大数据项目工程实践项目

班级：22 数据科学与大数据技术 1、2、3 班
课程名：大数据项目工程实践
题目：大数据项目工程实践项目
项目名称：XXX 大数据分析

1.项目概述

1.1 项目背景

1.2 项目功能

1.3 关键技术

1.4 运行环境

2.系统架构设计

2.1 系统组成

2.2 系统协作方式

2.3 系统网络拓扑

3.数据采集与数据存储

3.1 数据采集

要求每组在网站上爬取供本组分析使用的原始数据；

(1) 可以用八爪鱼等爬取工具，也可以自己写程序爬取数据；
(2) 禁用公开的数据集；数据量需 1.22 万条以上；小组间数据不能重复；
(3) 在报告中需说明数据来源、数据量、爬取时间、数据各字段等信息，并展示带字段名的至少前 5 条原始数据；

3.2 数据预处理

(1)对数据各字段与记录进行处理，供后面操作使用；
(2)需有数据预处理的过程，并说明预处理后的各字段及展示预处理后的带字段名的至少前 5 条数据；

3.3 数据上传与数据存储

(1)预处理后的数据上传至 HDFS，根据需要可导入 Hive；

3.4 数据库操作

3.4.1 数据导入

（1）要求：数据需导入 Hbase（或 Mongodb）（或 Neo4j 与 redis(可部分数据与字段，导入数据不少于 2000 条)）数据库，并说明数据库与表的结构等；

3.4.2 数据库操作

在程序或界面中，根据用户输入的字段值，使用 Java 或 Python 编程完成：

插入数据、删除数据、修改数据、查询数据（2 种以上不同字段查询或不同字段组合查询）；

（1）插入数据
（2）删除数据
（3）修改数据
（4）查询数据
①按 id 查询
②按 name 查询
③按 age 与 name 查询
…….

要求 1：每个不同操作需含相应的功能菜单，其中查询操作需有至少二个功能菜单；

说明 1：以上**数据爬取、数据预处理、数据上传、数据存储，数据库导入与数据库操作，每一步需有小标题、说明、过程代码、成功图示（重要！）**。

4. 大数据统计数据分析

4.1 Mapreduce 数据分析（3+）；

分析目的、代码、运行情况、结果；等

4.2 Hive 数据分析（4+）；

分析目的、代码、运行情况、结果；等

4.3 Spark SQL 数据分析（4+）；

分析目的、代码、运行情况、结果；等

……

说明 1：分析结果导出到本地或 HDFS 或 Mysql 等，用于后面的可视化；

说明 2：**在报告中对每个统计分析需含：小标题、分析目的、代码、执行成功图示、执行结果图示（重要！）**。

5. 实时数据分析

5.1 Flink 流数据实时分析（2+）：

说明：数据源可选取本地或 Kafka（至少一个）等，如数据不够，可模拟产生一些与你的数据结构相同的数据；

5.2 Spark 流数据实时分析（2+）

说明：使用 Spark streaming 或 Structured Streaming 进行流数据实时分析；

数据源可选取本地或 Kafka（至少一个）等，如数据不够，可模拟产生一些与你的数据结构相同的数据；

5.3 流数据实时分析与前端数据实时可视化（1+）

说明：使用 Spark streaming 或 Flink 进行流数据实时分析；Spark（或 Flink）+Kafka 实时分析与前端实时展示可视化数据（1+）

数据源可选取本地或 Kafka（至少一个）等，如数据不够，可模拟产生一些与你的数据结构相同的数据；

可参考如下流程：

(1) 对原始数据集进行预处理

(2) 将预处理后的数据发送至 Kafka

(3) Spark 从 Kafka 获取数据，实时处理，结果发送至 Kafka

(4) Flask 构建的 Web 程序从 Kafka 获取处理后的数据

(5) Flask-SockerIO 实时推送数据至客户端

(6) 客户端 Socket.io.js 实时获取数据

(7) 客户端 Highcharts.js 实时展示数据

说明：**在报告中对每个步骤需含：小标题、功能、代码、执行成功图示、执行结果图示（重要！）**。

6. 高级数据分析

6.1 分类

6.2 聚类

6.3 回归

6.4 推荐

……

说明：利用 Spark MLlib 进行机器学习：

(1) 对你的数据，可选取分类、回归、聚类、推荐等机器学习算法(至少选取二个算法，2+)进行高级数据分析；

(2) 要求模型完整（含小标题，功能，算法，分析过程，模型代码，运行过程与结果，模型评估,结论等）。

7. 数据可视化

7.1 开发环境

7.1.1 FastAPI 开发环境（或 Flask 开发环境）

7.1.2 本部分功能

7.2 数据可视化 1

7.2.1 功能说明

7.2.2 项目结构

7.2.3 HTML 模板

7.2.4 Python 后端代码

7.2.5 运行与页面效果展示

7.3 数据可视化 2

7.3.1 功能说明

7.3.2 运行与页面效果展示

7.4 数据可视化 3

7.4.1 功能说明

7.4.2 运行与页面效果展示

7.5 数据可视化 4

7.5.1 功能说明

7.5.2 运行与页面效果展示

……

要求：

要求 1：可选择 FastAPI 与 ECharts 数据可视化，或是 Flask 与 ECharts 数据可视化

要求 2：前面数据分析的结果要求导入到本地或 mysql, 用 FastAPI 或 Flask 从本地文本文件或 mysql 读取数据，并在前端使用 ECharts 实现数据可视化；

要求 3：数据分析的结果需按上面要求 Web 可视化（最少 4 个）.

要求 4：其中 7.2 按要求给出完整说明与完整代码，其它按要求即可

说明：在报告中对每个可视化需含：小标题、可视化说明、代码、执行结果图示（重要！）。

8. 总结

说明本项目完成的主要功能、体会等；

9.其它

(1) 报告要求

报告内容：每个步骤需有清晰过程，展现完整处理过程与结论；**含小标题、代码、运行成功图示、成功完成结果图示等（重要！）**。

报告格式：（含封面、目录、内容、总结）

分工：需在总结部分的开始处，写明小组中每位同学的分工，

格式为：如：张三，负责：4.1 Mapreduce 数据分析，6.2 聚类，7. 数据可视化

封面：

课程名：大数据项目工程实践

题目：大数据项目工程实践项目

项目名称：XXX 大数据分析

班级学号姓名
(2) 程序要求
程序部分是报告的支撑材料，需把所有代码与所用命令上交；
说明：如有爬虫程序需上交代码文件；
JAVA 程序：上交工程文件；
Python 程序：上交工程文件或程序文件；
Spark 程序：上交代码文件，或把所用命令对应报告中标题号，汇总到一个文本文件上交；
Flink 程序：上交代码文件，或把所用命令对应报告中标题号，汇总到一个文本文件上交；
Hive 统计分析：把所用命令编号对应报告中标题号，汇总到一个文本文件上交；
可视化：上交代码文件；
其它：上交代码文件，或把所用命令对应报告中标题号，汇总到一个文本文件上交。
(3) 数据集要求
数据集部分是报告的支撑材料，上交原数据集与预处理后的数据集；
(4) 上交要求
上交内容：程序+数据+报告(pdf 格式)；含“程序+数据+报告”电子版与报告打印版；
班级需在规定时间，统一上交；在规定时间未上交的同学视为缺考；
(5)难度要求
要求程序正确，报告规范；完成以上任务，为完成本项目的基本要求；增加实时分析与机器学习等功能，可增加得分，为较好完成项目工作。
(6)进度要求
要求：本项目共 8 周时间完成（第九周至第十六周）：
第九周确定小组成员（1-3 人）与功能需求（第九周周末上交功能需求报告）；
第十周至第十四周项目设计；
第十四周完成项目完整设计（从第十一周开始检查进度与设计问题，第十四周周末每组演示项目功能）；
第十五周完善项目功能（第十五周周末每组演示项目功能）；
第十六周撰写报告；
第十六周周末每小组演示答辩（初定）。

附录
参考

第 12 章

医药大数据案例分析

本章通过医药电商大数据平台的分析和开发，介绍如何开发一个基于 Hadoop 平台的大数据系统以及大数据的可视化问题，让读者在实践中学习大数据相关技术，掌握大数据相关技术和大数据系统的开发原理，从而利用现有大数据技术解决现实中相关的问题。

12.1 项目概述

近几年，电子商务的崛起给零售行业造成的竞争压力越来越大，一方面是由于移动信息技术日渐成熟、物流快递行业迅速发展以及政策的支持；另一方面在于“互联网+”存在的客观竞争优势，低成本运营、信息化应用带来的较高运营效率和创新模式带来的短期利好等因素；最后，长久以来，传统零售业习惯了粗犷经营，加上医药零售业半封闭的政策环境，导致了传统的销售模式和电商企业竞争时劣势更加明显。

医药电子商务是以医药企业、医疗机构、支付机构、医药信息服务提供商等为网络成员，通过互联网技术，为用户提供安全、可靠、开放并易于维护的医药电子商务平台。随着大数据技术的发展，硬件设备不断升级，计算能力不断增强，大数据技术逐步被引入各个行业。在医药产品销售过程中，通过利用大数据技术在海量数据计算、统计、分析等方面的优势，开展精准化营销和实现线上商品优化，例如基于顾客行为和消费习惯，开展个性化营销，提高成交率；根据医药产品销售数据和库存数据来优化医药商品，提高销售率，降低过期损耗，优化商品组合，充分挖掘数据的价值，提高企业的综合竞争力。因此，通过建立医药电商大数据分析平台采集医药电商平台数据、分析电商平台数据、可视化电商平台数据很有必要。

12.2 功能需求

为了让读者了解该医药电商大数据分析平台，下面介绍该医药电商大数据分析平台的功能需求，后面章节将针对部分功能设计、开发进行详细介绍。

（1）流量分析。按照每日、月度、年度分析用户的行为数据，如浏览量、访客数、访问次数、平均访问深度等。

（2）经营状况分析。按照月度或年度对销售状况进行统计，统计指标包括下单金额、下单客户数、下单量、下单商品件数、客单价。

（3）大数据可视化系统。所有的分析结果最终通过大数据可视化系统进行展示，整个大数据分析系统建立在 Hadoop 之上，用户可以直接通过可视化的界面查询分布式数据库 HBase 中的数据，并进行展示。

12.3 软件关键技术

医药电商大数据分析平台的关键技术有以下几种。

（1）Hadoop 作为分布式计算平台。

（2）HBase 作为分布式数据存储数据库。

（3）Bootstrap 作为页面搭建框架。

（4）jQuery 进行后台交互操作。

（5）EChart 实现数据可视化。

12.4 效果展示

前端系统设计完成后，系统的效果展示如图 12-1 和图 12-2 所示。

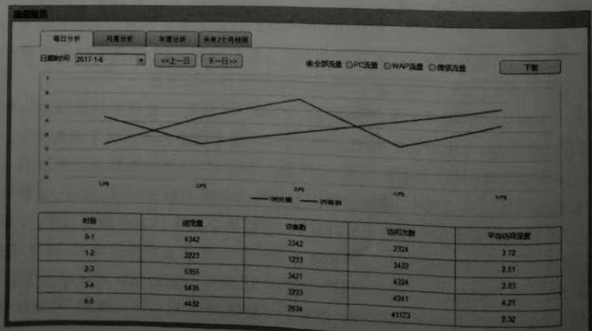

图 12-1 流量概览每日分析结果

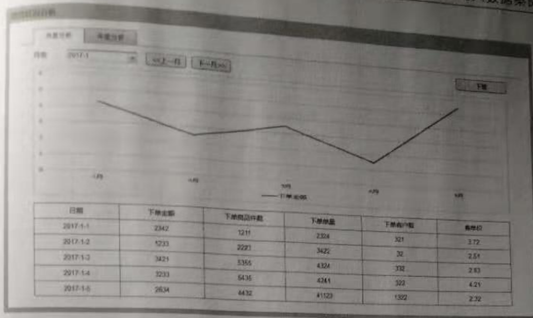

图 12-2　经营状况月度分析结果

12.5　系统构架设计

12.5.1　系统组成

医药大数据分析平台方案主要分为 3 个部分：大数据采集子系统、大数据统计分析子系统和大数据报表呈现子系统，如表 12-1 所示。

表 12-1　系统组成

子　系　统	系 统 定 义	交 互 接 口
大数据采集子系统	系统以离线批处理方式，推送采集结果数据给大数据分析平台	(1) 采集大数据接收的格式 (2) 大数据接口定义
大数据统计分析子系统	具有接收采集系统的数据、客户行为分析、不同药品的精准预测算法、药品推荐算法等特色功能。生成分析结果数据	(1) 大数据的存储 (2) 客户行为模型 (3) 流量分析模型 (4) 统计分析模型
大数据报表呈现子系统	采用 Web 的方案，进行大数据分析，结果以报表、图表的方式呈现给医药电子商务商家	以交互接口、调用报表数据的方式获取需要的结果

12.5.2　系统协作方式

医药大数据系统的子系统间的协作方式如图 12-3 所示。

12.5.3　系统网络拓扑

大数据分析系统的网络拓扑图如图 12-4 所示。

医药电商系统以批处理方式，推送采集数据给大数据分析平台，存储到 Hadoop 服务器集群，大数据报表服务器通过交换机和集群相连。

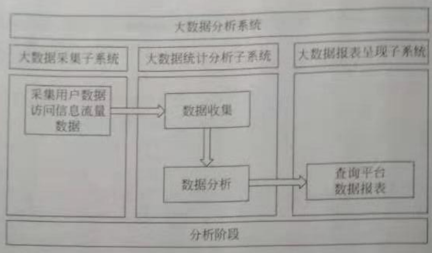

图 12-3 系统协作图

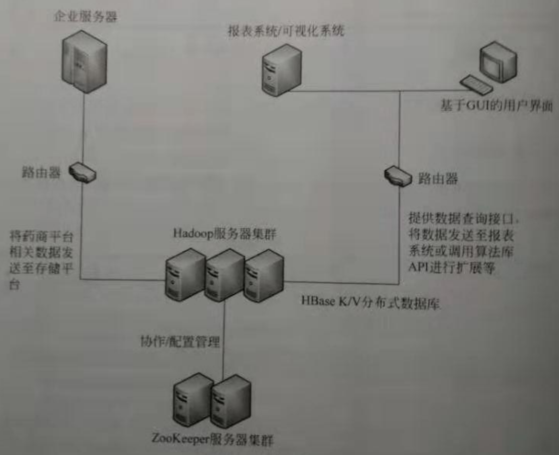

图 12-4 系统网络拓扑图

12.5.4 系统建设方案

1. 数据采集子系统建设方案

1）流量数据建设方案

医药电商平台采用在页脚添加跟踪脚本的方案，能够获取所有访问用户的访问URL、时间、USER AGENT 串，并可根据 cookie 获得用户是否曾经访问过站点，以及记录用户的 ID 来获取是访客还是会员在浏览站点。当用户访问时，PC 平台通过页面跟踪

本将用户的信息异步发送至大数据平台，数据结构如表 12-2 所示。

表 12-2 用户信息数据结构表

字段名	含义	描述
sessionId	会话 ID	一次连续的访问为一个会话，若用户关闭了浏览器重新打开，则为一个新的会话
userId	登录用户的 ID	登录后的用户将具备该信息
trackUid	用户标识	用户标识记录，在机器上永久地记录该标识，只要机器不清理缓存，该标记就永远存在，并在下次访问时发送给服务器。若没有该标识，则表示新用户
userAgent	用户 UA	判断用户属于哪个终端，用于区分 PC/WAP/微信
referer	要访问的页面地址	当访问商品数据时，给出商品页面的访问 URL 规则，包含伪静态的和直接访问的地址。若判断请求在这些地址，则表示在访问商品详细页，并从地址获得商品的 ID

2）订单数据建设方案

医药电商平台在用户生成订单，以及订单进行支付或货到付款订单确认时将对应的数据发送给大数据分析子系统，并在订单实际支付时通过接口通知大数据分析子系统该订单已完成支付，用于分析成交订单和流量之间的关系。订单数据建设方案包含的数据结构如表 12-3 和表 12-4 所示。

表 12-3 订单信息表

字段名	含义	描述
orderId	订单 ID	订单的 ID 编号，数据库序号
userId	下单的用户 ID	
orderNum	订单编号	冗余项，用于核对数据
payment	支付金额	订单打折促销后实际收取的支付金额
productTotalAmount	订单商品金额	订单商品的销售金额
isCod	是否货到付款	T 为货到付款，否则为先款订单
orderStatus	订单状态	0=刚生成订单；1=用户已支付订单；2=用户已确认订单。款到发货订单推送两次的状态序列是[0,1]，货到付款订单推送两次的状态序列是[0,2]

表 12-4 订单项信息表

字段名	含义	描述
orderItemId	订单项 ID	订单项的流水序号
orderId	订单 ID	
productId	商品 ID	在进行 URL 分析时，提取商品 ID 后可以和商品关联
productUnitPrice	商品单价	订单打折促销后实际收取的支付金额
num	商品数量	订单商品的销售金额

2. 大数据统计分析子系统建设方案

集群搭建过程请参考前面章节，此处不再赘述，由于本书应用于教学，因此选用 3 台 PC

服务器，其中包括一台 Namenode+ResourseManager，一台 Datanode+SecondNamenode，一台 Datanode，并可以根据需求动态扩展。

（1）Hadoop 选择。由于可靠性需求和容错性需求，本书选择 Hadoop-2.7.3，zookeeper-3.4.9 和 HBase-1.3.1。

（2）JDK。系统自带的 Java 不需要卸载，使用 jdk-8u131-linux-x64.tar.gz 即可。

3. 大数据报表呈现子系统建设方案

报表数据查询接口使用 webservice 进行数据查询。

12.6 数据存储设计

结合医药电商数据的具体特点和上述的设计及优化策略，为了满足用户进行流量分析、销售分析、药品推荐等需求，从而设计流量数据表、订单数据表、会员评价表，具体内容如表 12-5～表 12-7 所示。

1. 流量数据表

表名：tb_data；

行键：由数据来源的平台类型标识、用户访问时间、用户 ID（注册用户的电话号码，具有很好的离散性）后 4 位组合而成。

列族：名称为 cf，使用单列族设计。

列所包含的信息：会话 ID（sessionId）、用户登录 ID（userId）、用户标识（trackUid）、用户 UA（userAgent）、访问的页面地址（referer）等，还可以根据业务的需要动态扩展。

表 12-5 流量数据表

RowKey	列族 cf				
<platformtype><clicktime><userId>	sessionId	userId	trackUid	userAgent	referer

2. 订单数据表

表名：tb_order；

行键：由订单 ID 和用户 ID 后 4 位组合而成。

列族：名称为 cf，使用单列族设计。

列所包含的信息：订单 ID（orderId）、下单用户 ID（userId）、订单编号（orderNum）、支付总金额（Payment）、订单商品总金额（totalAmount）、支付方式（isCod）和订单状态（orderStatus）等。

表 12-6 订单数据表

RowKey	列族 cf						
<orderId><userId>	orderId	userId	orderNum	Payment	totalAmount	isCod	orderStatus

4. 订单项数据表

表名：[illegible]

行键：由订单ID和用户ID后4位随机数组成。

列族：名称为cf，使用单列族设计。

列名包含的信息：商品ID(productId0)、商品单价(unitPrice0)、商品数据(num0)等。如果订单中包含多个商品，则对应增加相应的列项即可。

表 12-7 订单项数据表

RowKey	列族 cf						
<orderId><userId>	productId0	unitPrice0	num0	productId1	unitPrice1	num1	…

12.7 数据分析

本节描述了利用Eclipse开发工具，分析用户流量数据、用户订单数据以及数据存储实现的详细过程；基于Hbase数据库，对存储流量数据、存储订单数据进行了测试。具体实施步骤如下。

(1) 打开Eclipse，右击Project Explorer项目列表空白处，选择New→Dynamic Web Project选项，这里命名为test01，如图12-5所示。

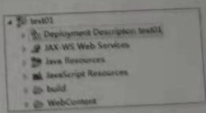

图 12-5 项目 test01

(2) 导入HBase全部的jar包。

(3) 添加本地访问集群所需要的配置文件core-site.xml、hdfs-site.xml及log4j.properties。

① 添加core-site.xml配置文件，代码如下。

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at
    http://www.apache.org/licenses/LICENSE-2.0
  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
```