基于大数据平台的二手车信息采集分析系统设计与实现

杨增春

重庆工商职业学院电子信息工程学院

[摘　要]随着私家车数量的不断增多，闲置的车辆也不断增多，消费者在购买二手车时会参照汽车的各个参数进行比对，这时如果汽车的数据杂乱无章地堆积在一起，会极大降低成功交易率，此时二手车的数据分析工作显得尤为重要。本系统主要将Hive数据仓库和数据展示功能结合起来，Hive数据仓库可以满足对海量二手车数据的存储需求，同时具有可扩展性强、离线分析的特点，数据展示主要从汽车价格、里程、类型、折旧程度等方面进行分析，为商家和个人提供准确、客观的数据结果展示。

[关键词]二手车；数据采集；大数据

【DOI】10.12252/j.issn.2096-627X.2021.12.1622

一、引言

二手车信息采集分析系统能够从实际出发解决现有市场上存在的许多交易痛点，站在商家的角度，消费者对二手车的接受程度远远低于市场数量，导致汽车难出手；站在消费者的角度，二手车交易的情况难以把控，尤其是信息不对称，消费者无法掌握二手车的真实车况，多数时候在购买时更像是一场博弈，消费者也因此对二手车的期待值降低。

本文主要使用Spark对真实数据对汽车情况和汽车价值进行客观分析，继而将分析后的数据在Echart页面上清晰展示出来，为消费者展示出客观真实的汽车数据，能够快速将有价值的信息呈现在消费者面前，帮助他们做出最适合自己的决定。

二、二手车大数据需求分析

经过调查，考虑二手车行业中的商业数据特点，本项目系统的主要功能模块包括：数据采集、数据清洗、数据存储、数据分析、数据展示。这五个模块是互相关联的。数据采集是整个项目系统的基础，采集到的汽车数据会传给下一步做数据清洗，清洗完后的数据会进行会存储到Hive数据仓库中，使用Spark sql大数据计算框架对在数据仓库中的数据进行统计分析。最后通过echart组件对统计分析结果进行展示。

三、系统设计

（一）系统总体架构

系统总体架构主要一下功能：

1.数据采集层，使用爬虫框架Scrapy采集特定站点的二手车数据信息，并将爬取到的数据保存下来，之后导出数据做数据预处理和数据分析工作。

2.数据存取层，从网站上获取到的原始数据经过数据预处理后，导入到Hive数据仓库中。

3.功能实现层，提供对数据的统计分析和展示功能。

（二）系统功能设计

1.数据采集

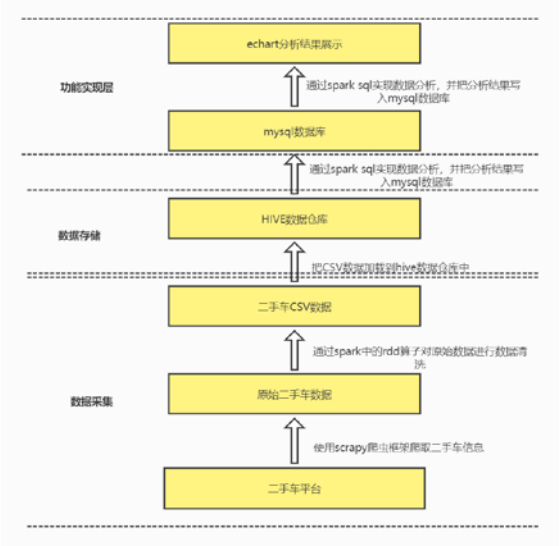

图1 系统架构图

数据采集模块承担了整个项目系统的数据收集工作，后续的所有工作都是在此基础上开展的，在采集工作开展前需要明确采集源、采集方法和采集目的。

2.数据清洗

数据清洗功能主要是对车源的各个参数中的异常数据进行处理，针对人们购买汽车的主要参数进行分类，包括汽车型号、自动挡、手动挡、驱动方式、排放标准、汽车年份等。

3.数据存储

本系统使用Hive存储数据，HIVE数据仓库，能使用 SQL 读取、写入和管理存在于分布式存储架构上的大数据集,结构可以映射到已经存储的数据上,用户连接 Hive 可以使用命令行工具和 JDBC 驱动。

Hive的数据是存在HDFS上的，可以通过MapReduce进行计算，并且提交在yarn上运行的，当他的计算能力不够时，可以通过横向扩机器解决数据暴增的问题。Hive的数据存储在HDFS上的。Hive底层使用的MapReduce执行引擎的，老版本1.x默认支持MapReduce，新版本2.x默认支持Spark。但是可以通过参数设置底层执行引擎是MapReduce或者Spark。这个

对用户的不感知的。MapReduce适合离线处理，他的执行效率不是非常高。Hive和MySQL的对比，如表1所示。

表1 Hive与传统数据库对比

对比项	MySql	Hive
查询语言	SQL	Hive Ql
数据存储位置	块设备、本地文件系统	HDFS
数据格式	系统决定	用户自定义
数据更新	支持	不支持
事务	支持	不支持
执行延迟	低	高
可扩展性	低	高
多表插入	不支持	支持
数据规模	小	大

所以，Hive特别适合在本系统中存储结构化的大数据。

4.数据分析

本系统使用Spark SQL进行数据分析，它主要用于结构化数据处理和对Spark数据执行类SQL的查询。Spark给Hadoop插上一对翅膀，让Hadoop跑地更快。

主要有以下特点：

提供标准化的SQL支持和子查询支持

支持多种数据源： Hive、 RDD、Parquet、 JSON、JDBC等。

通过对Spark SQL的了解，所以本项目采用它对Hive数据仓库中的数据进行数据分析。

5.数据展示

在系统展示层，本系统使用开源框架Echart进行数据展示，该框架的主要特点如下：

丰富的可视化类型：提供了常规的柱状图、饼图、折线图，用于统计的盒形图。

无需编程：不需要复杂的编程，只需配置简单的json语句。

资源多，免费：Echart资源特别丰富，同时免费。

四、系统实现

（一）数据采集

本系统的数据获取以“二手车之家”网站为爬取目标，爬取二手车的基本信息，数据总计5000条，数据类型包括整数、小数、中英文字符、地址链接等。爬取的字段有汽车名、里程数、原价、折扣价、变速箱、上牌时间、排量、排放标准、车身颜色、车型、年检到期时间、保险到期时间、燃油编号、发动机、车辆所在地、驱动方式、链接地址、发布时间、过户次数。

1.数据清洗

处理采集好的异常数据，比如空值、异常值、特殊字符串等。

并把数据转CSV格式。

2.数据加载

把已经出来好的CSV格式的数据，通过hive命令加载到hive数据仓库中。命令格式如：load data local inpath ’/opt/datas/car.txt’ into table stg.car。

3.数据分析

使用Spark Sql框架对hive数据仓库中的数据进行分析处理。

4.数据展示

通过前端echart框架对分析结果进行展示。

五、系统测试

二手车信息采集分析系统采用软件测试方法，这种测试方法与传统的测试方法基本相同，但也存在一些差异

软件的黑盒测试也被称功能测试，这里主要是对需求阶段提出的功能进行测试，检查各个功能是否都能正常运行，不会出现逻辑错误。在进行测试时，仅对界面的操作进行检查，在这个过程中，无需考虑系统的代码结构，完全模拟用户可能进行的实际操作。

白盒测试又称结构测试，与黑盒测试相反，进行白盒测试时，需要清楚地了解代码，根据代码实现的功能编写测试用例，通常用于检查异常处理分支是否正确，是否能够跳转到指定页面，主要用于软件单元测试中。

通过对系统的功能测试，能够正常实现二手车数据的爬取、清洗、转换、分析和分析结果展示。

参考文献：

[1]徐锐.基于某二手车交易平台的产品个性化推荐方法及其系统[D].杭州:浙江理工大学, 2020.

[2]汪九康.我国汽车二手车销售网上平台发展前景[J].中外企业家,2019,(33):1.

[3]李晓瑜.协同过滤推荐算法研究[J].计算机与数字工程, 2019, 47(9):6.

[4]金晶,怀丽波.基于标签和协同过滤的改进推荐算法研究[J].延边大学学报:自然科学版, 2019, 45(3):7.

[5]吕劲.基于特征优化组合SVM的二手车价格预测研究[D]. 武汉:中南财经政法大学, 2019.