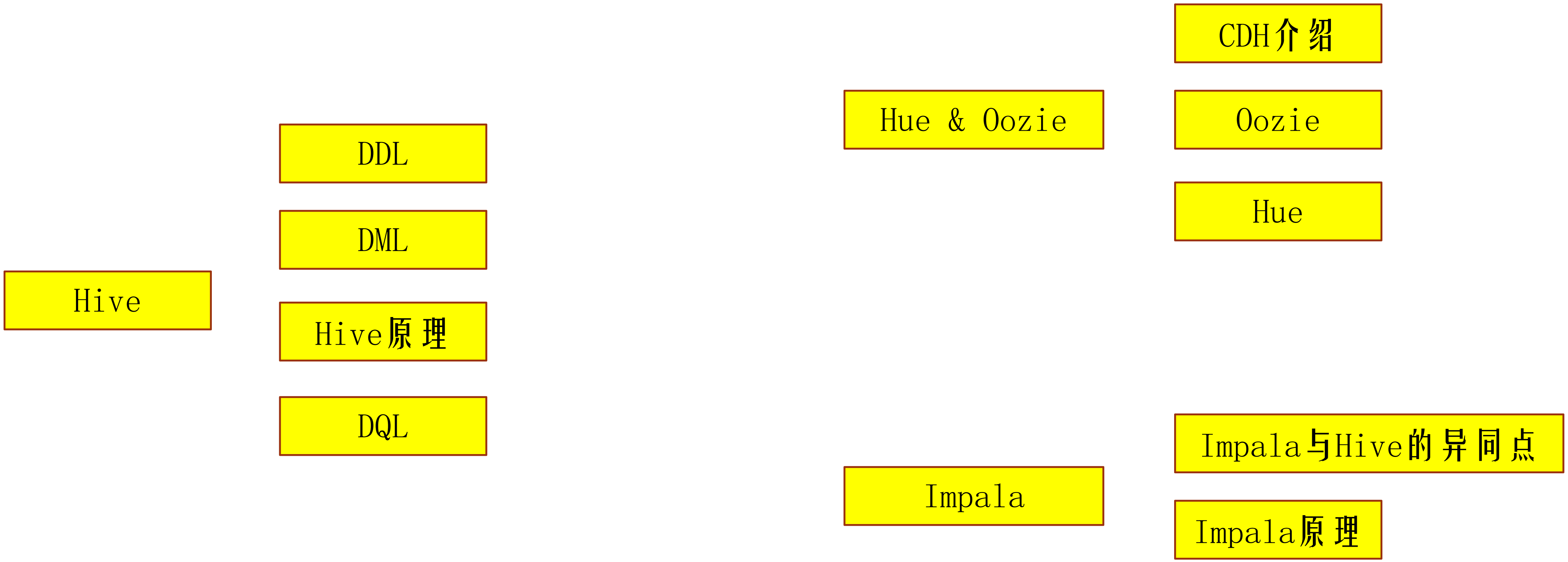
Hive - Impala

老汤

课程内容



- Hive
  - DDL
  - DML
  - Hive原理
  - DQL
- Hue & Oozie
  - CDH介绍
  - Oozie
  - Hue
- Impala
  - Impala与Hive的异同点
  - Impala原理

Hive数据模型 - DDL

1、Database

2、Table

3、Data Types

4、Partition

5、View

File Formats

1、File Formats

2、Serialization and Deserialization formats

Hive数据加载 - DML

1、Load files into tables

2、Inserting data into Hive tables from queries

3、Inserting data into dynamic partitions

4、Writing data into files from queries

5、Inserting values into tables from SQL

6、Bucket

6、skew

Hive数据查询 - DQL

1、SELECT

2、JOIN

3、Data Aggregation

4、Functions

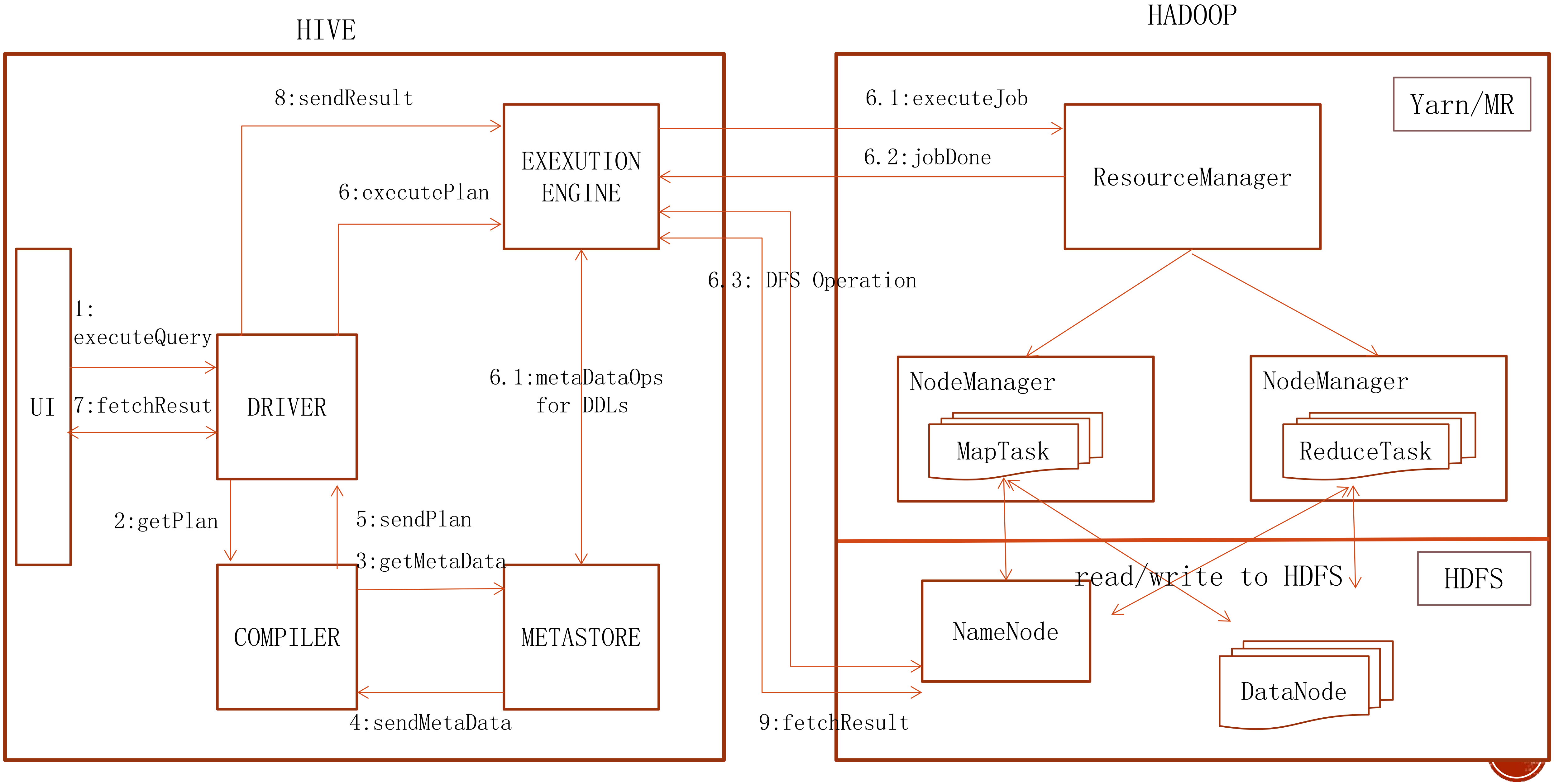
5、UDF、UDAF以及UDTF

Hive集成其他的系统

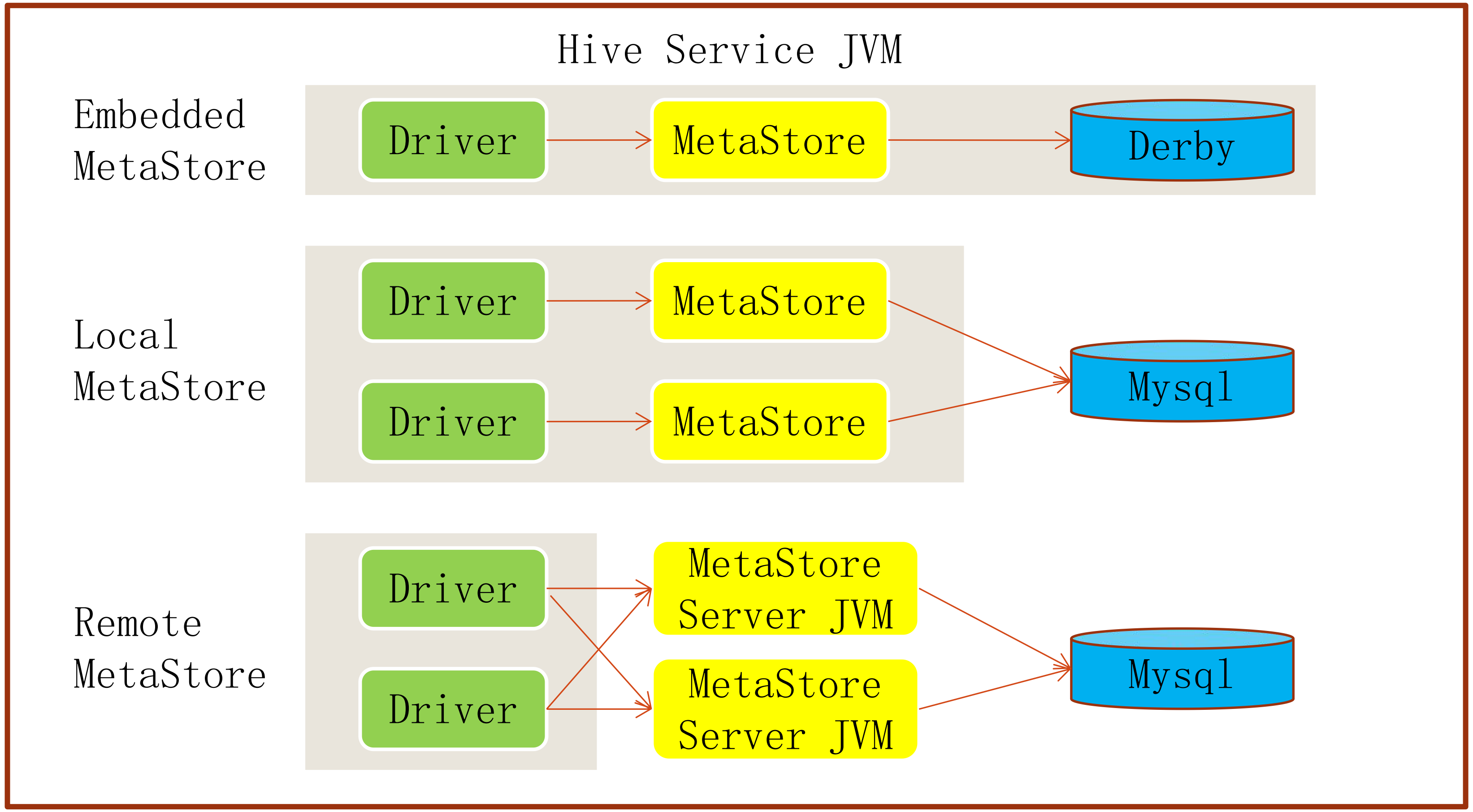
1、Hive On Spark

2、Hive On HBase

Hive On MR架构


HIVE
HADOOP
8:sendResult
EXEXUTION ENGINE
6:executePlan
1: executeQuery
UI
7:fetchResut
DRIVER
6.1:metaDataOps for DDLs
2:getPlan
5:sendPlan
3:getMetaData
COMPILER
METASTORE
4:sendMetaData
6.1:executeJob
6.2:jobDone
6.3: DFS Operation
9:fetchResult
Yarn/MR
ResourceManager
NodeManager
MapTask
NodeManager
ReduceTask
read/write to HDFS
HDFS
NameNode
DataNode

Hive安装的三种模式



Hive安装 - Remote MetaStore模式

在master机器上启动hive的remote metastore

```
mkdir ~/bigdata/apache-hive-2.3.3-bin/logs
nohup hive --service metastore > ~/bigdata/apache-hive-2.3.3-bin/logs/metastore.log 2>&1 &
```

在master上的$HIVE_HOME/conf/hive-site.xml中增加如下的配置：

```
<!-- thrift://<host_name>:<port> 默认端口是9083 -->
<property>
      <name>hive.metastore.uris</name>
      <value>thrift://master:9083</value>
      <description>Thrift uri for the remote metastore. Used by metastore client to connect to
    remote metastore.</description>
</property>
```

使得在master上启动的hive CLI走remote metastore服务访问元数据

Hive安装 - slave1访问hive

1、scp apache-hive-2.3.3-bin.tar.gz hadoop-twq@slave1:~/bigdata/

2、ssh登陆到slave1， 执行：

```
cd ~/bigdata
tar -xvf apache-hive-2.3.3-bin.tar.gz
cd apache-hive-2.3.3-bin/conf/
```

vi hive-site.xml，增加如下配置：

```
<configuration>
    <!-- thrift://<host_name>:<port> 默认端口是9083 -->
    <property>
      <name>hive.metastore.uris</name>
      <value>thrift://master:9083</value>
      <description>Thrift uri for the remote metastore. Used by metastore client to connect to remote metastore.</description>
    </property>
    <!--  hive表的默认存储路径 -->
    <property>
      <name>hive.metastore.warehouse.dir</name>
      <value>/user/hive/warehouse</value>
      <description>location of default database for the warehouse</description>
    </property>
</configuration>
```

Hive安装 - slave1访问hive

3、cp hive-env.sh.template hive-env.sh

```
vi hive-env.sh
HADOOP_HOME=/home/hadoop-twq/bigdata/hadoop-2.7.5
```

4、在slave1中配置hive环境变量：

```
vi ~/.bash_profile
export HIVE_HOME=/home/hadoop-twq/bigdata/apache-hive-2.3.3-bin
source ~/.bash_profile
```

5、将mysql的jdbc驱动包mysql-connector-java-5.*-bin.jar上传到~/bigdata/apache-hive-2.3.3-bin/lib下
将master有的mysql-jar包拷贝到slave1上相应的目录，在master机器上执行：

```
scp ~/bigdata/apache-hive-2.3.3-bin/lib/mysql-connector-java-5.1.44-bin.jar hadoop-twq@slave1:~/bigdata/apache-hive-2.3.3-bin/lib/
```

6、在master机器上启动hive的metastore

```
mkdir ~/bigdata/apache-hive-2.3.3-bin/logs
nohup hive -service metastore > ~/bigdata/apache-hive-2.3.3-bin/logs/metastore.log 2>&1 &
```

7、在slave1上执行hive命令

HiveServer2

1、停止master机器上的hiveserver2： ps -aux| grep hiveserver2

2、在slave1中打开hiveserver2:

```
mkdir ~/bigdata/apache-hive-2.3.3-bin/logs
nohup $HIVE_HOME/bin/hiveserver2 > ~/bigdata/apache-hive-2.3.3-bin/logs/hiveserver2.log 2>&1 &
```

3、在master上通过beeline的方式访问hive:

```
beeline
beeline> !connect jdbc:hive2://slave1:10000
Connecting to jdbc:hive2://slave1:10000
Enter username for jdbc:hive2://slave1:10000: hadoop-twq
Enter password for jdbc:hive2://slave1:10000:
0: jdbc:hive2://slave1:10000> show databases;
+-----------------+--+
| database_name   |
+-----------------+--+
| default         |
| dml             |
| hive_learning   |
| hive_test       |
| twq             |
+-----------------+--+
```

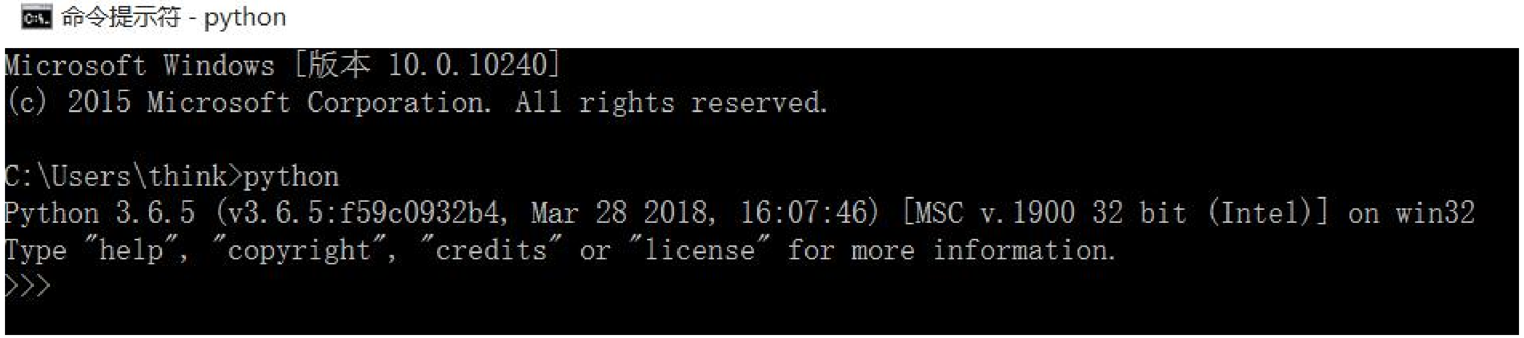
Python3开发环境安装

1.1、下载Python：https://www.python.org/

1.2、双击python-3.6.5.exe，进行傻瓜式安装

1.3、配置环境变量以及Path

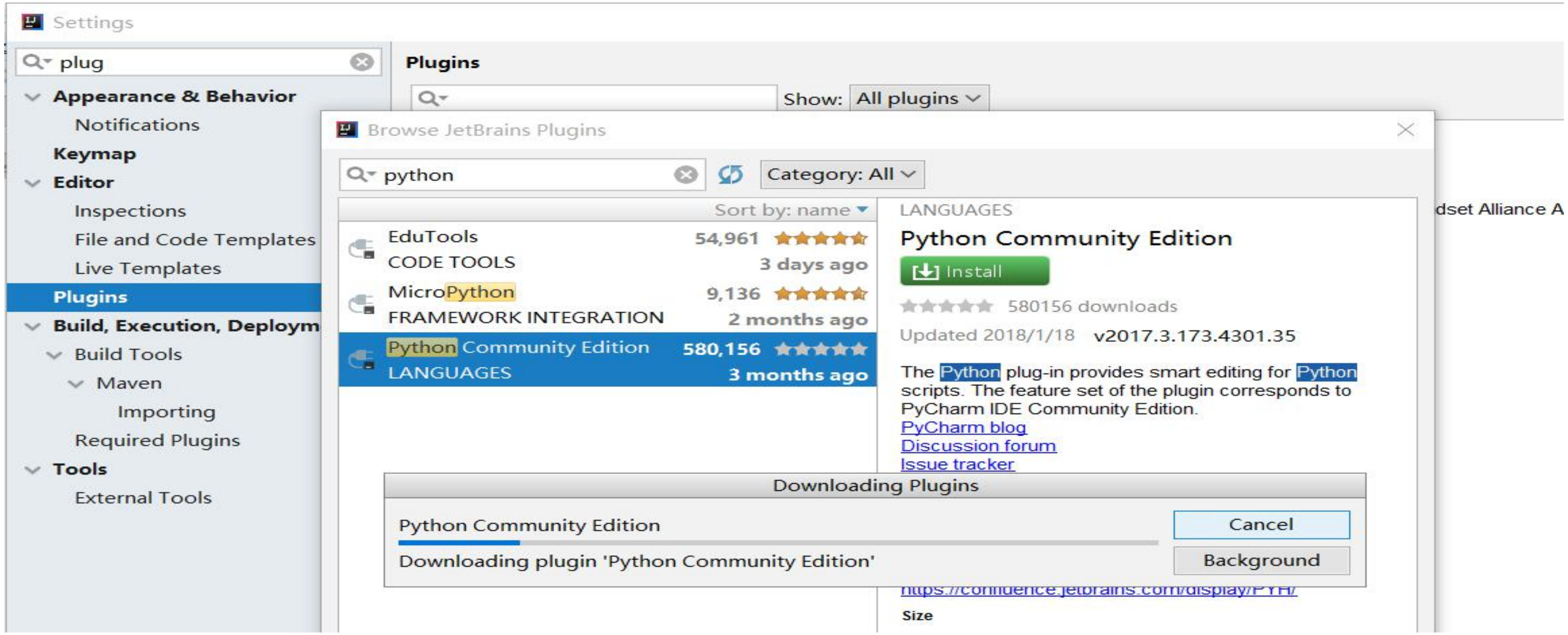
1.4、打开cmd进行测试：



```
Microsoft Windows [版本 10.0.10240]
(c) 2015 Microsoft Corporation. All rights reserved.

C:\Users\think>python
Python 3.6.5 (v3.6.5:f59c0932b4, Mar 28 2018, 16:07:46) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

Python3开发环境安装

2.1、IntelliJ IDEA中安装python插件，并重启IDEA



Python Community Edition

580156 downloads

Updated 2018/1/18 v2017.3.173.4301.35

The Python plug-in provides smart editing for Python scripts. The feature set of the plugin corresponds to PyCharm IDE Community Edition.

PyCharm blog

Discussion forum

Issue tracker

https://confluence.jetbrains.com/display/PYH/

Downloading Plugins

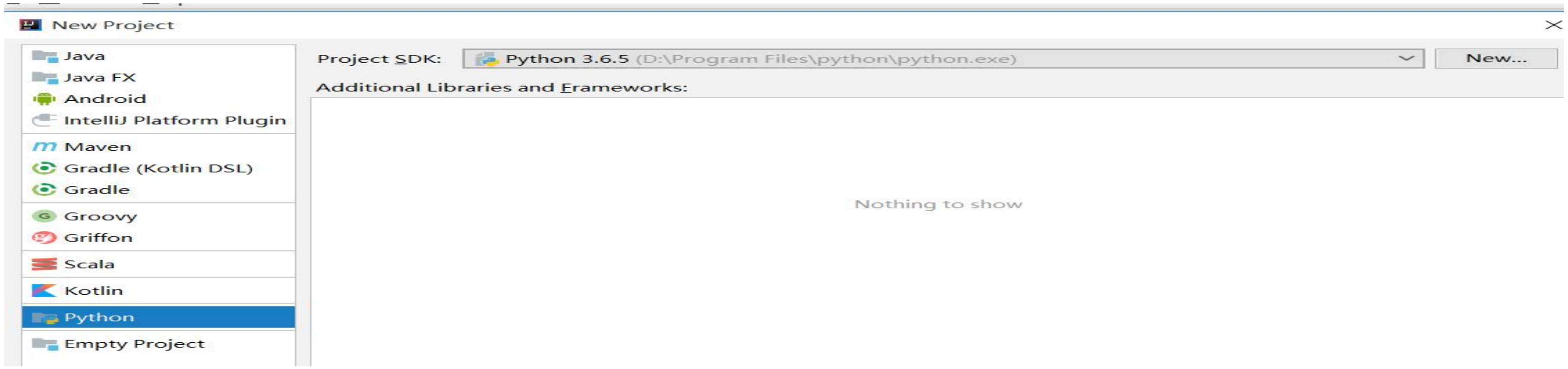
Python Community Edition

Downloading plugin 'Python Community Edition'

如果这种方法不能安装的话，则可以在https://plugins.jetbrains.com/plugin/631-python中下载插件进行安装

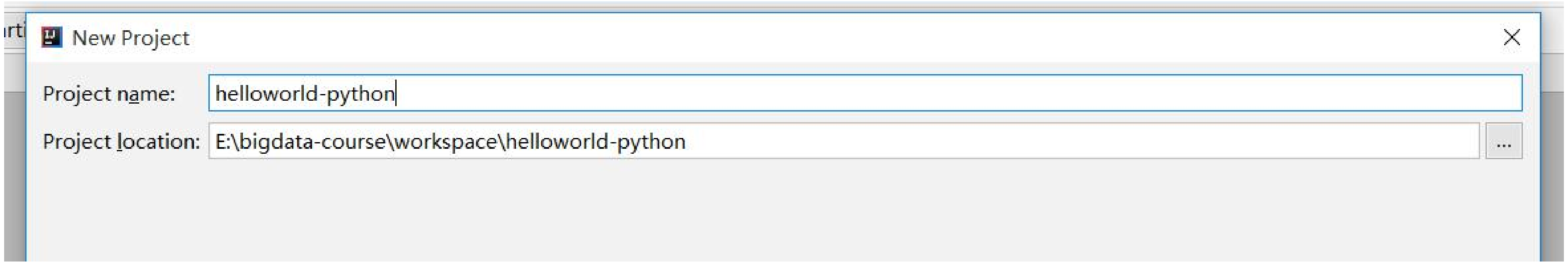
Python3开发环境安装

2.2、创建python项目并运行



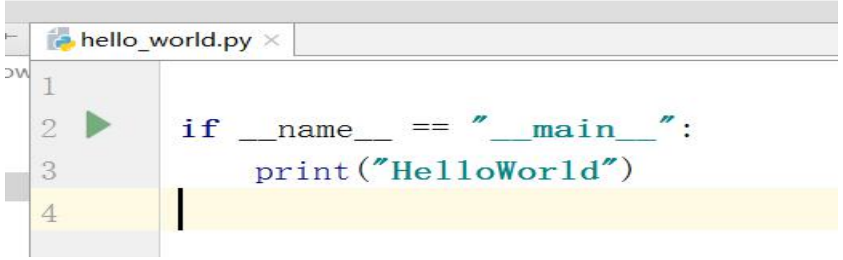
Project SDK: Python 3.6.5 (D:\Program Files\python\python.exe)

Additional Libraries and Frameworks:



Project name: helloworld-python

Project location: E:\bigdata-course\workspace\helloworld-python



```
if __name__ == "__main__":
    print("HelloWorld")
```

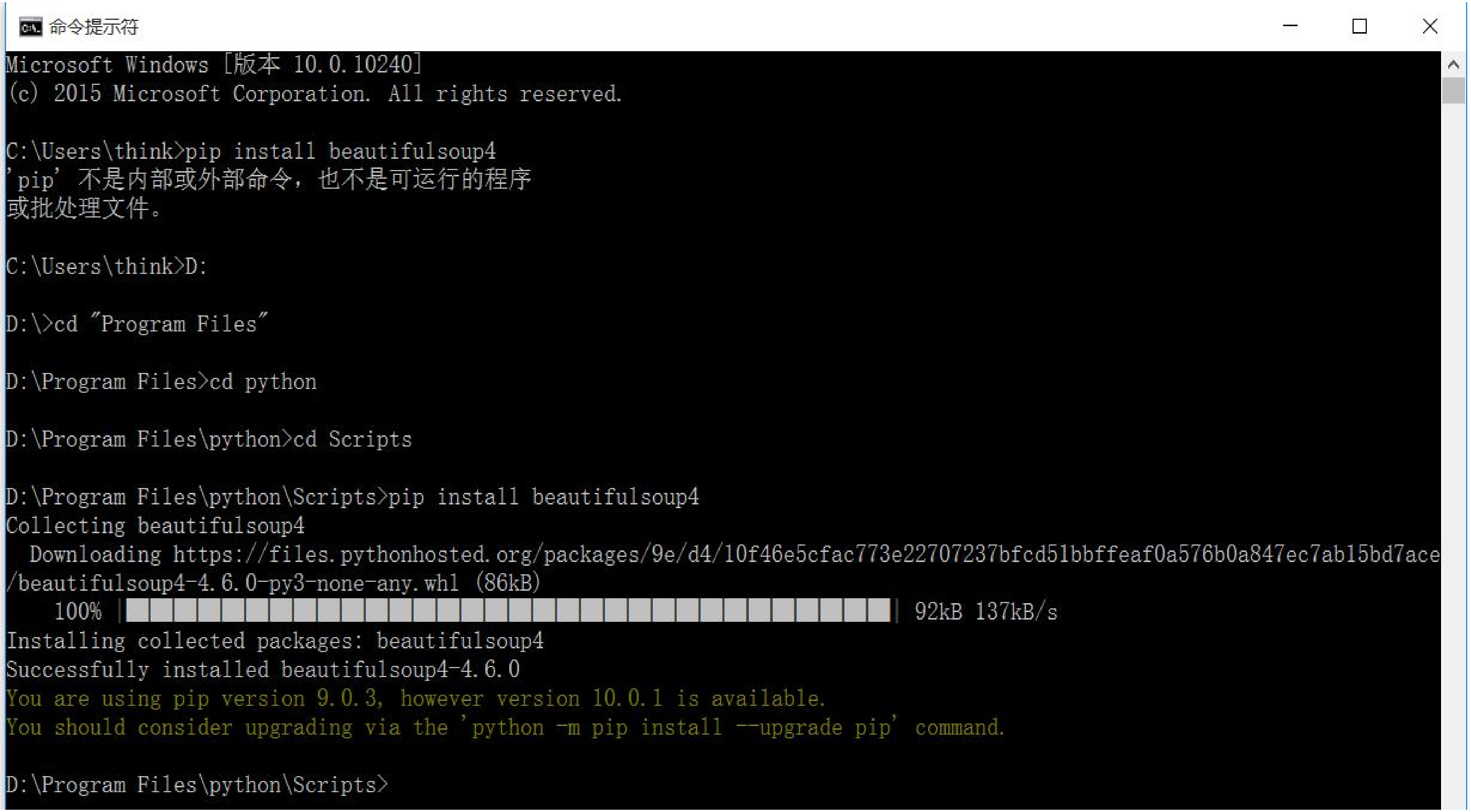
Python3开发爬虫爬出豆瓣电影的数据

chromedriver下载的页面：
https://chromedriver.storage.googleapis.com/index.html?path=2.37/

```
pip install selenium
```

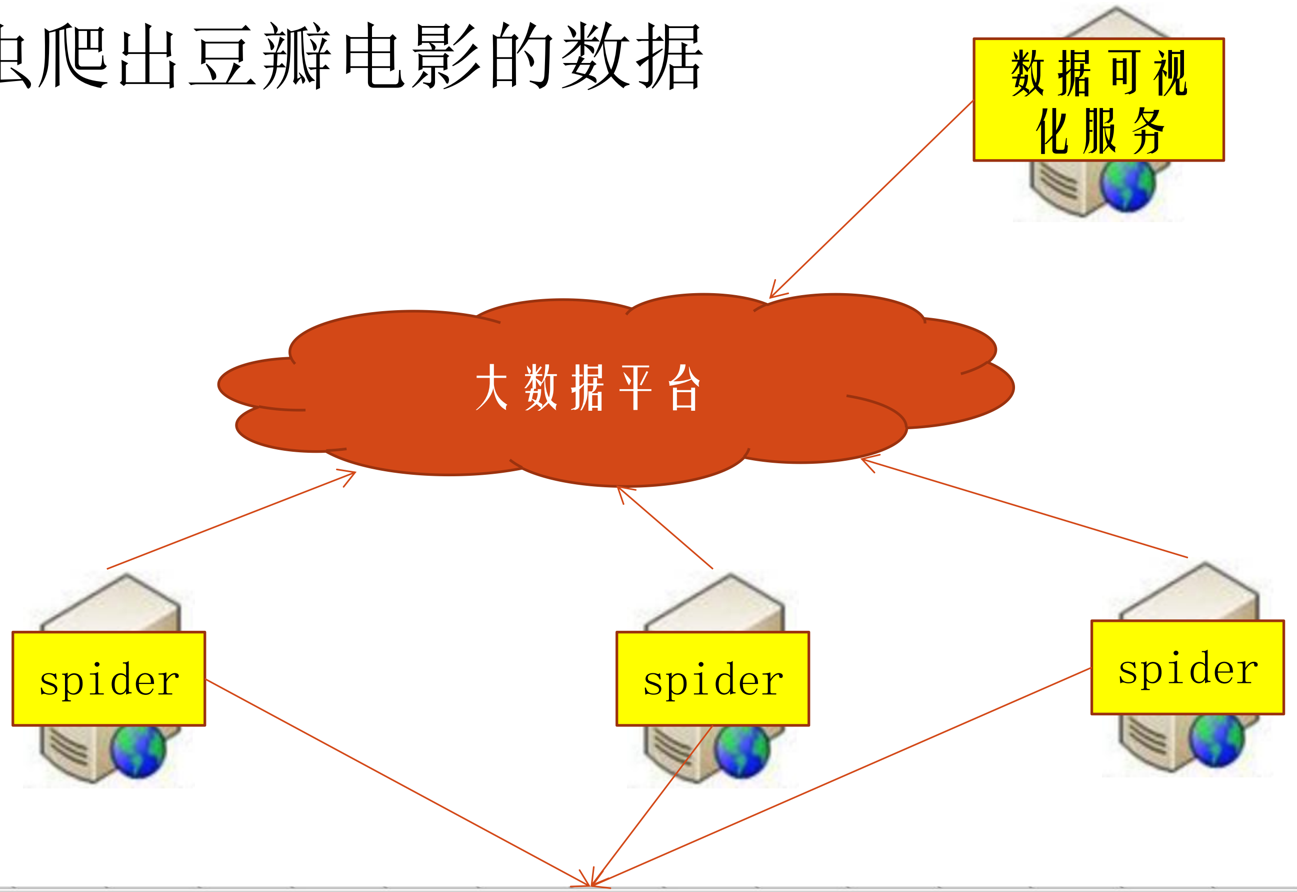
Python3开发环境安装

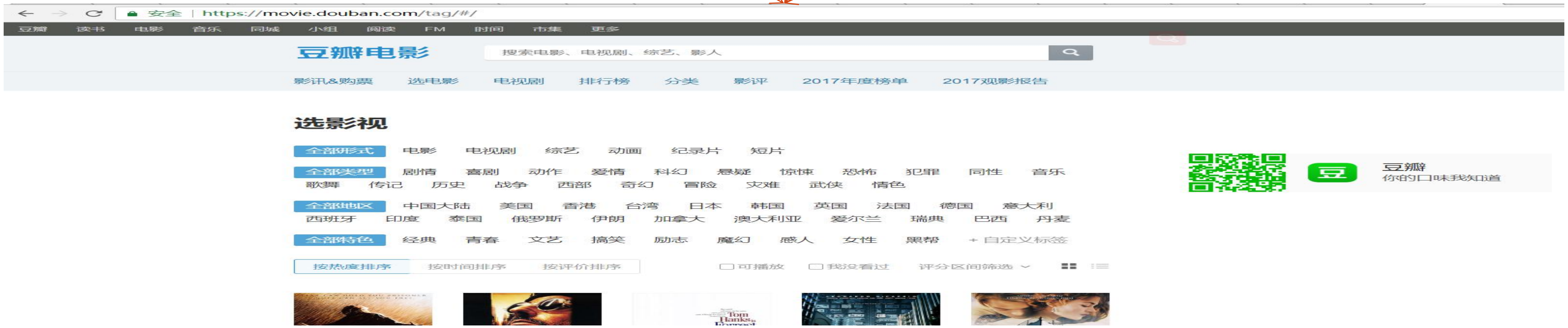
2.3、依赖包的安装



```
Microsoft Windows [版本 10.0.10240]
(c) 2015 Microsoft Corporation. All rights reserved.

C:\Users\think>pip install beautifulsoup4
'pip' 不是内部或外部命令，也不是可运行的程序
或批处理文件。

C:\Users\think>D:

D:\>cd "Program Files"

D:\Program Files>cd python

D:\Program Files\python>cd Scripts

D:\Program Files\python\Scripts>pip install beautifulsoup4
Collecting beautifulsoup4
  Downloading https://files.pythonhosted.org/packages/9e/d4/10f46e5cfac773e22707237bfcd51bbffeaf0a576b0a847ec7ab15bd7ace
/beautifulsoup4-4.6.0-py3-none-any.whl (86kB)
    100% |████████████████████████████████| 92kB 137kB/s
Installing collected packages: beautifulsoup4
Successfully installed beautifulsoup4-4.6.0
You are using pip version 9.0.3, however version 10.0.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.

D:\Program Files\python\Scripts>
```

Python3开发爬虫爬出豆瓣电影的数据




安全
https://movie.douban.com/tag/#/
豆瓣 读书 电影 音乐 同城 小组 阅读 FM 时间 市集 更多
豆瓣电影
搜索电影、电视剧、综艺、影人
影讯&购票 选电影 电视剧 排行榜 分类 影评 2017年度榜单 2017观影报告

选影视

全部形式 电影 电视剧 综艺 动画 纪录片 短片

全部类型 剧情 喜剧 动作 爱情 科幻 悬疑 惊悚 恐怖 犯罪 同性 音乐 歌舞 传记 历史 战争 西部 奇幻 冒险 灾难 武侠 情色

全部地区 中国大陆 美国 香港 台湾 日本 韩国 英国 法国 德国 意大利 西班牙 印度 泰国 俄罗斯 伊朗 加拿大 澳大利亚 爱尔兰 瑞典 巴西 丹麦

全部特色 经典 青春 文艺 搞笑 励志 魔幻 感人 女性 黑帮 + 自定义标签

按热度排序 按时间排序 按评价排序
□可播放 □我没看过 评分区间筛选
豆瓣
你的口味我知道
Tom Hanks

上传数据

1、本地hadoop配置JAVA_HOME，修改%HADOOP_HOME%\etc\hadoop\hadoop-env.cmd

```
set JAVA_HOME="D:\Program Files"\Java\jdk1.8.0_161
```

2、在master上执行：

```
hadoop fs -mkdir /user/hadoop-twq/hive-course/douban
hadoop fs -chmod -R 757 /user/hadoop-twq/hive-course/douban
```

3、在本地打开cmd执行：

```
E:
cd bigdata-course\workspace\sql-on-hadoop\doban-analysis\spider\file_output
hadoop fs -put movie-video.csv hdfs://master-dev:9999/user/hadoop-twq/hive-course/douban
hadoop fs -put links\movie-video_links.csv hdfs://master-dev:9999/user/hadoop-twq/hive-course/douban/links
```