setting, 选中 Build, Execution, Deployment -> Compiler -> java, 将 Target bytecode version 改为 1.8 或 8
- 确认一下project stucture中的Project SDK, Modules SDK都是1.8

3. 导入依赖jar包

- 在Project Stucture中Modules选中 Dependencies 后点击下方的 + 号, 选择「JARs or directories」, 把hadoop目录下的share/hadoop/下的几个文件夹里的jar包都添进去



4. 在pom.xml中添加依赖 我的是这样写的

```
<?xml version="1.0" encoding="UTF-8"?>
<project xmlns="http://maven.apache.org/POM/4.0.0"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://maven.apache.org/POM/4.0.0
http://maven.apache.org/xsd/maven-4.0.0.xsd">
```



```

<modelVersion>4.0.0</modelVersion>

<groupId>hts.nju.edu.cn</groupId>
<artifactId>testWordCount</artifactId>
<version>1.0-SNAPSHOT</version>

<dependencies>
  <!-- https://mvnrepository.com/artifact/junit/junit -->
  <dependency>
    <groupId>junit</groupId>
    <artifactId>junit</artifactId>
    <version>4.11</version>
    <scope>test</scope>
  </dependency>
  <!--&lt;!&dash; https://mvnrepository.com/artifact/commons-logging/commons-logging &dash;&gt;-->
  <dependency>
    <groupId>commons-logging</groupId>
    <artifactId>commons-logging</artifactId>
    <version>1.1.3</version>
  </dependency>
  <!--&lt;!&dash; https://mvnrepository.com/artifact/org.apache.hadoop/hadoop-common &dash;&gt;-->
  <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-common</artifactId>
    <version>3.0.0</version>
  </dependency>
  <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-hdfs</artifactId>
    <version>3.0.0</version>
  </dependency>
  <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-client</artifactId>
    <version>3.0.0</version>
  </dependency>
</dependencies>
</project>

```

- 添加后刷新maven下载依赖，注意有的下载错误的，不方，可以去本地maven仓库删除掉再重新下载或许就好了。

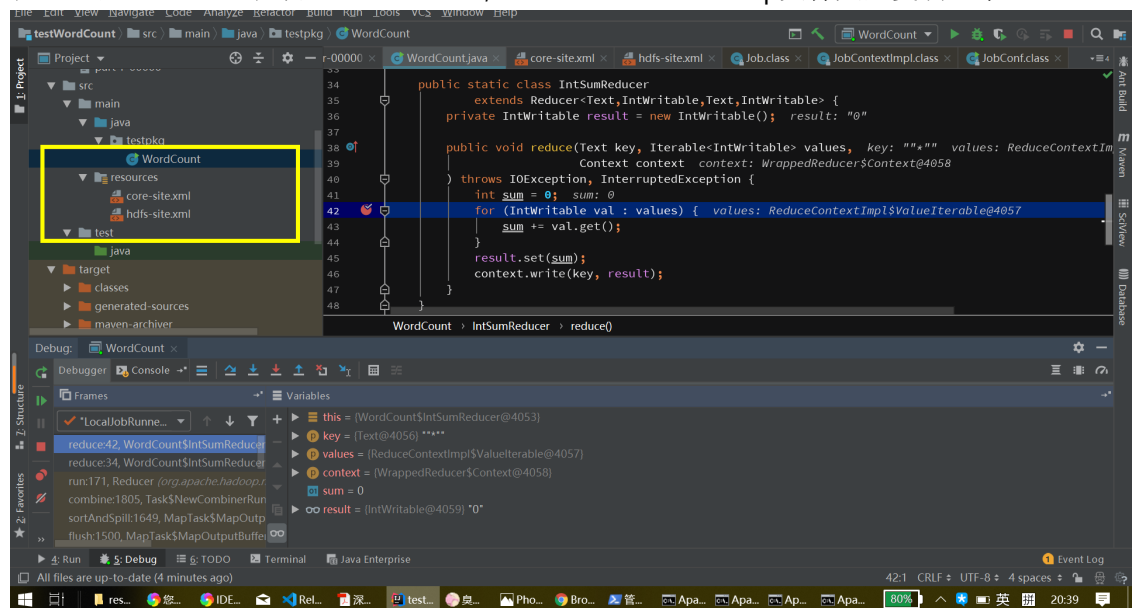
5. 编写hadoop java代码WordCount

- 单机模式
 - 和src同级建立一个input文件夹，里面随便建立两个多行txt文件作为输入
 - 在IntelliJ菜单栏中选择Run->Edit Configurations，在弹出来的对话框中点击+，新建一个Application配置。配置Main class为WordCount（可以点击右边的...选择），Program

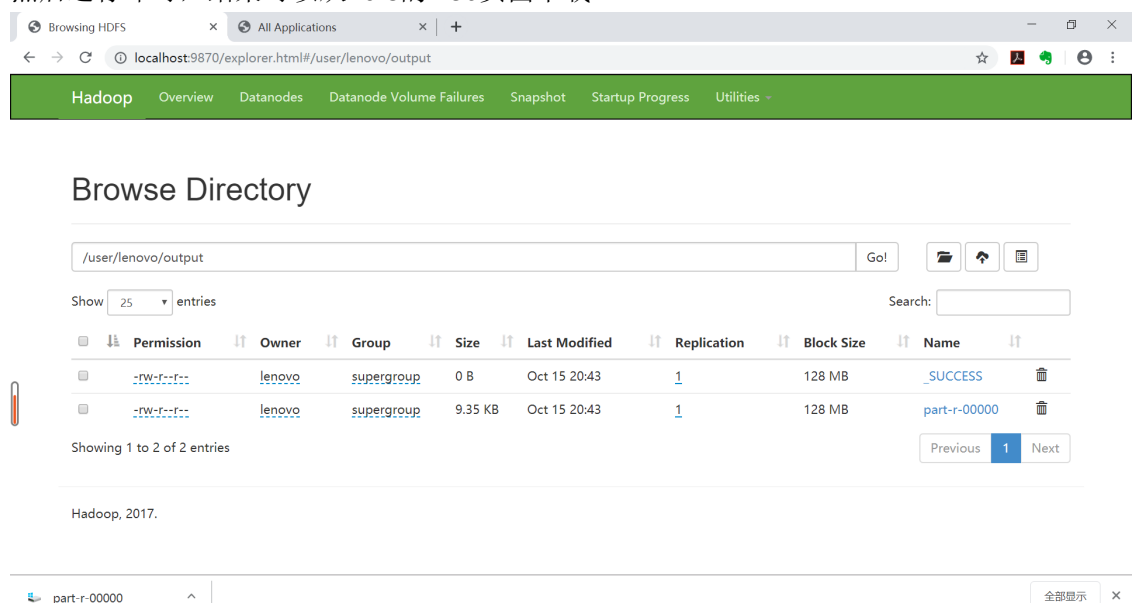
arguments为input/ output/, 即输入路径为刚才创建的input文件夹, 输出为output, 点击ok, 然后build, run就可以了

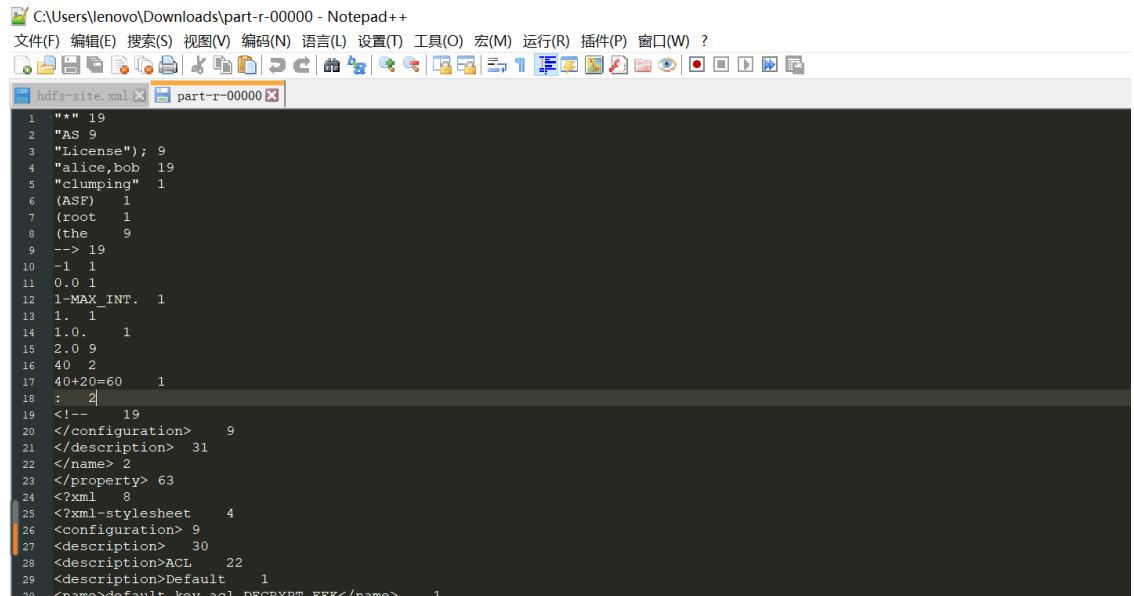
- 伪分布式

- 命令行start-all.cmd启动hdfs
- 在IDEA里resources中添加core-site.xml, hdfs-site.xml (hadoop文件夹里复制过来)



- 然后运行即可, 结果可以从hdfs的web页面下载





- 注意由于hadoop设定，下次运行前务必删除output文件夹

Intellij IDEA配置HDFS可视化插件

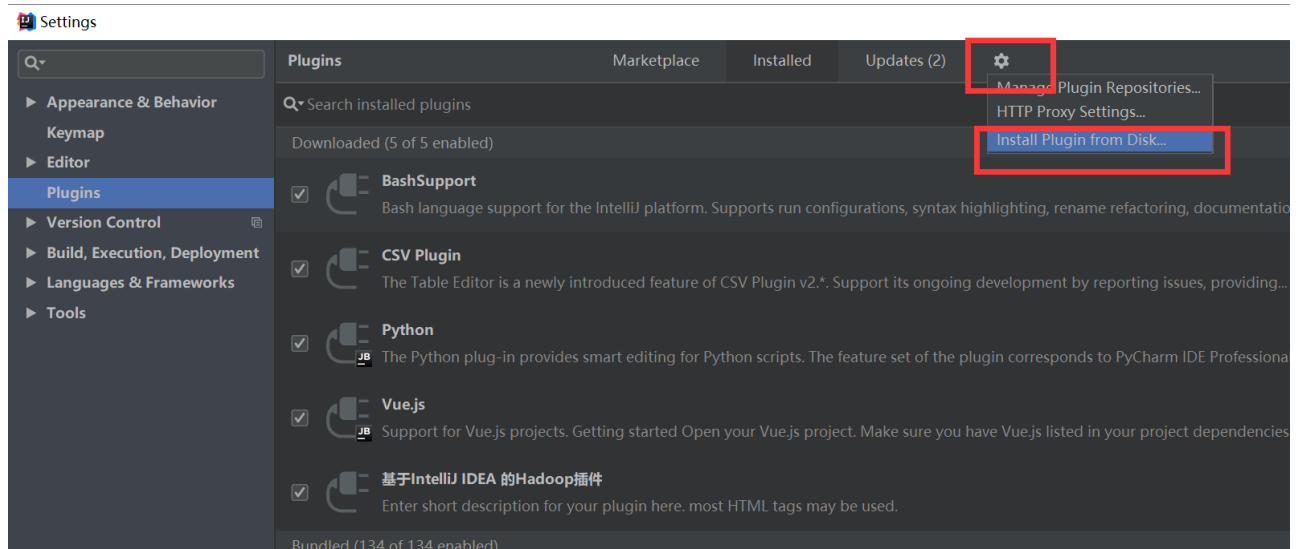
1. 下载HadoopIntellijPlugin

- github地址: <https://github.com/fangyuzhong2016/HadoopIntellijPlugin> **Github**上的源码先编译再用 网上也有别人编译好的zip但是要收费很不爽
- 编译: 如果是要开发则需要新建IDEA Plugin项目, 但是编译只要将github导入maven中, 修改pom.xml (主要是hadoop版本改成自己的, IDEA安装目录改成自己的, 就可以了。然后

```
mvn clean
mvn assembly:assembly
```

第一次出错, 重新clean一下好了, 找到target下的.zip文件为编译好的插件安装包, 无需解压。

2. 安装HadoopIntellijPlugin File-Settings-Editor-Plugins

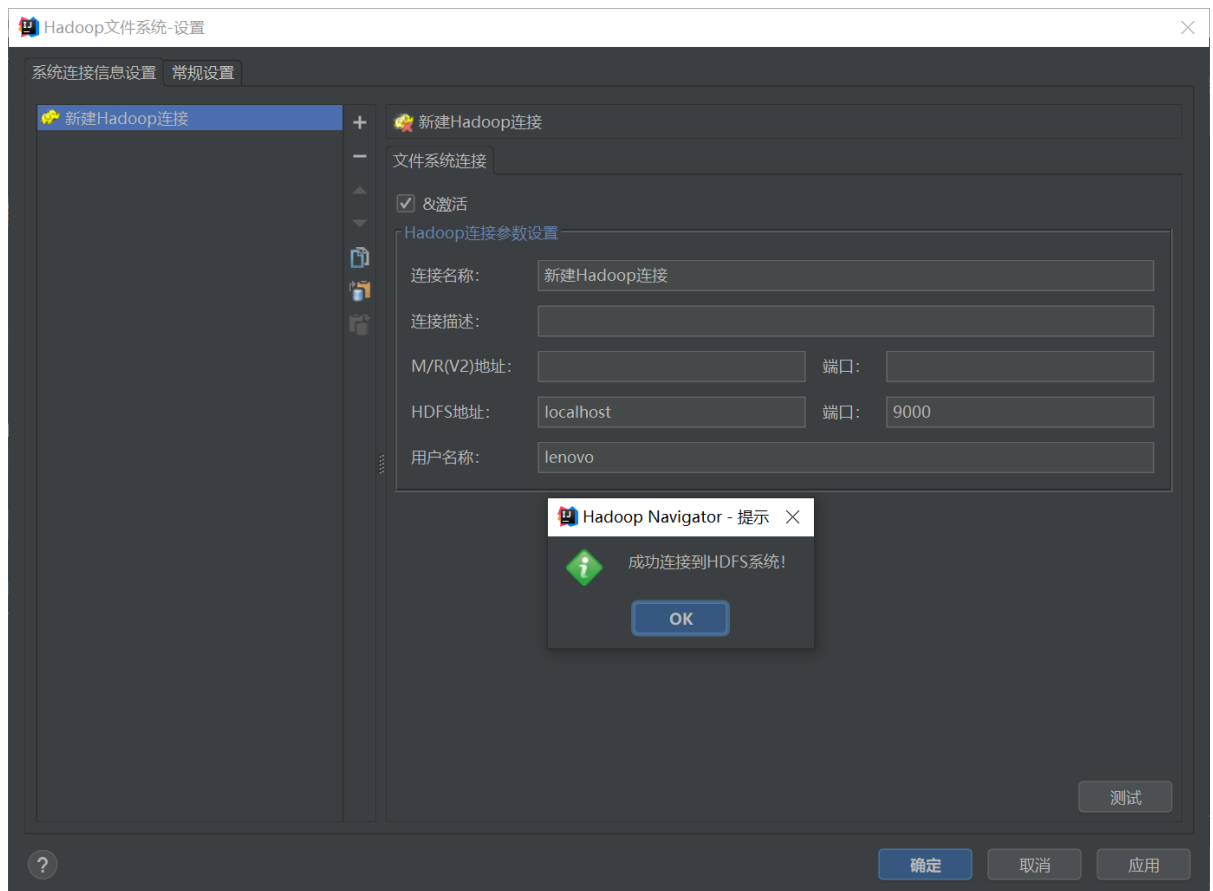


3. 按提示要重启IDEA

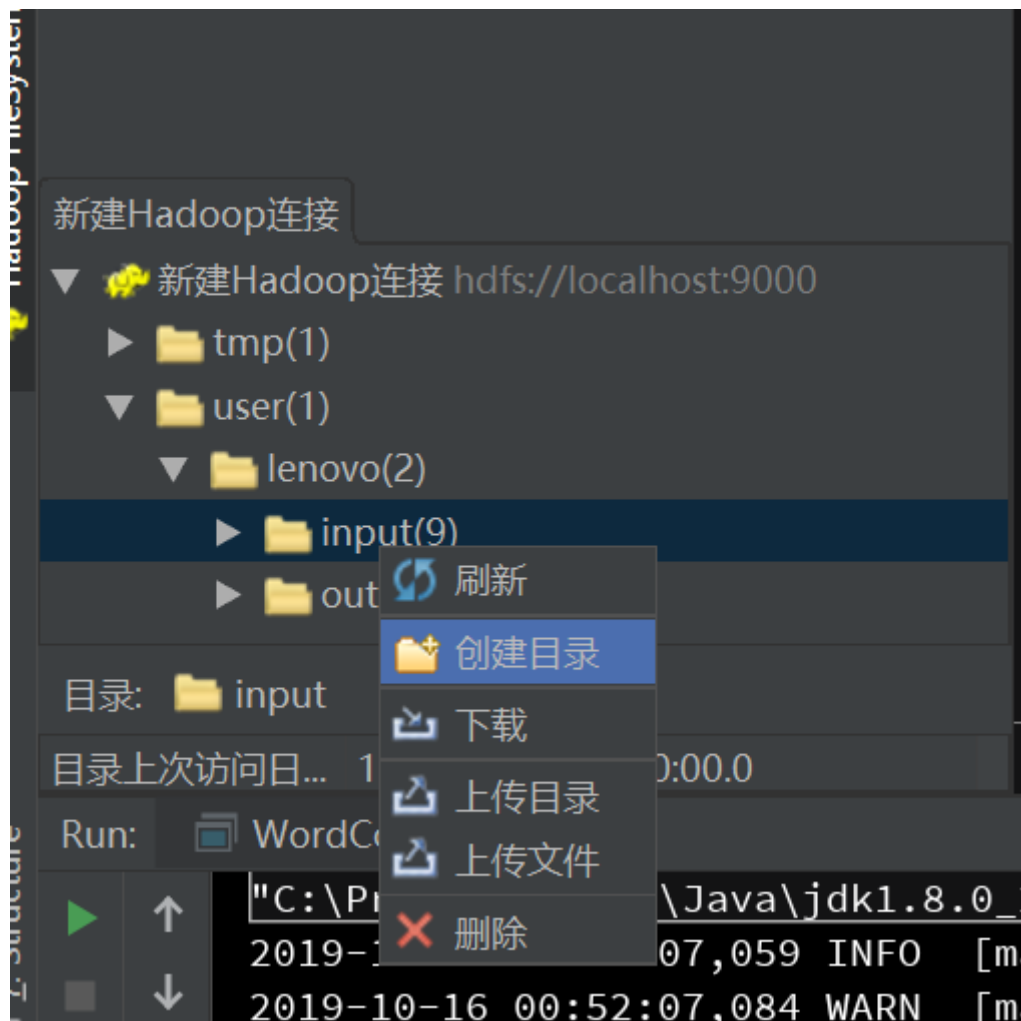
- （不要忘了关命令行java否则又启动不起来了）

4. 使用HadoopIntelliJPlugin

- 两个地方进入
- 新建连接，点击左栏边的 + 号，填写正确的HDFS地址和端口



- 成功口可以看到界面



这样码代码的环境就比较宜人了哈哈哈哈哈