# 第02阶段_数据采集及处理



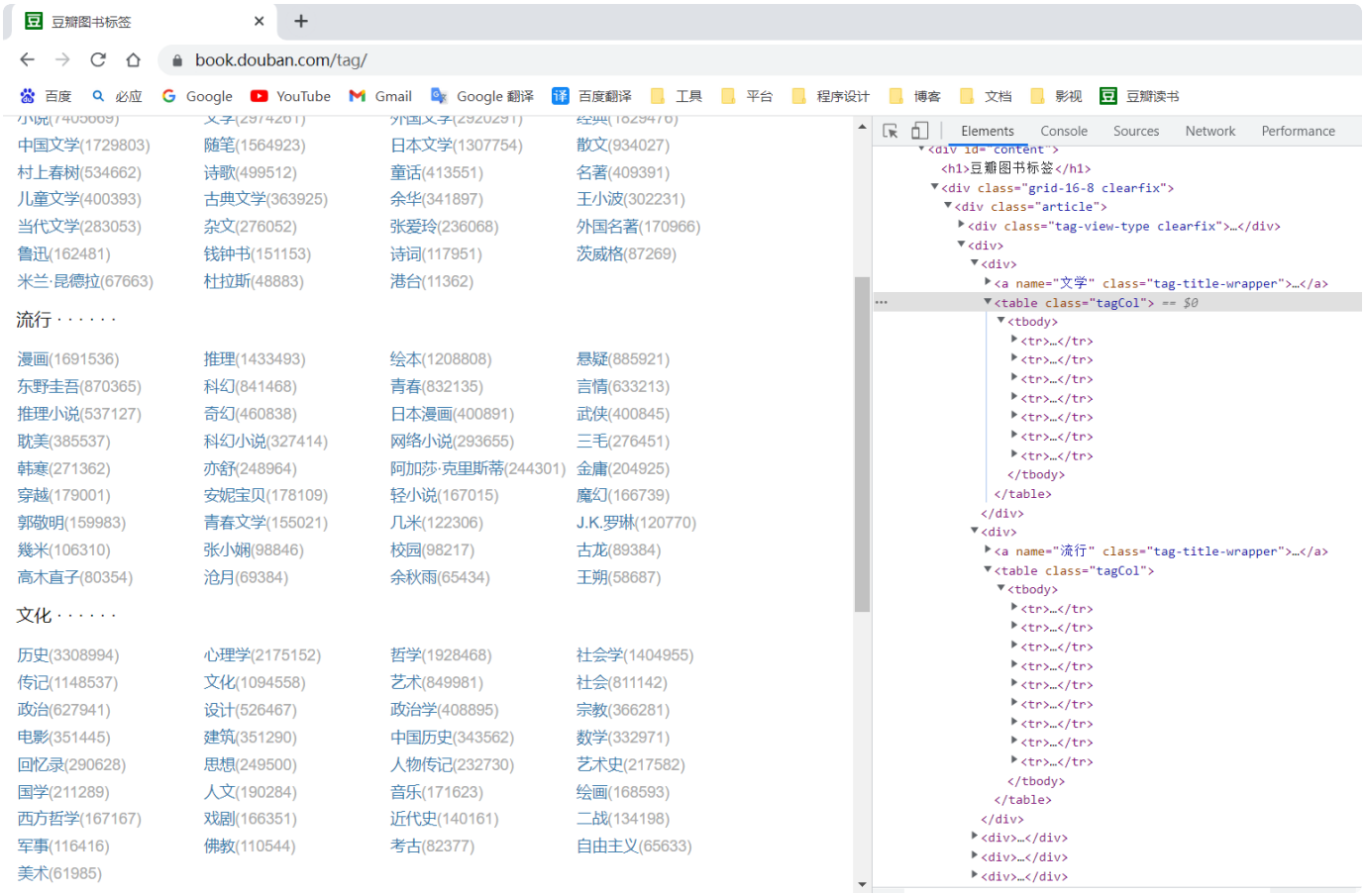
# 1. 数据采集

## 1.1 数据源

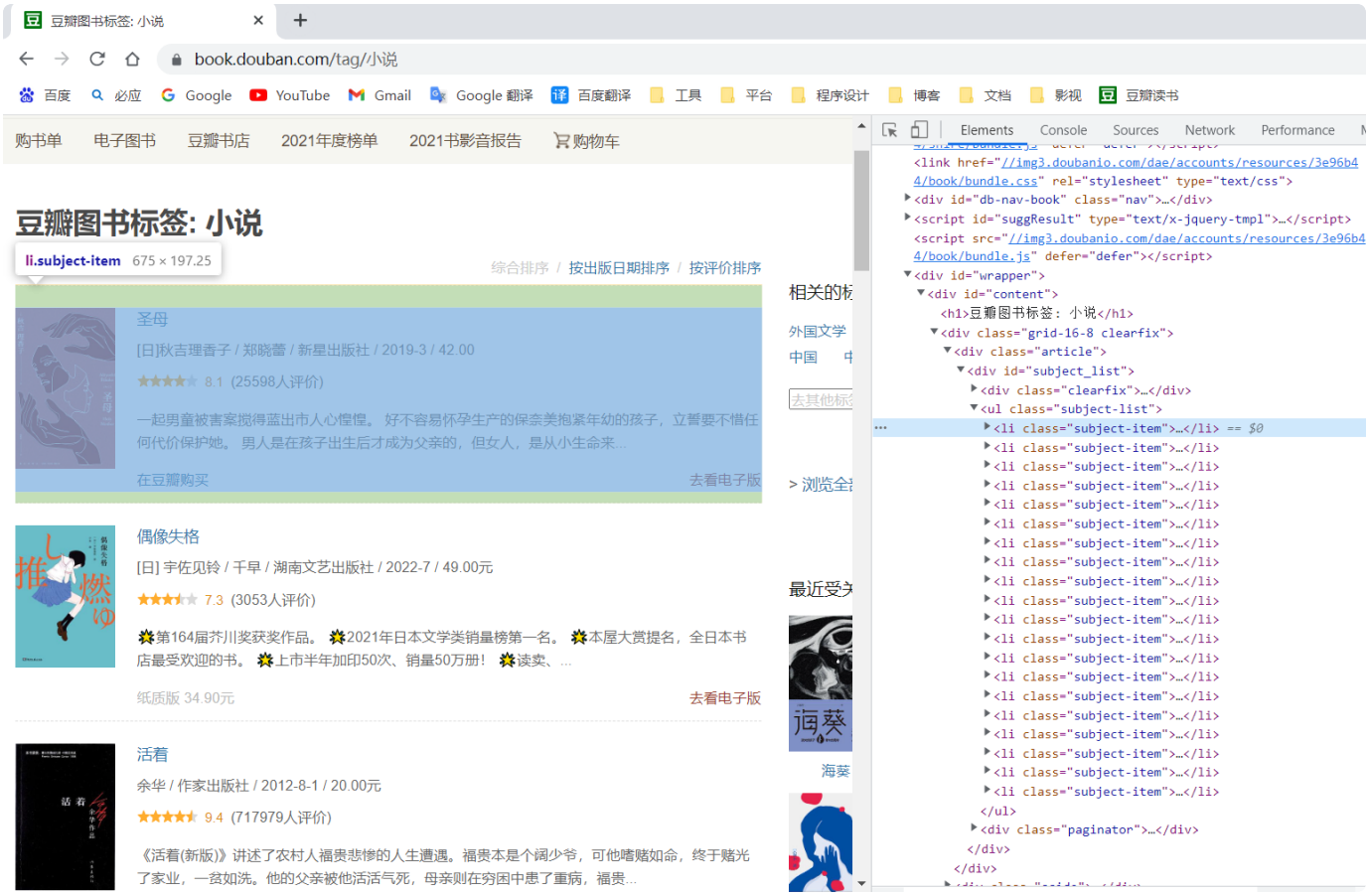
`url` : https://book.douban.com/tag/

## 1.2 网页数据结构

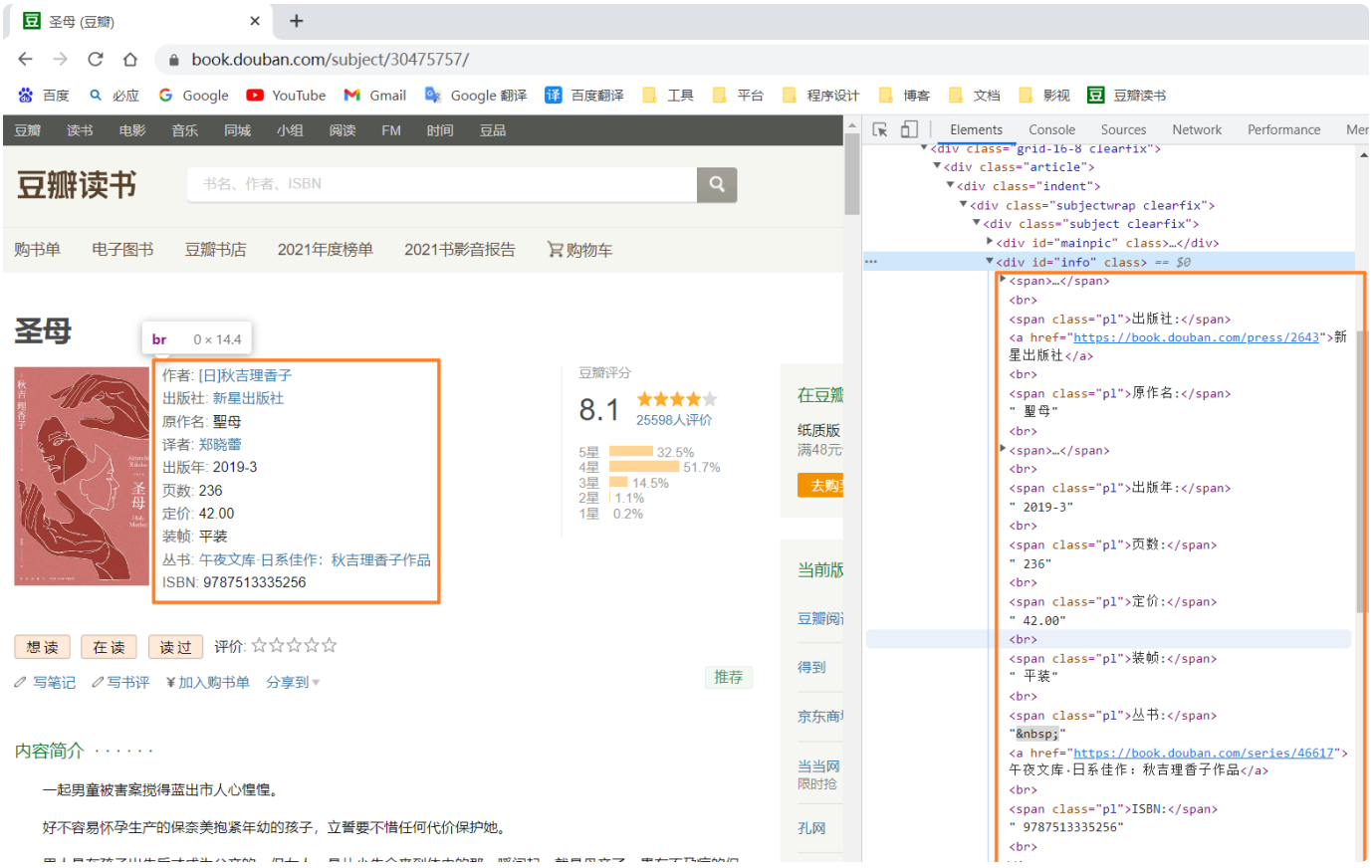
一级页面


豆瓣图书标签
book.douban.com/tag/
百度 必应 Google YouTube Gmail Google 翻译 百度翻译 工具 平台 程序设计 博客 文档 影视 豆瓣读书
中国文学(1729803) 随笔(1564923) 日本文学(1307754) 散文(934027)
村上春树(534662) 诗歌(499512) 童话(413551) 名著(409391)
儿童文学(400393) 古典文学(363925) 余华(341897) 王小波(302231)
当代文学(283053) 杂文(276052) 张爱玲(236068) 外国名著(170966)
鲁迅(162481) 钱钟书(151153) 诗词(117951) 茨威格(87269)
米兰·昆德拉(67663) 杜拉斯(48883) 港台(11362)
流行
漫画(1691536) 推理(1433493) 绘本(1208808) 悬疑(885921)
东野圭吾(870365) 科幻(841468) 青春(832135) 言情(633213)
推理小说(537127) 奇幻(460838) 日本漫画(400891) 武侠(400845)
耽美(385537) 科幻小说(327414) 网络小说(293655) 三毛(276451)
韩寒(271362) 亦舒(248964) 阿加莎·克里斯蒂(244301) 金庸(204925)
穿越(179001) 安妮宝贝(178109) 轻小说(167015) 魔幻(166739)
郭敬明(159983) 青春文学(155021) 几米(122306) J.K.罗琳(120770)
幾米(106310) 张小娴(98846) 校园(98217) 古龙(89384)
高木直子(80354) 沧月(69384) 余秋雨(65434) 王朔(58687)
文化
历史(3308994) 心理学(2175152) 哲学(1928468) 社会学(1404955)
传记(1148537) 文化(1094558) 艺术(849981) 社会(811142)
政治(627941) 设计(526467) 政治学(408895) 宗教(366281)
电影(351445) 建筑(351290) 中国历史(343562) 数学(332971)
回忆录(290628) 思想(249500) 人物传记(232730) 艺术史(217582)
国学(211289) 人文(190284) 音乐(171623) 绘画(168593)
西方哲学(167167) 戏剧(166351) 近代史(140161) 二战(134198)
军事(116416) 佛教(110544) 考古(82377) 自由主义(65633)
美术(61985)
Elements Console Sources Network Performance
<h1>豆瓣图书标签</h1>
<table class="tagCol"> == $0

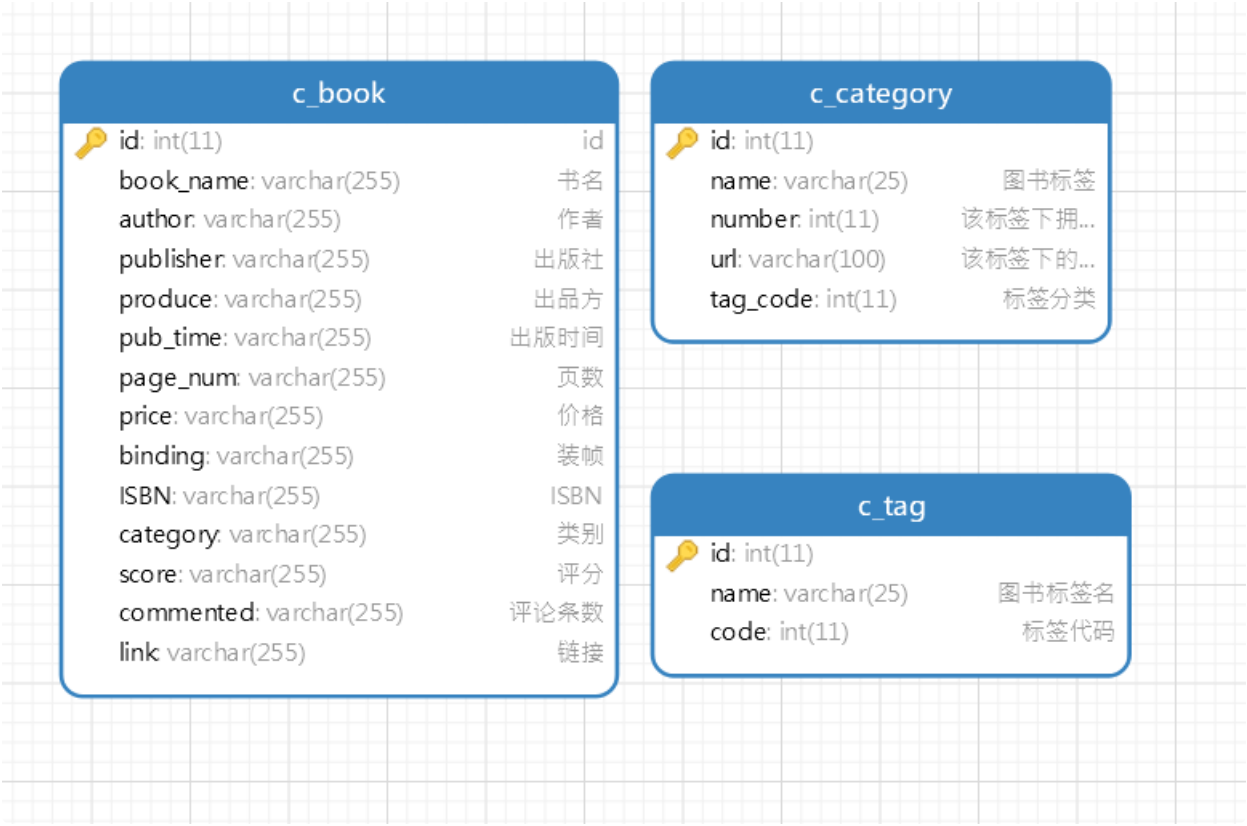
## 二级页面


豆瓣图书标签: 小说
book.douban.com/tag/小说
豆瓣图书标签: 小说
li.subject-item 675 × 197.25
圣母
[日]秋吉理香子 / 郑晓蕾 / 新星出版社 / 2019-3 / 42.00
8.1 (25598人评价)
偶像失格
[日] 宇佐见铃 / 千早 / 湖南文艺出版社 / 2022-7 / 49.00元
7.3 (3053人评价)
活着
余华 / 作家出版社 / 2012-8-1 / 20.00元
9.4 (717979人评价)
Elements
Console
Sources
Network
Performance

## 三级页面


圣母 (豆瓣)
book.douban.com/subject/30475757/
豆瓣读书
圣母
br 0 × 14.4
作者: [日]秋吉理香子
出版社: 新星出版社
原作名: 聖母
译者: 郑晓蕾
出版年: 2019-3
页数: 236
定价: 42.00
装帧: 平装
丛书: 午夜文库·日系佳作：秋吉理香子作品
ISBN: 9787513335256
豆瓣评分
8.1
25598人评价
想读
在读
读过
内容简介
一起男童被害案搅得蓝出市人心惶惶。
好不容易怀孕生产的保奈美抱紧年幼的孩子，立誓要不惜任何代价保护她。
Elements
Console
Sources
Network
Performance

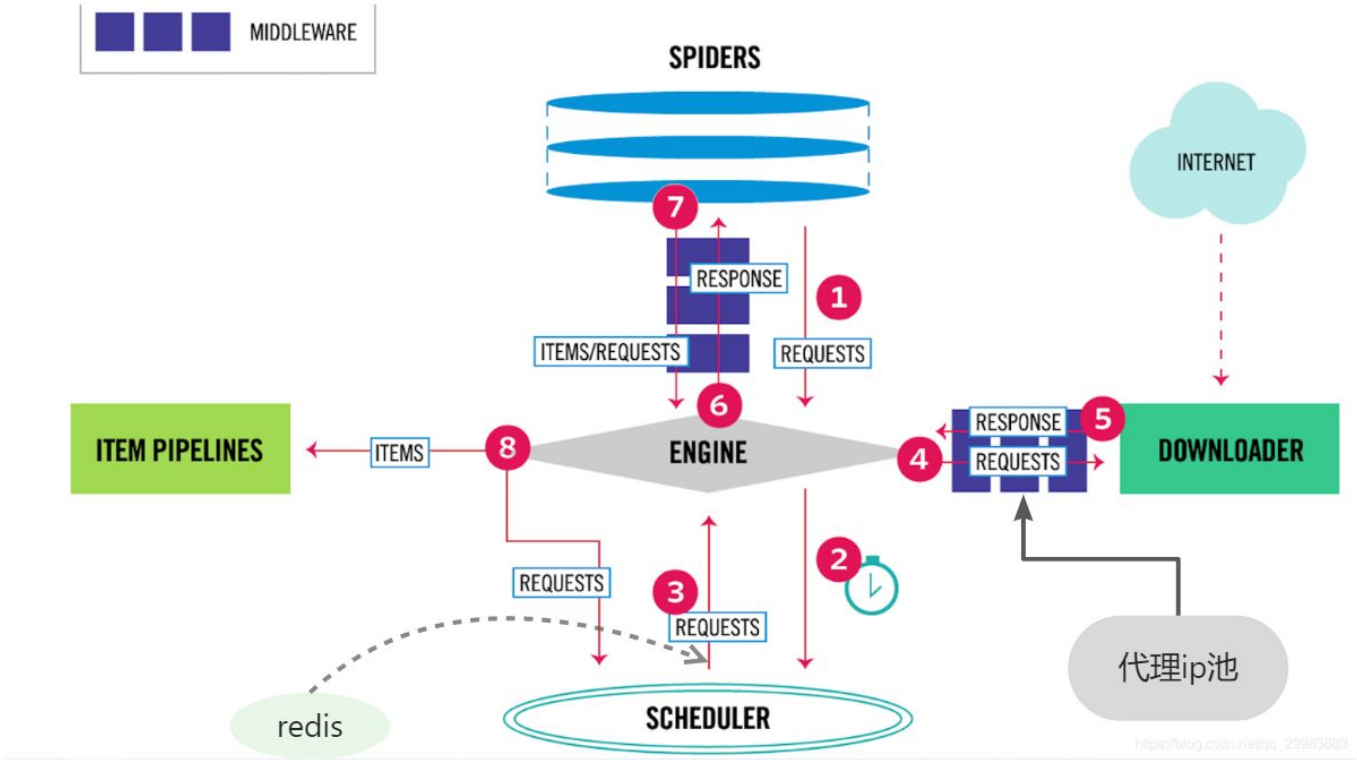
## 1.3 采集字段&&数据库设计




## 1.4 爬虫程序架构

爬虫框架: scrapy + redis + mysql

架构图:


MIDDLEWARE
SPIDERS
INTERNET
7
RESPONSE
1
ITEMS/REQUESTS
REQUESTS
6
ITEM PIPELINES
ITEMS
8
ENGINE
RESPONSE
5
4
REQUESTS
DOWNLOADER
2
REQUESTS
3
REQUESTS
代理ip池
redis
SCHEDULER

## 1.5 程序代码结构

```
├─README.md
├─test                                                    # 测试代码文件
|  ├─extract.py
├─log                                                     # 日志文件
|  ├─book-2022-09-20.log
├─crawler                                                 # 爬虫文件夹
|    ├─category.py                                       # 一级页面采集程序
|    ├─__init__.py
|    ├─douban_book                                       # scrapy项目
|    |      ├─scrapy.cfg
|    |      ├─douban_book
|    |      |      ├─items.py                           # 定义采集字段
|    |      |      ├─middlewares.py                     # 中间件, 配置代理ip
|    |      |      ├─pipelines.py                       # 管道文件, 保存数据到mysql
|    |      |      ├─settings.py
|    |      |      ├─__init__.py
|    |      |      ├─spiders
|    |      |      |    ├─book.py                       # spider主文件
|    |      |      |    ├─run.py                        # scrapy启动文件
|    |      |      |    ├─__init__.py
├─config                                                  # msyql,redis等配置
|    ├─setting.py
|    ├─__init__.py
├─common
|    ├─common_enum.py
|    ├─mylogger.py                                       # 日志工具类
|    ├─mysql.py                                          # msyql CRUD工具类
|    ├─proxy.py                                          # 代理ip提取工具类
|    ├─__init__.py
|    ├─西刺代理ip_demo
```

## 1.6 遇到的问题

| 问题 | 解决方法 |
| --- | --- |
| 网站封禁ip | 在middlewares.py中配置代理ip |
| url重复爬取 | 使用redis缓存, 在爬虫程序启动时, 将数据库中的url缓存到redis, 在提取到url时, 判断是否存在于redis,如果存在, 则跳过不爬取 |
| 页面数据解析 | xpath与css语法解析源码提取数据比较困难, 改用re正则表达式的方式提取网页数据 |

# 2 数据预处理

## 2.1 导入数据到hdfs

使用sqoop将采集到mysql中的数据集, 导入到hdfs

1. 由于sqoop导入数据到hdfs字段顺序与mysql不一致, 为了spark程序能够精准读取数据,

   所以指定 `query` 参数, 将bigdata数据库下c_book表中, 如下字段按顺序导入hdfs:

   book_name, author, publisher, produce, pub_time, page_num,

   price, binding, ISBN, category, score, commented

2. 指定字段分割符为" `\001` ", 是由于采集到的数据中可能含有" `,` ", " `\t` "等特殊字符, 若使用这些作为分割符, 可能导致spark切割每一行得到的array长度不一致(即列数不一致).

```
sqoop import \
--connect jdbc:mysql://master:3306/bigdata \
--username root \
--password 123456 \
--query 'select id, book_name, author, publisher,produce,pub_time,page_num,price,binding,ISBN,category,score,commented,link from c_book where 1=1 and $CONDITIONS' \
--target-dir /bigdata/book/ \
--delete-target-dir \
--fields-terminated-by '\001' \
--hive-drop-import-delims \
--split-by id \
--m 5
```

## 2.2 idea开发spark程序关于hdfs的高可用访问配置

Scala | 复制代码

```scala
val spark: SparkSession.Builder = SparkSession
.builder()
.appName("app")
.getOrCreate()

spark.sparkContext.hadoopConfiguration.set("fs.defaultFS", "hdfs://mycluster")
spark.sparkContext.hadoopConfiguration.set("dfs.nameservices", "mycluster")

spark.sparkContext.hadoopConfiguration.set("dfs.ha.namenodes.mycluster", "master,slave1,slave2")
spark.sparkContext.hadoopConfiguration.set("dfs.namenode.rpc-address.mycluster.master", "master:8020")
spark.sparkContext.hadoopConfiguration.set("dfs.namenode.rpc-address.mycluster.slave1", "slave1:8020")
spark.sparkContext.hadoopConfiguration.set("dfs.namenode.rpc-address.mycluster.slave2", "slave2:8020")
spark.sparkContext.hadoopConfiguration.set("dfs.client.failover.proxy.provider.mycluster", "org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider")
```

## 2.3 程序代码结构

```
spark                                                    Python | 复制代码
1  ├─pom.xml                                                  # pon依赖
2  ├─README.md                                                #
3  |  ├─main
4  |  |  ├─scala
5  |  |  |   ├─com
6  |  |  |   |  ├─zys
7  |  |  |   |  |  ├─spark
8  |  |  |   |  |  |   ├─util
9  |  |  |   |  |  |   |  ├─MyUdf.scala                      # 自定义udf函数
10 |  |  |   |  |  |   |  └SparkUtils.scala                  # 配置sparksession
11 |  |  |   |  |  |   ├─preprocess
12 |  |  |   |  |  |   |     └run.scala                      # 数据处理主 程序
13 |  |  |   |  |  |   ├─pojo
14 |  |  |   |  |  |   |  └Book.scala                        # 样例类, 对应hdfs
   文件的各列
15 |  |  |   |  |  |   ├─config
16 |  |  |   |  |  |   |   └ApplicationConfig.scala          # 配置类
17 |  |  |   |  |  |   ├─common
18 |  |  ├─resources
19 |  |  |     ├─config.properties                           # 配置文件
20 |  |  |     └log4j.properties
```

## 2.4 spark-submit

1. 修改config.properties文件

   app.is.local=false

2. 再idea中使用maven将程序打成jar包
3. 提交jar包到spark-yarn

```
1  bin/spark-submit \
2  --class com.zys.spark.preprocess.run \
3  --master yarn \
4  --deploy-mode cluster \
5  /root/book.jar \
6  10
```

4. 运行成功

```
22/10/02 21:15:35 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:36 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:37 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:38 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:39 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:40 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:41 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:42 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:43 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:44 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:45 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:46 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:47 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:48 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:49 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:50 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:51 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:52 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:53 INFO Client: Application report for application_1664689047924_0002 (state: RUNNING)
22/10/02 21:15:54 INFO Client: Application report for application_1664689047924_0002 (state: FINISHED)
22/10/02 21:15:55 INFO Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: slave2
         ApplicationMaster RPC port: 33696
         queue: default
         start time: 1664716411919
         final status: SUCCEEDED
         tracking URL: http://slave2:8088/proxy/application_1664689047924_0002/
         user: root
22/10/02 21:15:55 INFO ShutdownHookManager: Shutdown hook called
22/10/02 21:15:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-0f020003-9739-44d3-9c1a-4a064e735ec4
22/10/02 21:15:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-6bcd4703-2729-4768-8b0b-6a8901ed9a34
[root@master spark]#
```

# 3. 遇到的一些BUG

## 3.1 msyql

1. Can't connect to local MySQL server through socket '/var/lib/mysql/mysql.sock' (111)

```
# 解决方法
rm -f /var/lib/mysql/mysql.sock
systemctl restart mysqld
```

## 3.2 hdfs

1. INFO retry.RetryInvocationHandler: org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.ipc.StandbyException): Operation category READ is not supported in state standby. Visit https://s.apache.org/sbnn-error

解决方法: https://118k.tistory.com/1052

hdfs-site.xml XML 复制代码

```xml
<property>
    <name>dfs.ha.allow.stale.reads</name>
    <value>true</value>
</property>
```

## 3.3 hive&&spark

1. window下spark写入数据到hive权限被拒

   赋予相应的权限