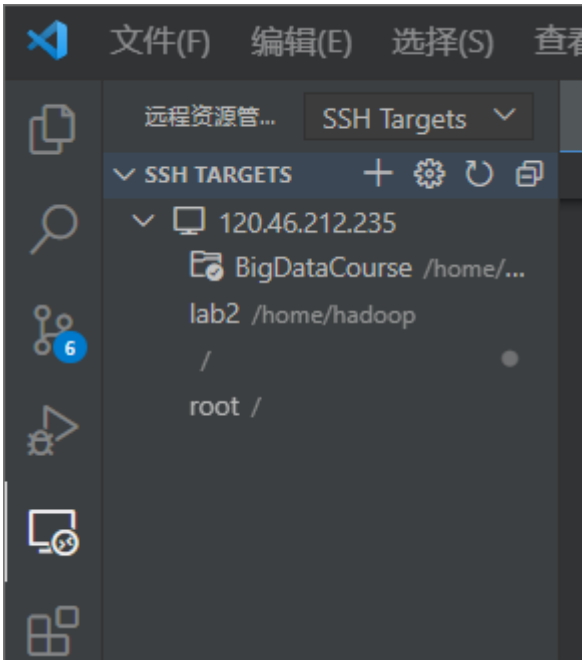
1 开发环境配置

1.1 VS Code配置远程开发环境

如下图所示，直接在左端创建SSH远程连接即可。



1.2 Git配置

1. 安装Git
   - 安装依赖库

     ```
     yum install curl-devel expat-devel gettext-devel openssl-devel zlib-devel
     ```

     ```
     yum install gcc-c++ perl-ExtUtils-MakeMaker
     ```

   - 下载源码包

     ```
     cd /usr/src
     wget http://mirrors.edge.kernel.org/pub/software/scm/git/git-2.36.1.tar.gz
     ```

   - 解压、编译、安装

     ```
     tar -zxvf git-2.36.1.tar.gz
     cd git-2.36.1
     ./configure --prefix=/usr/local
     make
     make install
     ```

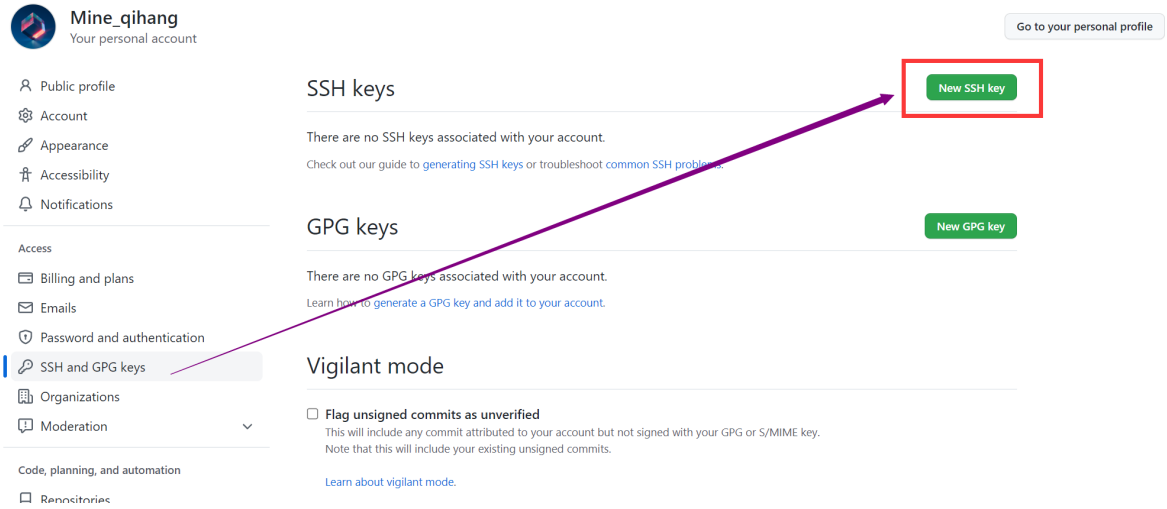
   - 刷新环境变量

     ```
     source /etc/profile
     ```

- 查看git版本

```
git --version
```

2. 运行`git config --global user.name "自己github的用户名"`设置用户名
3. 运行`git config --global user.email "自己github的邮箱"`设置邮箱
4. 在github中的个人设置界面新建SSH Keys并将`/home/hadoop/.ssh/id_rsa.pub`中内容粘入，保存即可。



5. 运行`git clone git@github.com:MineQihang/BigDataCourse.git`拉取仓库即可（此为私人仓库，需要先申请成为开发者）
6. 进入`lab2`文件夹再进行下一步

1.3 Python配置（也可使用Anaconda配虚拟环境）

1. 更新pip

```
sudo python3 -m pip install -i https://pypi.tuna.tsinghua.edu.cn/simple --upgrade pip
```

2. 使用清华镜像

```
sudo pip3 config set global.index-url https://pypi.tuna.tsinghua.edu.cn/simple
```

3. 安装相关库

```
sudo pip3 install -r requirements.txt
```

1.4 安装字体文件

```
sudo yum install wqy-microhei-fonts
```

1.5 安装驱动文件

1. 安装google-chrome

```
sudo yum -y install google-chrome-stable
```

> 如果找不到google-chrome-stable，请按照下方指令执行：
>
> 在目录 /etc/yum.repos.d/ 下新建文件 google-chrome.repo:
>
> ```
> sudo vim /etc/yum.repos.d/google-chrome.repo
> ```
>
> 写入如下内容:
>
> ```
> [google-chrome]
> name=google-chrome
> baseurl=http://dl.google.com/linux/chrome/rpm/stable/$basearch
> enabled=1
> gpgcheck=1
> gpgkey=https://dl-ssl.google.com/linux/linux_signing_key.pub
> ```

2. 安装chromedrive

   chromedrive版本要和google-chrome对应，所以我们先查看google-chrome版本号：

```
google-chrome-stable --version
```

   记录版本号后，去https://npm.taobao.org/mirrors/chromedriver/ 下载对应的驱动。**请修改下方的版本号**（如果没有，那么选择离本版本最近的即可）

```
sudo rm -f chromedriver_linux64.zip*
wget https://npm.taobao.org/mirrors/chromedriver/[版本号]/chromedriver_linux64.zip
```

   - 先删除之前的chromedriver

```
sudo rm -f /usr/bin/chromedriver
sudo rm -f /usr/local/bin/chromedriver
sudo rm -f /usr/local/share/chromedriver
```

   - unzip 解压

   解压出文件chromedriver

```
unzip chromedriver_linux64.zip
```

   - 赋予777权限

```
chmod 777 chromedriver
```

   - mv 移动到/usr/bin/路径

```
sudo mv chromedriver /usr/bin/
```

1.6 设置日志级别

由于`spark`在运行时会打印非常多的日志，为了便于调试观察，我们设置日志级别为`WARN`。

以下为全局设置日志级别方式，你也可在代码中临时设置`sc.setLogLevel("WARN")`（详见`ex3.pdf`）。

1. 切换到`conf`目录

```
cd /usr/local/spark/conf
```

2. 设置配置文件

```
cp log4j.properties.template log4j.properties
vim log4j.properties
```

修改`log4j.rootCategory=WARN,console`

2 编码WordCount.py

2.2 提交到Spark运行

```
/usr/local/spark/bin/spark-submit /home/hadoop/BigDataCourse/lab2/WordCount.py
```

```
[hadoop@ecs-bigdata lab2]$ /usr/local/spark/bin/spark-submit /home/hadoop/BigDataCourse/lab2/WordCount.py
22/10/16 10:51:36 WARN Utils: Your hostname, ecs-bigdata resolves to a loopback address: 127.0.0.1; using 192
.168.0.59 instead (on interface eth0)
22/10/16 10:51:36 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
22/10/16 10:51:36 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using buil
tin-java classes where applicable
Building prefix dict from the default dictionary ...
Dumping model to file cache /tmp/jieba.cache
Loading model cost 0.970 seconds.
Prefix dict has been built successfully.
22/10/16 10:51:55 WARN TaskSetManager: Stage 1 contains a task of very large size (10038 KiB). The maximum re
commended task size is 1000 KiB.
[('身高', 2627), ('家庭', 2018), ('父母', 2002), ('性格', 1882), ('男生', 1640), ('朋友', 1618), ('条件', 156
8), ('学历', 1445), ('女生', 1380), ('感情', 1301)]
```

遇到的问题

```
[hadoop@slave01 lib]$ /usr/local/spark/bin/spark-submit /home/hadoop/BigDataCourse/lab2/WordCount.py
log4j:WARN No appenders could be found for logger (org.apache.spark.util.ShutdownHookManager).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Traceback (most recent call last):
  File "/home/hadoop/BigDataCourse/lab2/WordCount.py", line 84, in <module>
    sorted(rdd.reduceByKey(lambda x, y : x + y).collect())
  File "/usr/local/spark/python/lib/pyspark.zip/pyspark/rdd.py", line 1197, in collect
  File "/usr/local/spark/python/lib/py4j-0.10.9.5-src.zip/py4j/java_gateway.py", line 1322, in __call__
  File "/usr/local/spark/python/lib/py4j-0.10.9.5-src.zip/py4j/protocol.py", line 328, in get_return_value
py4j.protocol.Py4JJavaError: An error occurred while calling z:org.apache.spark.api.python.PythonRDD.collectAndServe.
: org.apache.spark.SparkException: Job aborted due to stage failure: Task 1 in stage 0.0 failed 1 times, most recent failure: Lost task 1.0 in stage 0.0 (TID 1) (slave01 executor driver): org.apache.spark.api.python.PythonException: Traceback (most recent call last):
  File "/usr/local/spark/python/lib/pyspark.zip/pyspark/worker.py", line 668, in main
    func, profiler, deserializer, serializer = read_command(pickleSer, infile)
  File "/usr/local/spark/python/lib/pyspark.zip/pyspark/worker.py", line 85, in read_command
    command = serializer._read_with_length(file)
  File "/usr/local/spark/python/lib/pyspark.zip/pyspark/serializers.py", line 173, in _read_with_length
    return self.loads(obj)
  File "/usr/local/spark/python/lib/pyspark.zip/pyspark/serializers.py", line 452, in loads
    return pickle.loads(obj, encoding=encoding)
  File "/usr/local/spark/python/lib/pyspark.zip/pyspark/cloudpickle/cloudpickle.py", line 590, in _create_parametrized_type_hint
    return origin[args]
  File "/usr/lib64/python3.6/typing.py", line 682, in inner
    return func(*args, **kwds)
  File "/usr/lib64/python3.6/typing.py", line 1131, in __getitem__
    _check_generic(self, params)
  File "/usr/lib64/python3.6/typing.py", line 662, in _check_generic
    ("many" if alen > elen else "few", repr(cls), alen, elen))
TypeError: Too many parameters for typing.Iterable; actual 2, expected 1
```

版本问题

换成Spark3.1.2

```
22/10/16 11:22:44 WARN TaskSetManager: Stage 1 contains a task of very large size (10038 KiB). The maximum recommended task size is 1000 KiB.
Traceback (most recent call last):
  File "/home/hadoop/BigDataCourse/lab2/WordCount.py", line 75, in <module>
    resRdd = wordcount(isvisualize=True)
  File "/home/hadoop/BigDataCourse/lab2/WordCount.py", line 67, in wordcount
    v.drawPie(pieDic)
  File "/home/hadoop/BigDataCourse/lab2/visualize.py", line 91, in drawPie
    make_snapshot(snapshot, pie_position().render(), SAVAPATH + '饼图可视化.png')
  File "/usr/local/lib/python3.6/site-packages/pyecharts/render/snapshot.py", line 52, in make_snapshot
    save_as_png(image_data, output_name)
  File "/usr/local/lib/python3.6/site-packages/pyecharts/render/snapshot.py", line 77, in save_as_png
    with open(output_name, "wb") as f:
PermissionError: [Errno 13] Permission denied: '/home/hadoop/BigDataCourse/lab2/results/饼图可视化.png'
```

```
sudo chmod 777 BigDataCourse -R
```

设置权限即可