金融大数据实验四

211275012 郭洋溢

本次实验我选择使用pyspark和jupyter notebook来完成，以下是实验步骤

Spark配置

1. 单机模式

1.1 下载Anaconda3

下载 Anaconda3-2021.11-Linux-x86\_64.sh：

wget https://repo.anaconda.com/archive/Anaconda3-2021.11-Linux-x86\_64.sh

解压安装包

sudo tar -zxf ~/Anaconda3-2021.11-Linux-x86\_64.sh -C /home/gyy

安装Anaconda3：

|  |
| --- |
| sh ./Anaconda3-2021.11-Linux-x86\_64.sh |

接着然后一直回车然后空格，等出现接受许可界面，输入 yes接收许可。

1.2 配置文件

配置国内源：

|  |
| --- |
| sudo vim ~/.condarc |

追加以下内容：

|  |
| --- |
| channels:  - defaults  show\_channel\_urls: true  default\_channels:  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/r  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/msys2  custom\_channels:  conda-forge: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud  msys2: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud  bioconda: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud  menpo: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud  pytorch: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud  simpleitk: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud |

创建pyspark环境:

conda create -n pyspark python=3.6

conda activate pyspark

在pyspark环境中使用pip下载pyhive、 pyspark、jieba包：

1.3 Spark Local模式搭建

下载 spark-2.4.8，解压安装包spark-2.4.8-bin-without-hadoop至路径 /home/gyy：

sudo tar -zxf ~/spark-2.4.8-bin-without-hadoop -C /home/gyy

重命名：

|  |
| --- |
| mv spark-2.4.8-bin-without-hadoop/ spark-2.4.8 |

配置环境变量和相关文件

vi ~/.bashrc修改 .bashrc 文件，在文件末尾追加：

|  |
| --- |
| # 默认启动pyspark虚拟环境  conda activate pyspark  # JAVA\_HOME  export JAVA\_HOME=/home/gyy/jdk1.8.0\_381  # PYSPARK\_PYTHON  export PYSPARK\_PYTHON=/home/gyy/anaconda3/envs/pyspark/bin/python  export PATH=$JAVA\_HOME/bin:$PATH  export SPARK\_HOME=/home/gyy/spark  export PATH=$SPARK\_HOME/bin:$PATH |

配置 /etc/profile ，文件末尾追加：

export SPARK\_HOME=/home/gyy/spark-2.4.8

# HADOOP\_CONF\_DIR

export HADOOP\_CONF\_DIR=$HADOOP\_HOME/home/gyy/hadoop-3.3.6

# PYSPARK\_PYTHON

export PYSPARK\_PYTHON=/home/gyy/anaconda3/envs/pyspark/bin/python

export

PATH=$JAVA\_HOME/bin:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin:$SPARK\_HOME/bin:$PATH

使环境变量文件立即生效：

source /etc/profile

source ~/.bashrc

配置spark-env.sh：

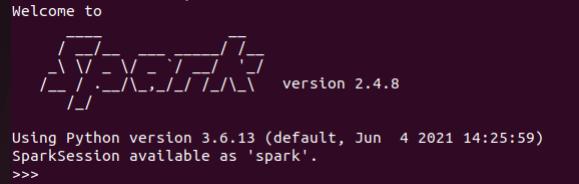
|  |
| --- |
| cd /opt/spark-2.4.0/conf  cp spark-env.sh.template spark-env.sh |

在该文件最后追加以下内容：

export JAVA\_HOME=/home/gyy/jdk1.8.0\_381

export SPARK\_DIST\_CLASSPATH=$(/home/gyy/hadoop-3.3.6/bin/hadoop classpath)

验证Spark是否安装成功：



可见spark可以正常开启

WebUI截图：

