安完Hadoop之后再来安装spark。

1. sudo apt update

2. wget <https://dlcdn.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz>， 这个版本好像才跟Hadoop3之后的兼容

3. tar xvf spark-3.0.3-bin-hadoop2.7.tgz， 解压下载的spark文件

4. sudo mv spark-3.0.3-bin-hadoop2.7/ /opt/spark， 把spark文件夹移动到/opt/spark里面

5. sudo nano ~/.profile 打开配置文件，加入

export SPARK\_HOME=/opt/spark

export PATH=$PATH:$SPARK\_HOME/bin:$SPARK\_HOME/sbin

export PYSPARK\_PYTHON=/usr/bin/python3

退出后 source ~/.profile来是路径变量生效。

6. 进入/opt/spark/bin然后输pyspark来运行spark shell，运行成功

A computer screen with white text

Description automatically generated

-以Standalone模式在Hadoop上运行spark

进入/opt/spark/sbin然后输入start-msater.sh打开cluster，然后因为我们用到vm这一步本来应该返回master node port的但是没有，所以要自己拿。

wget <http://localhost:8080> 下载index.html 然后cat index.html | grep master来找到带有master的内容



后面spark://这一段就是master node url

然后运行/opt/spark/bin/pypark –-master spark://… （这一段替换成你拿到的对应的url）就可以用standalone的模式在master node上运行pyspark了

A computer screen with text on it

Description automatically generated

为了拿到hadoop http url, 用命令 hdfs getconf -confkey dfs.namenode.http-address

通过webhdfs来访问hdfs, 用命令 hdfs dfs -ls webhdfs://<hadoop http url>/<path>

传输back-up plan:

使用scp，现在本地生成key pair, ssh-keygen -t rsa,拷贝 ~/.ssh/id\_rsa.pub里的内容到remote主机的 ~/.ssh/authorized\_keys里面 然后使用scp命令就可以传输文件了。

hdfs://h-QF5214M:9000 represents a HDFS endpoint

dfs.replication specifies the default block replication factor for files stored in HDFS.