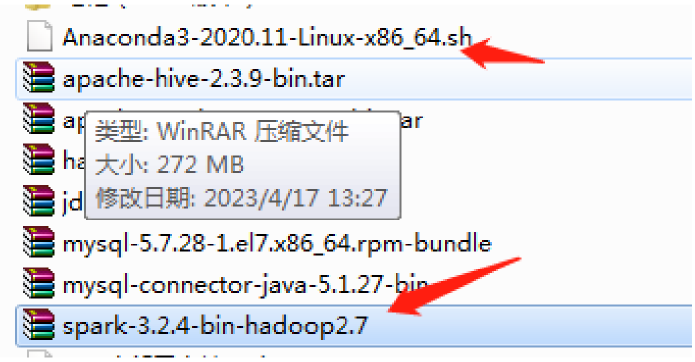
1.Anaconda&spark安装

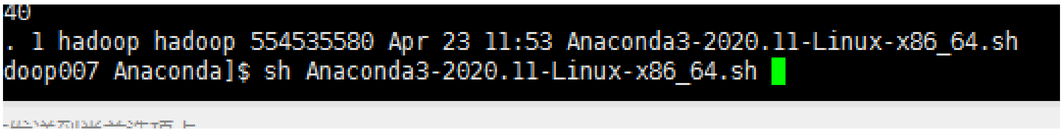
Anaconda&spark安装

1、上传Anaconda和spark安装包到服务器



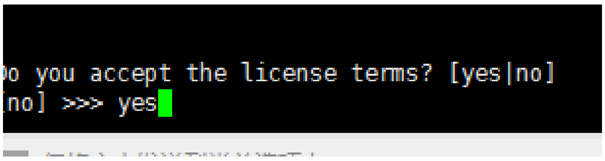
Anaconda3-2020.11-Linux-x86_64.sh
apache-hive-2.3.9-bin.tar
ap
ha
jd
类型: WinRAR 压缩文件
大小: 272 MB
修改日期: 2023/4/17 13:27
mysql-5.7.28-1.el7.x86_64.rpm-bundle
mysql-connector-java-5.1.27-bin
spark-3.2.4-bin-hadoop2.7

安装Anaconda无需解压，直接执行Anaconda脚本



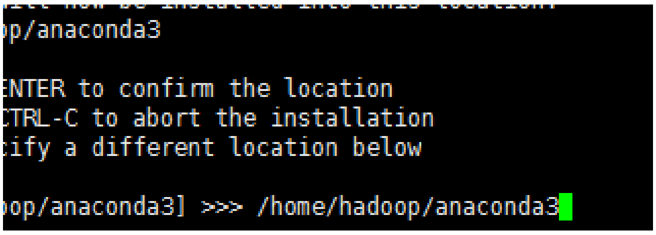
```
40
. 1 hadoop hadoop 554535580 Apr 23 11:53 Anaconda3-2020.11-Linux-x86_64.sh
doop007 Anaconda]$ sh Anaconda3-2020.11-Linux-x86_64.sh
```

授权写yes



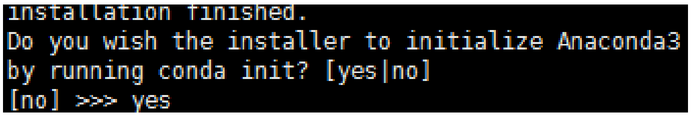
```
o you accept the license terms? [yes|no]
[no] >>> yes
```

安装目录填/home/hadoop/anaconda3



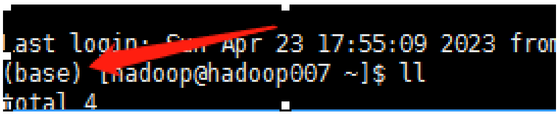
```
op/anaconda3

NTER to confirm the location
TRL-C to abort the installation
ify a different location below

op/anaconda3] >>> /home/hadoop/anaconda3
```

是否初始化选择yes



```
installation finished.
Do you wish the installer to initialize Anaconda3
by running conda init? [yes|no]
[no] >>> yes
```

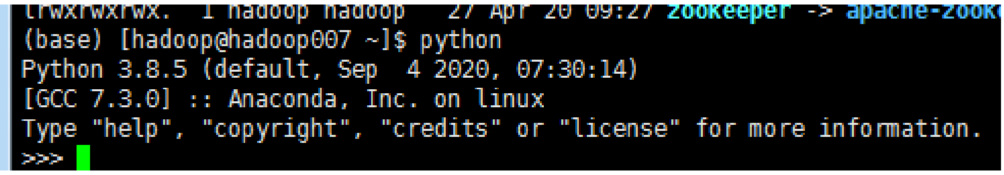
2、安装完成后重新打开一个xshell窗口，如果出现base，说明安装成功



```
ast login: Sun Apr 23 17:55:09 2023 from
(base) [hadoop@hadoop007 ~]$ ll
total 4
```

如果没有base显示，执行conda activate命令看是否能进入base环境

进入成功到虚拟环境之后检查Python环境是否正常，测试Python代码



```
lrwxrwxrwx.  1 hadoop hadoop   27 Apr 20 09:27 zookeeper -> apache-zook
(base) [hadoop@hadoop007 ~]$ python
Python 3.8.5 (default, Sep  4 2020, 07:30:14)
[GCC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> 
```

3、创建一个基于Python3.8的虚拟环境

```
conda create -n pyspark python=3.8
```

切换到pyspark环境命令：conda activate pyspark
退出pyspark环境命令，（再次退出会退出base环境）：conda deactivate
默认不进入虚拟环境（base）命令：conda config --set auto_activate_base false

其他节点同步安装Anaconda环境，并创建pyspark虚拟环境，重复上述步骤。

4、安装spark on yarn模式。

注意：

spark on yarn只需要在Yarn集群其中一个节点上安装Spark即可，该节点可作为提交Spark应用程序到YARN集群的客户端。Spark本身的Master节点和Worker节点不需要启动，由Yarn集群统一调度。

因此，只需在主节点安装spark，并配置环境变量即可，但是其他节点需要安装Python环境来运行Python代码

4.1、解压spark安装包

```
tar -zxvf spark-3.2.4-bin-hadoop2.7.tgz -C /home/hadoop/
```

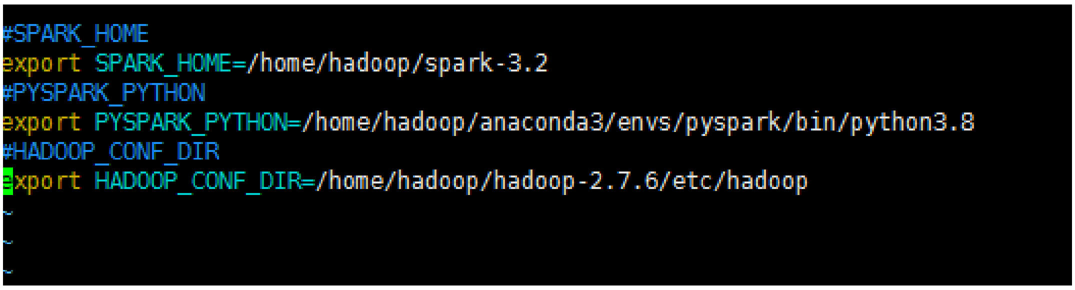
4.2、创建spark的软连接,cd到hadoop用户目录下

```
ln -s spark-3.2.4-bin-hadoop2.7 spark-3.2
```

4.3、配置环境变量到/etc/profile.d/my_env.sh,(配置完source一下：命令:source /etc/profile)

配置内容:

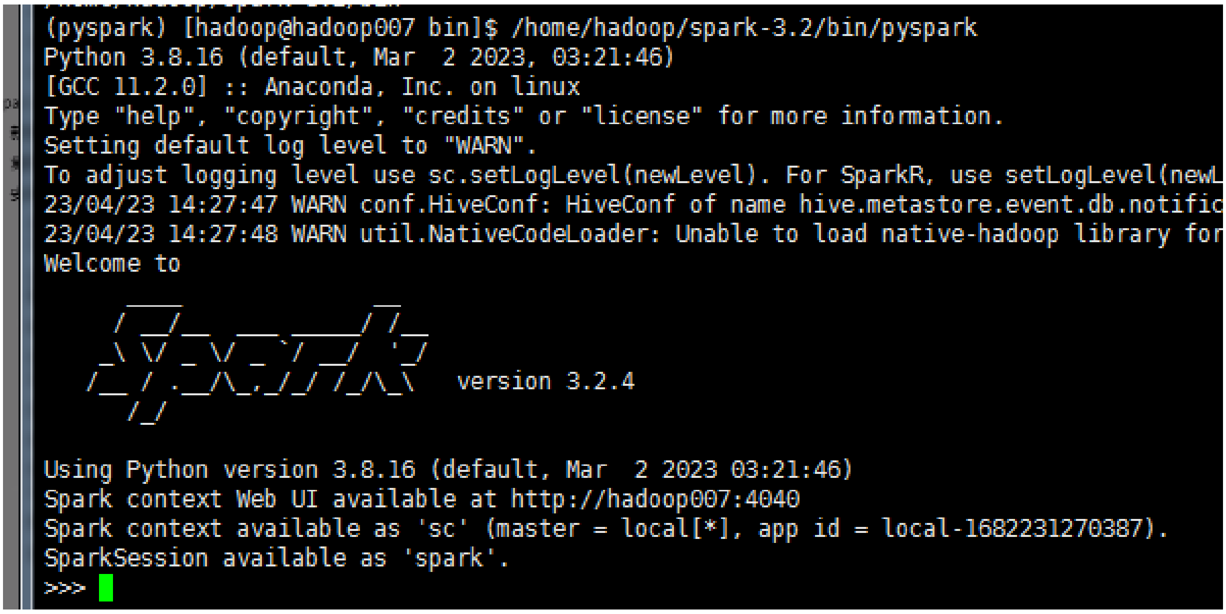
```
#SPARK_HOME
export SPARK_HOME=/home/hadoop/spark-3.2
#PYSPARK_PYTHON
export PYSPARK_PYTHON=/home/hadoop/anaconda3/envs/pyspark/bin/python3.8
#HADOOP_CONF_DIR
export HADOOP_CONF_DIR=/home/hadoop/hadoop-2.7.6/etc/hadoop
```



```
#SPARK_HOME
export SPARK_HOME=/home/hadoop/spark-3.2
#PYSPARK_PYTHON
export PYSPARK_PYTHON=/home/hadoop/anaconda3/envs/pyspark/bin/python3.8
#HADOOP_CONF_DIR
export HADOOP_CONF_DIR=/home/hadoop/hadoop-2.7.6/etc/hadoop
~
~
~
```

4.4、测试启动spark（启动pyspark一定要切换到anaconda的虚拟环境conda activate pyspark）

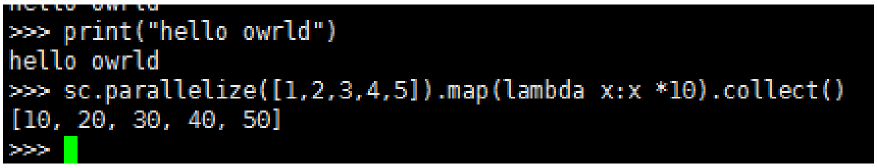
启动命令： `/home/hadoop/spark-3.2/bin/pyspark` （此时启动方式是local模式）



```
(pyspark) [hadoop@hadoop007 bin]$ /home/hadoop/spark-3.2/bin/pyspark
Python 3.8.16 (default, Mar  2 2023, 03:21:46)
[GCC 11.2.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newL
23/04/23 14:27:47 WARN conf.HiveConf: HiveConf of name hive.metastore.event.db.notific
23/04/23 14:27:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.2.4
      /_/

Using Python version 3.8.16 (default, Mar  2 2023 03:21:46)
Spark context Web UI available at http://hadoop007:4040
Spark context available as 'sc' (master = local[*], app id = local-1682231270387).
SparkSession available as 'spark'.
>>> 
```

测试是否能正常计算：

测试代码：`sc.parallelize([1,2,3,4]).map(lambda x:x * 10).collect()`



```
>>> print("hello owrld")
hello owrld
>>> sc.parallelize([1,2,3,4,5]).map(lambda x:x *10).collect()
[10, 20, 30, 40, 50]
>>> 
```

4.5、进入到spark的conf目录下。将park-env.sh.template改为spark-env.sh并添加内容。

```
mv spark-env.sh.template spark-env.sh
vim spark-env.sh
```

添加内容：

#JAVA安装目录

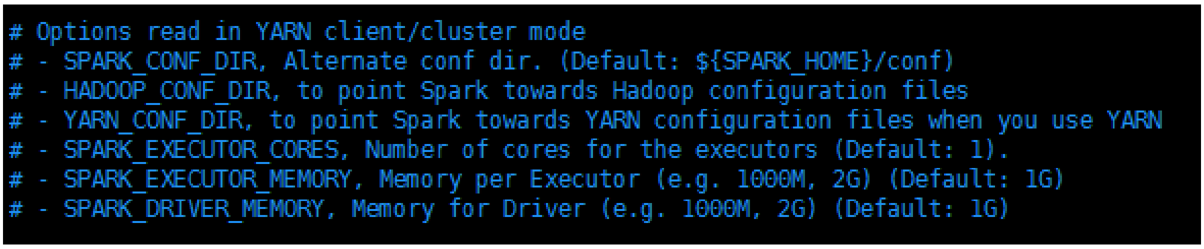
JAVA_HOME=/home/hadoop/jdk1.8.0_181

# HADOOP配置目录

HADOOP_CONF_DIR=/home/hadoop/hadoop-2.7.6/etc/hadoop

YARN_CONF_DIR=/home/hadoop/hadoop-2.7.6/etc/hadoop

4.6、官spark官方给出的yarn集群所需配置图示（了解即可）



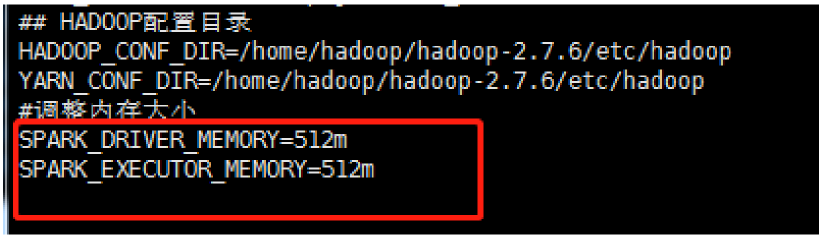
```
# Options read in YARN client/cluster mode
# - SPARK_CONF_DIR, Alternate conf dir. (Default: ${SPARK_HOME}/conf)
# - HADOOP_CONF_DIR, to point Spark towards Hadoop configuration files
# - YARN_CONF_DIR, to point Spark towards YARN configuration files when you use YARN
# - SPARK_EXECUTOR_CORES, Number of cores for the executors (Default: 1).
# - SPARK_EXECUTOR_MEMORY, Memory per Executor (e.g. 1000M, 2G) (Default: 1G)
# - SPARK_DRIVER_MEMORY, Memory for Driver (e.g. 1000M, 2G) (Default: 1G)
```

5、调整启动内存和yarn容器内存大小（可选项，电脑内存足够大的无需做，最起码16GB以上内存）

配置此项原因：因为yarn默认申请的容器大小是2.1G虚拟内存，而启动spark是需要2.3G虚拟内存，因此yarn会直接kill掉该application，当然也可以禁用yarn会kill掉application的选项，但是不推荐，会造成服务器崩溃。

一篇解决此问题的优质博客链接：https://blog.csdn.net/L_15156024189/article/details/106647535

5.1、调整启动内存和容器内存大小，在spark-env.sh文件中添加此内容：



```
## HADOOP配置目录
HADOOP_CONF_DIR=/home/hadoop/hadoop-2.7.6/etc/hadoop
YARN_CONF_DIR=/home/hadoop/hadoop-2.7.6/etc/hadoop
#调整内存大小
SPARK_DRIVER_MEMORY=512m
SPARK_EXECUTOR_MEMORY=512m
```

5.2、在yarn-site.xml中将yarn.nodemanager.vmem-pmem-ratio的值改大并分发到另外两台机器。然后重启集群。

yarn-site.xml添加以下内容：

```
<property>
    <name>yarn.nodemanager.vmem-pmem-ratio</name>
    <value>3</value>
</property>
```

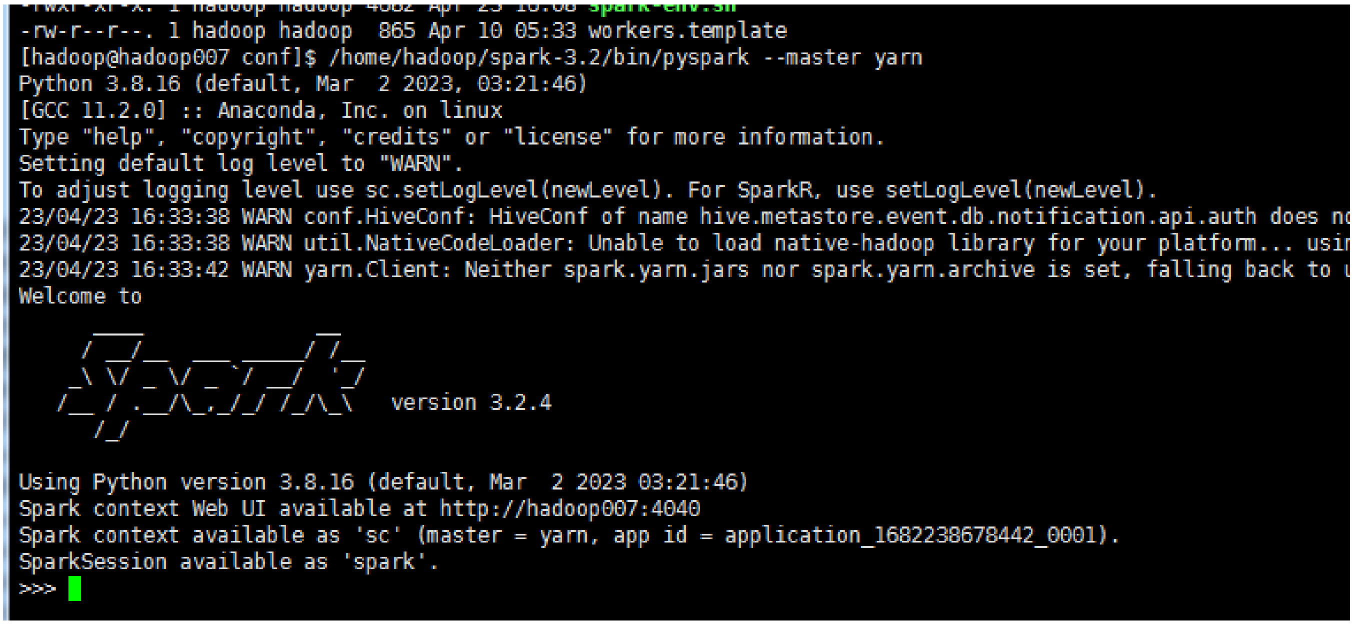
分发： （注意进入到hadoop的./etc/hadoop/目录下再使用`pwd`**）**

scp yarn-site.xml hadoop008:`pwd`/

scp yarn-site.xml hadoop009:`pwd`/

5.3、启动yarn模式

`/home/hadoop/spark-3.2/bin/pyspark --master yarn`



```
-rwxr-xr-x. 1 hadoop hadoop 4882 Apr 23 16:08 spark-env.sh
-rw-r--r--. 1 hadoop hadoop  865 Apr 10 05:33 workers.template
[hadoop@hadoop007 conf]$ /home/hadoop/spark-3.2/bin/pyspark --master yarn
Python 3.8.16 (default, Mar  2 2023, 03:21:46)
[GCC 11.2.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/04/23 16:33:38 WARN conf.HiveConf: HiveConf of name hive.metastore.event.db.notification.api.auth does no
23/04/23 16:33:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
23/04/23 16:33:42 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to u
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.2.4
      /_/

Using Python version 3.8.16 (default, Mar  2 2023 03:21:46)
Spark context Web UI available at http://hadoop007:4040
Spark context available as 'sc' (master = yarn, app id = application_1682238678442_0001).
SparkSession available as 'spark'.
>>> 
```