

**《分布式系统课程设计》**

**项目报告**

**项目名称：基于hive的电商数据分析平台的实现与应用**

班级： 大数据201

成员： 闫浩勇 刘文龙 杨发元 岳文韬 赵佳伟

学号: 2020310835 2020310812 2020310819 2020310816 2020310827

教师： 孟祥冰

信息工程学院计算机系

1. 项目介绍
2. 项目简介

随着移动互联网的快速发展以及智能手机的快速普及,越来越多的人们从线下购物转向了线上购物,因为线上商品相比线下实体店来说价格相对更加便宜,购买更方便,由此也催生了很多电商平台的飞速发展，在这竞争激烈的互联网电商平台背景下,电商企业如何抓住用户流量,提高用户粘性,增加自身的GMV,已经成了必须要考虑的问题。因此，将开展基于Hadoop + Hive+MySQL实现对电商平台的大数据分析。

主要分为数据采集模块,数据仓库和数据可视化模块这三个模块。数据采集模块主要负责采集用户的行为数据和业务数据,用户行为数据包括启动、页面、事件、曝光、错误这5类,业务数据包括订单相关的数据、用户相关的数据、商品相关的数据、活动相关的数据、地区相关的数据。数据仓库主要进行数据的清洗、建模、分析,整体分为了5层,分别是ODS层（原始数据层）,DWD层（数据明细层）,DWS层（数据服务层）,DWT层（数据主题层）,ADS层（数据应用层）。数据可视化模块主要是以ADS层结果数据为模板,在My SQL中创建相应的表,使用Sqoop工具定时将ADS层结果数据导出到My SQL中,并使用数据可视化工具对数据进行直观的展示。本平台采用Python、Flume、Sqoop作为数据采集工具,HDFS作为数据存储框架,Hive作为数据仓库工具,Spark作为Hive计算引擎,以目前大数据时代的电商企业数据分析需求为背景,搭建了一个全流程的电商大数据分析平台,涵盖数据采集、数据存储、数据分析、数据展示。

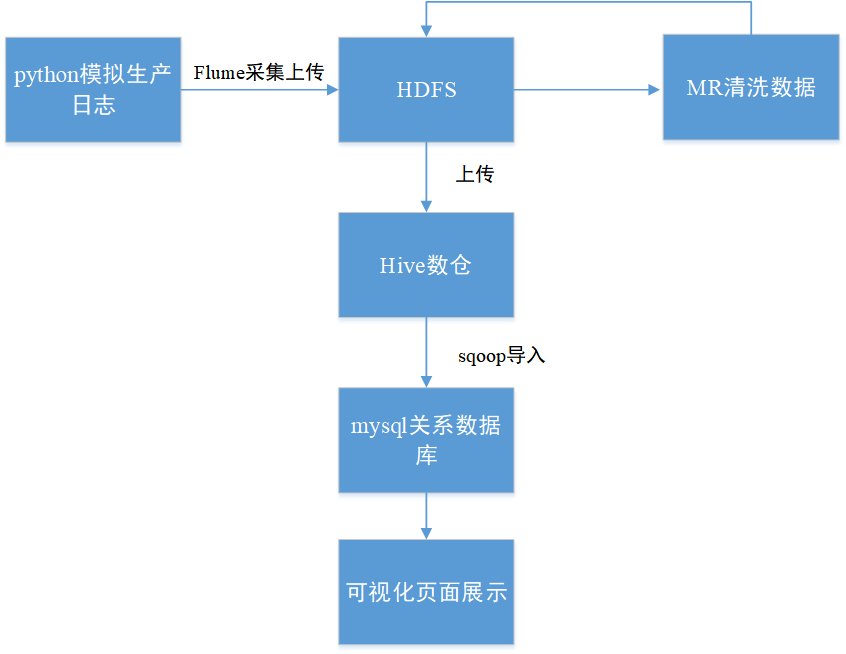
1. 项目背景

在目前的移动互联网加大数据时代，越来越多的人们开始用上了智能手机，选择在网上购物的人也越来越多，这种趋势也导致了许多电商企业的诞生。随着用户数量的日益增长，用户在电商 APP 上浏览、购物等操作产生的数据量也越来越多，但是有些电商企业还是用的传统的数据库，无法承接较大的数据量，也无法利用已有的数据对各种指标进行综合分析，这样很容易在与其他公司的竞争中落后。据不完全统计，2015 年止，已有注册超过1万家各式各样的电商企业，有超过3 亿用户在电商网站下留下埋点数据。随着电商“新零售”下的电商新布局以及 5G 时代的加速来临，到 2020 年，全球包含 PC、平板电脑、智能手机等联网设备预计将超过 300 亿台。实际上，随着物联网技术与可穿戴设备的飞速发展，终端设备会远远大于这个数量，随之而来各种数据密集型应用会产生巨量的数据。根据国际分析机构IDC 的统计，全球不同设备产生的数据量，到2020 年预计将会突破 40ZB。如此海量、持续、细粒度、多样化的数据，让各个行业都看到了数据的巨大潜在价值。“新零售”尤其给电商五联网公司带水了重大的机遇，人为电商五联网公司在数据密集型应用上具有其他行业和企业无法比拟的资源优势，其拥有着最精准，最全面的用户数据，特别是这其中还包括了用户身份信息、网络状态、终端信息、业务识别、位置信息、社交关系和消费信用等极具商业价值的数据。因此，将海量数据应用到企业的运营和业务的支撑中将是未来电商互联网公司向数据服务转型和抓住市场主动权的重要手段，而数据源的异构性以及 PB 乃至TB 数量级却对采用传统的企业数据仓库 EDW的多维分析带来了巨大的挑战。

1. 主要目标

通过Hadoop+Hive框架进行电子商务数据的分析，给用户推荐精准的隐式反馈推荐。从用户维度来讲，我们能快速满足用户需求，提高用户忠诚度。从平台角度来看，能给平台带来大量忠实用户，并且提高平台知名度，以及帮助平台实现长久盈利。

1. 架构图



1. 项目进展实施情况
2. 、项目时间节点
3. 项目架构设计图及成员详细分工完：2021.6.9.-2021.6.10
4. 开发软件准备及环境部署及前后端开发：2021.6.11-2021.6.20
5. 前后端代码整合并测试：2021.6.21-2021.6.23
6. 项目完善及发布，总结：2021.6.24-2021.6.24
7. 、开发软件准备及环境部署（2021.6.9.-2021.6.10）

准备软件：python，eclipse

环境部署：伪分布式，Hive，sqoop，flume，mapreduce

数据采集（ETL）

电商日志一般存储在日志服务器，通过 Flume 拉取到 HDFS 上，本文通过 编写python程序模拟日志数据。业务数据通过 Sqoop 从关系型数据库mysql 中读取数据,然后导入到HDFS。

数据清洗

使用 MapReduce 进行数据清洗。

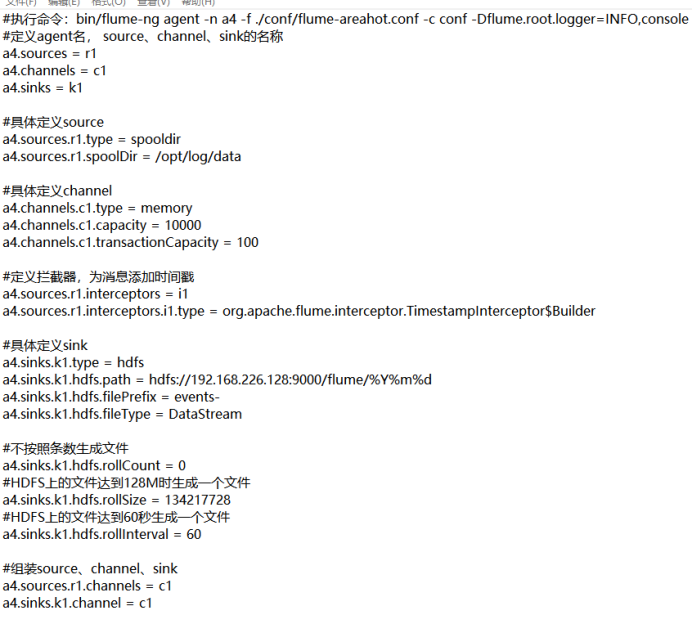
各区域热门商品计算

使用 Hive 进行数据的分析和处理。

前端：

1. 、部分截图
2. 数据采集

使用Flume采集用户点击日志，Flume配置文件（flume-areahot.conf）：



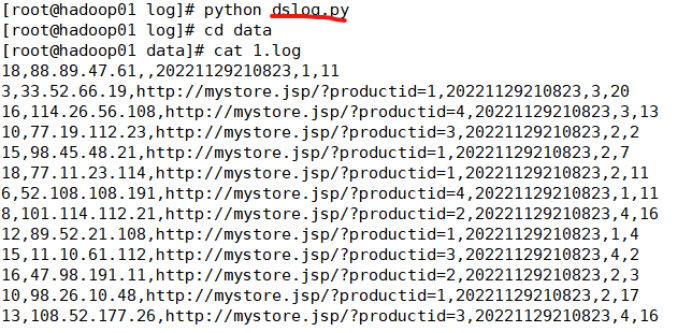
启动 Flume agent，在 Flume 的根目录下执行命令：bin/flume-ng agent -n a4 -f flume-areahot.conf -c conf -Dflume.root.logger=INFO,console。



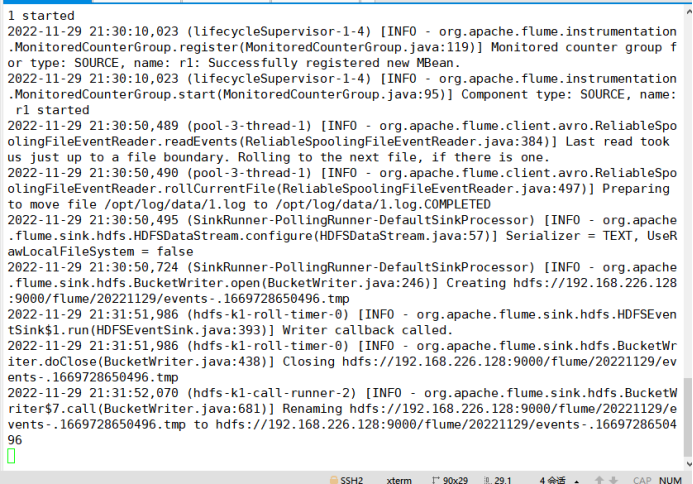
Dslog.py python代码，用于生成模拟用户点击日志文件。



再执行python dslog.py向 /data 目录里放入用户日志文件



Flume会将 /log/data 目录下的文件采集到 hdfs://hadoop01:9000/flume/ 当天日期目录下。

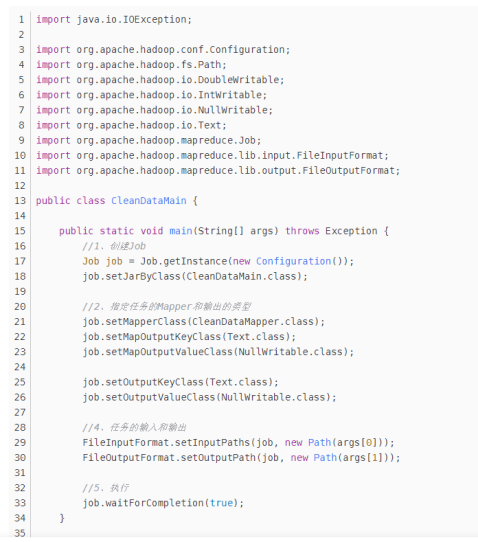


1. 数据清洗（剔除脏数据）

需要将用户点击日志里面对于商品的点击识别出来，过滤不满足6个字段的数据，过滤URL为空的数据，使用 MapReduce 程序进行数据的清洗，过滤出包含http开头的日志记录。

1. CleanDataMain.java及CleanDataMapper.java代码实现：

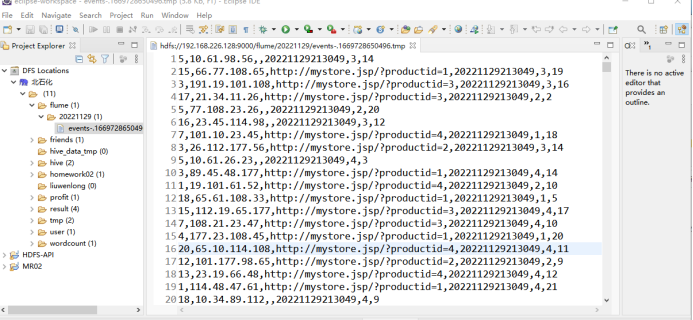
CleanDataMain.java：



CleanDataMapper.java：



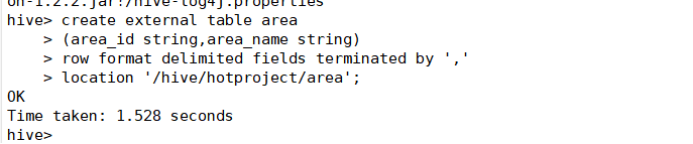
清洗前的数据截图：



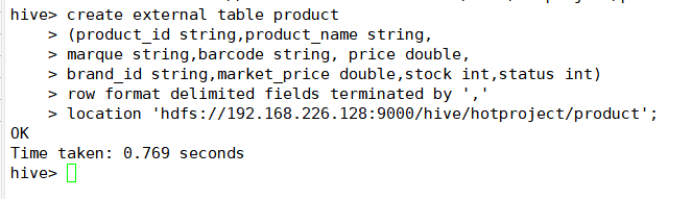
清洗后的数据截图：



1. 基于Hive和Spark SQL的各区域热门商品热度统计
2. 使用hive统计
3. 创建地区表



1. 创建商品表



1. 创建一个临时表，用于保存用户点击的初始日志

create external table clicklogTemp

(user\_id string,user\_ip string,url string,click\_time string,action\_type string,area\_id string)

row format delimited fields terminated by ','

location 'hdfs://192.168.226.128:9000/hive/hotproject/cleandata';

1. 创建用户点击日志表，需要从临时表中解析出product\_id

create external table clicklog

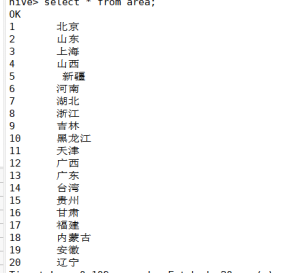
(user\_id string,user\_ip string,product\_id string,click\_time string,action\_type string,area\_id string)

row format delimited fields terminated by ','

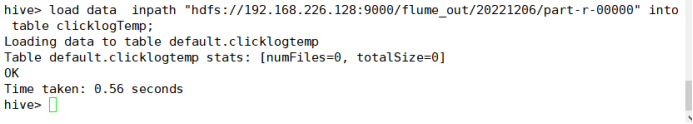
location 'hdfs://192.168.226.128:9000/hive/hotproject/clicklog';

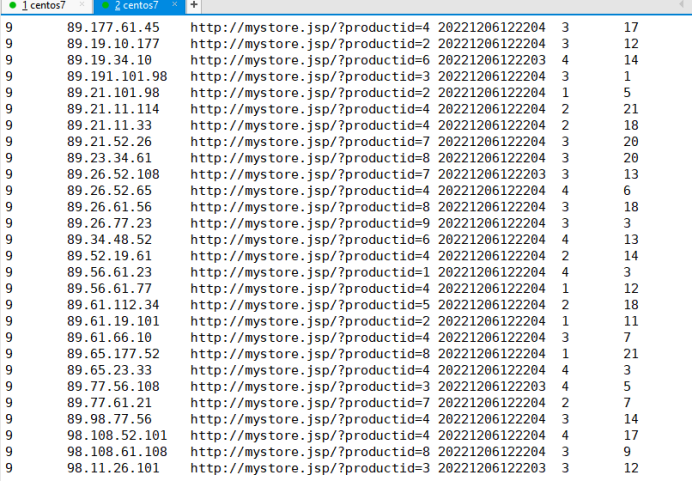
1. 导入日志数据

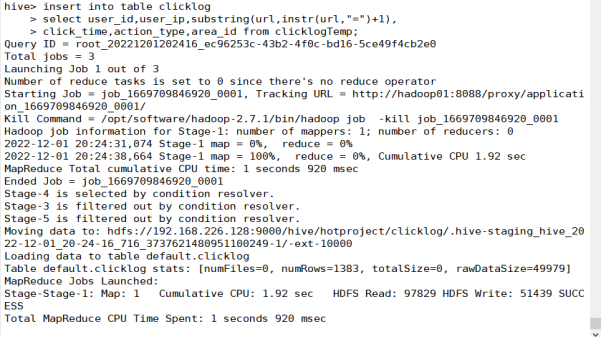
（load data inpath 'hdfs://hadoop01:9000/data/areainfo.txt' into table area）



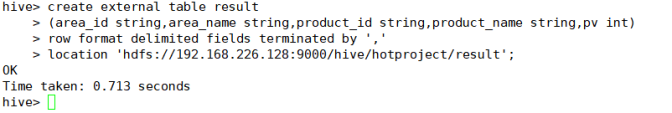
1. 日志通过flume采集到HDFS，日志数据从HDFS导入到hive的clicklogtemp表（临时表）中



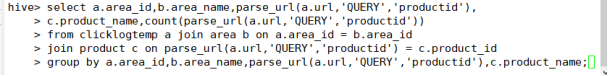




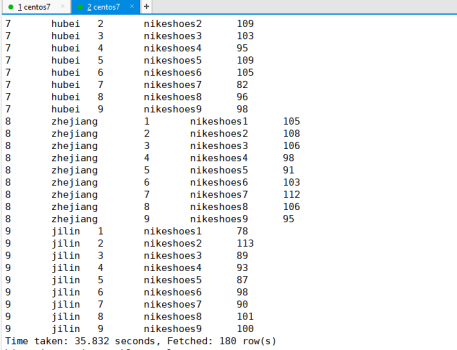
1. 创建结果表



1. 查询各地区的商品热度



查询结果：



1. 将hive分析结果导入结果表

insert into table result

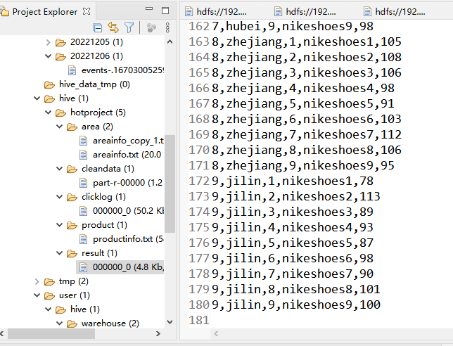
selecta.area\_id,b.area\_name,parse\_url(a.url,'QUERY','productid'),c.product\_name,count(parse\_url(a.url,'QUERY','productid'))

from clicklogtemp a join area b on a.area\_id = b.area\_id

join product c on parse\_url(a.url,'QUERY','productid') = c.product\_id

group by a.area\_id,b.area\_name,parse\_url(a.url,'QUERY','productid'),c.product\_name；

在HDFS中查看result表中内容：

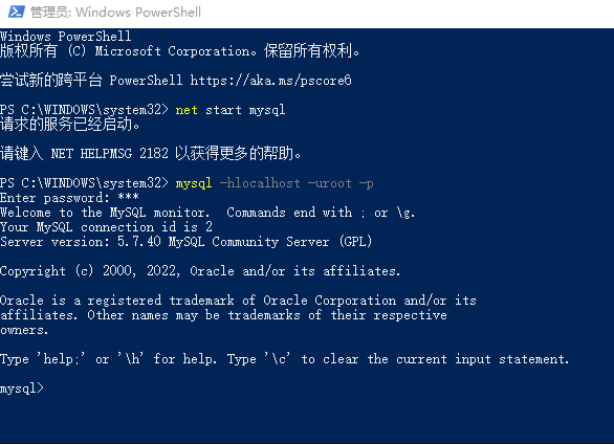


1. 使用Sqoop读入数据

将Hhive中的result表用Sqoop导入到MySQL数据库中

启动MySQL

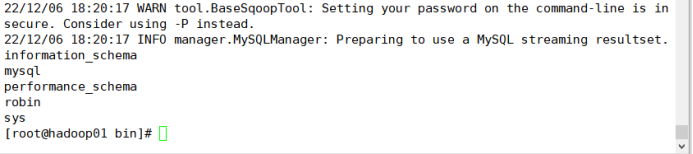
1. net start mysql
2. mysql –hlocalhost –uroot –p，输入密码启动MySQL成功。



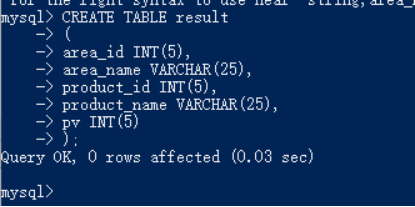
启动和使用Sqoop

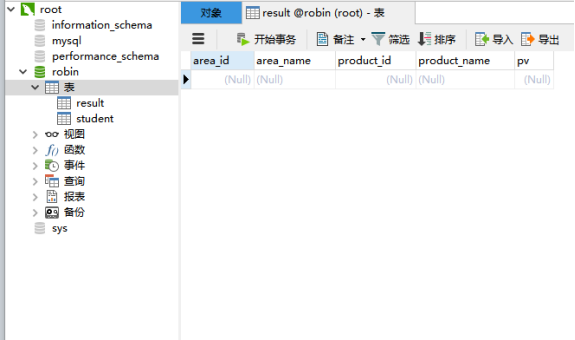
1. 进入sqoop路径的bin目录下，用sqoop查看MySQL上的数据库





1. 在MySQL中创建result表





1. 将HDFS（hive-result）上的数据导出到MySQL中

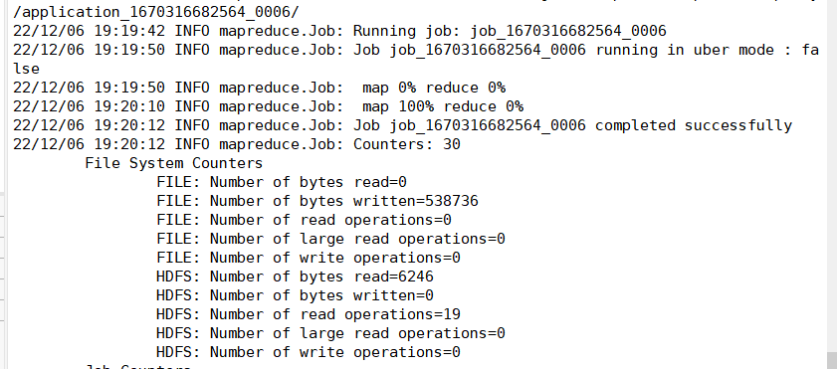
sh sqoop export

--connect “jdbc:mysql://192.168.124.11:3306/robin?characterEncoding=UTF-8”

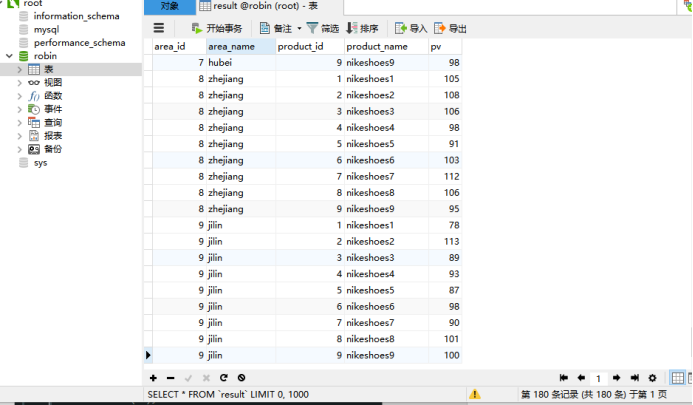
--username root --password 123 --table result

--export-dir 'hdfs://192.168.226.128:9000/hive/hotproject/result/000000\_0'

Sqoop显示导入成功



在MySQL中查看



1. 项目反馈
2. 、问题
3. 因为要访问数据库，所以会对数据库造成很大的压力，而且在真实的生产环境中，一般没有权限直接访问数据库。

2、创建hive表时，没有把表建在自己创建的robin数据库中，结果在sqoop导入时，不能按照hive-sqoop-mysql的导入方式导入，一直出错

3、.当hive表中含有中文，用sqoop把hive表导入到MySQL中时，频繁报错

4、load data inpath "/input/data/areainfo.txt" into table area;导入时 hadoop01和192.168.226.128 会报错。

create external table result

(areaid,areaname,productid,productname,pv)

row format delimited

fields terminated by ','

location '/hive/hotproject/ result';

（二）、解决办法

1、可以把数据导出成csv文件，放到日志服务器上，再通过Flume采集到HDFS上。假如有权限访问数据库，数据库也需要设置成读写分离的模式，来缓解压力。

2、只能使用hdfs-sqoop-mysql的方式导入数据

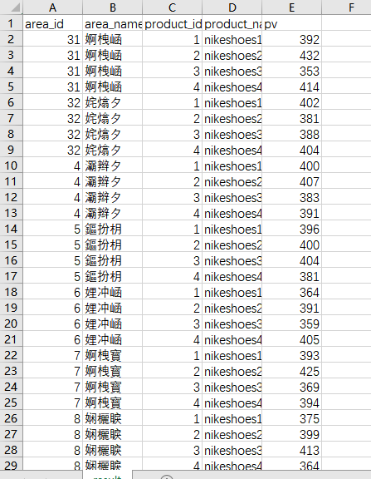
3、hive表的编码和MySQL表的编码不一致，当这两库中编码都提前设定为utf-8时，sqoop可以重新正常导入了。

4、要根据core-site.xml配置选择，一样就能正常导入了

1. 项目成果展示

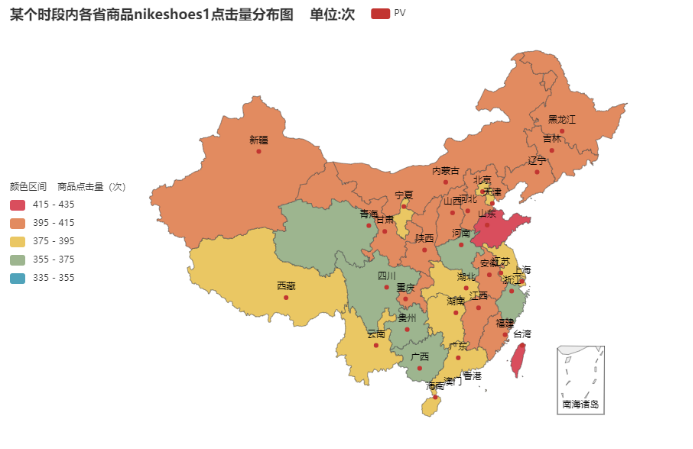
数据可视化：

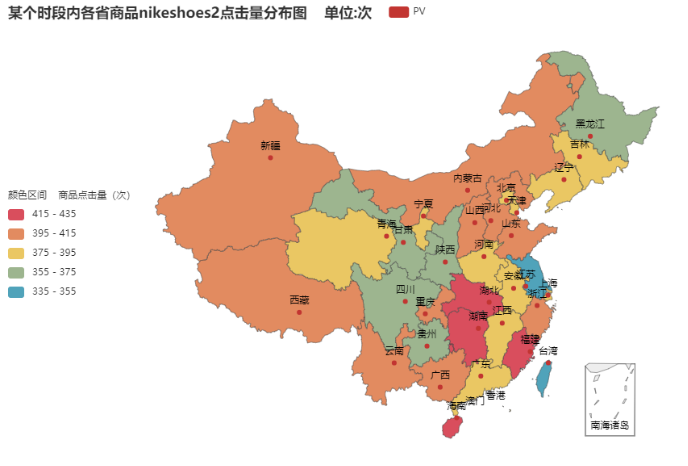
使用Python的pyecharts库进行数据可视化，从MySQL导出csv数据文件。

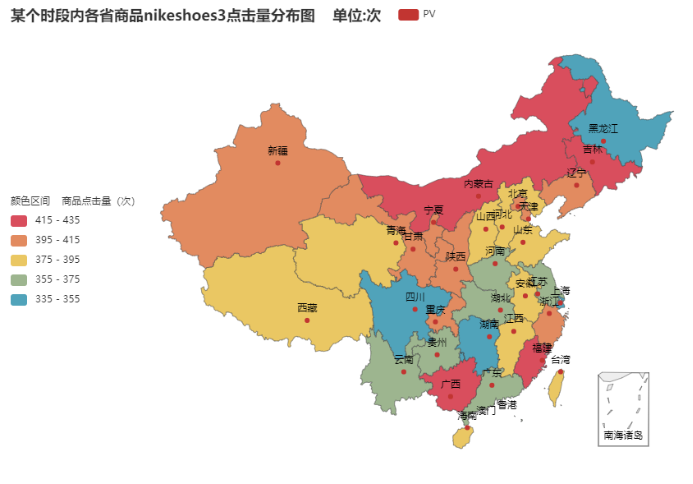


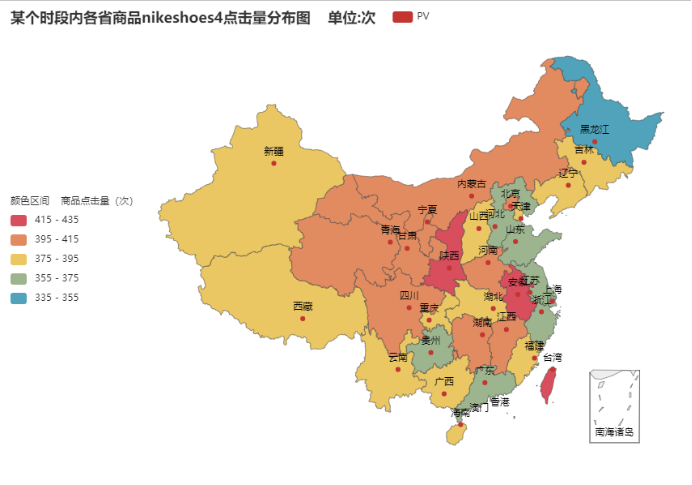
地图可视化：



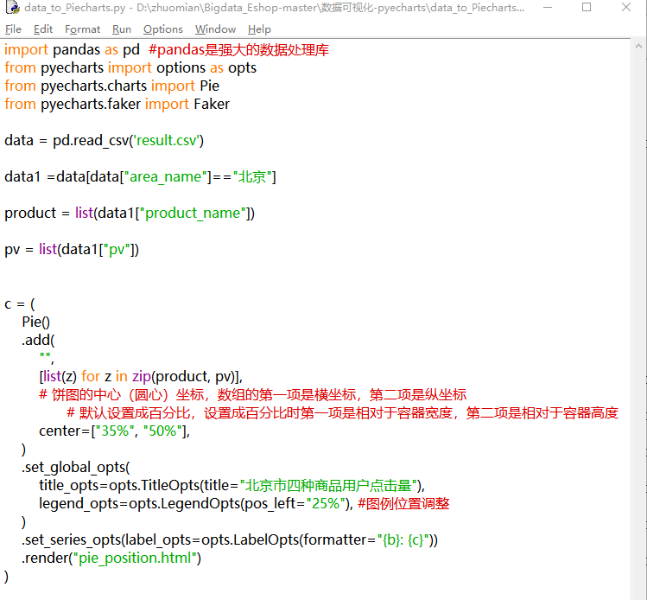


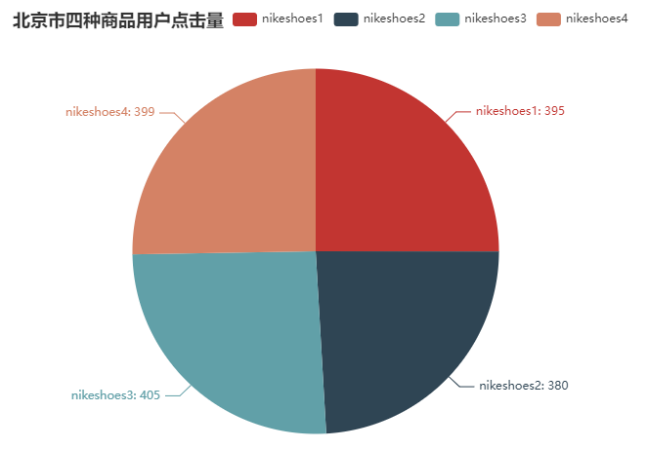






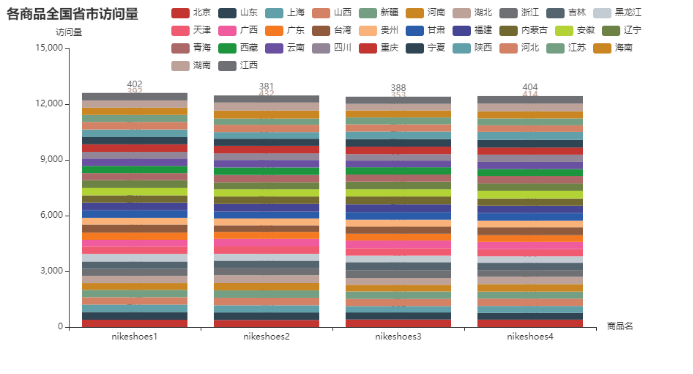
饼图可视化：





柱状图可视化：





1. 项目反思
2. 、总结

通过此次项目我们对电商数据进行分析，并创建了基于Hive的电商数据分析平台在这竞争激烈的互联网电商平台背景下,电商企业如何抓住用户流量,提高用户粘性,增加自身的GMV。并能通过各种指标的分析为企业运营决策提供帮助。

1. 、感悟

通过本次项目实训，我们对python，eclipse等软件进行了深一步的了解与巩固，且可以应用其做项目。熟悉分布式：Hadoop、Hive，sqoop，flume，mapreduce等环境部署以及对数据库sql语句的应用。

1. 、反思

本次项目组员之间分工明确，每个人都完成了各自的任务，虽然在各个阶段都遇到了困难，但是通过老师的指导和组员的努力所有问题都迎刃而解。