基于 Hadoop 的酒店客户数据统计分析应用研究

朱叶[1]，刘小满[1]

广西安全工程职业技术学院， 广西南宁 530000

摘要：随着 Hadoop 技术的日臻成熟，海量数据的复杂处理逐渐变得短时、高效。Hadoop 技术大大推动了当今社会的发展，给各行各业带来了重大变革。可以说，Hadoop 开启了我们人类科技史的新篇章。本文旨在通过 Hadoop 技术分析酒店客户数据的行为特征，从不同维度揭示了客户的预订特点和消费习惯，为酒店的优化方向提供了数据支持和决策依据。根据本文分析结果，酒店可以实施针对性的干预措施，如调整房型比例，优化服务内容等。

关键词：Hadoop；客户数据；统计分析；行为特征

Application Research of Statistical Analysis on Hotel Customer Data Based on Hadoop

Zhu Ye, Liu Xiaoman

Guangxi Vocational College of Safety Engineering,Nanning Guangxi, 530000

Abstract: With the development of Hadoop,the complex processing of massive data is gradually becoming shorter and more efficient. Hadoop technology has greatly promoted the development of our society, and bringing significant changes to many fields.We can say that Hadoop has opened a new chapter in the history of human technology. This article aims to analyze hotel customer data through Hadoop,revealing the booking characteristics and consumption habits of customers from different dimensions, and providing data support and decision-making basis for the direction of hotel. According to the analysis results of this article, hotels can implement targeted interventions, such as adjusting the proportion of different room types and optimizing the service content.

Key words: Hadoop;Customer Data;Statistical Analysis; Behavior Characteristics

DOI:10.69979/3041-0673.24.2.048

引言：

随着互联网技术的迅猛发展，线上酒店预订作为时下流行的出行规划方式，已经成为人们旅游、出差的必要事项。如何捕捉人们的酒店预订习惯和偏好，以提高酒店预订数量和增加酒店效益，成为酒店行业的焦点。线上酒店预订平台积累了大量的客户行为数据和业务数据，这些数据包含了客户的酒店预订习惯、房型偏好、消费记录等多维度信息。传统的数据分析工具，如 Excel、SPSS、Tableau 的缺点是不能对大量的数据样本进行处理，且效率低下。随着 Hadoop 技术的日臻成熟，我们逐渐过渡到了大数据人工智能时代，海量数据的复杂处理逐渐变的短时、高效。

Hadoop 是一个开源的分布式计算框架，旨在对大规模数据集进行存储和处理[1]，为海量数据的处理与分析提供了新的解决方案。本文旨在探讨利用 Hadoop 技术对客户酒店预订和消费情况进行统计分析，以期为酒店管理的优化指明方向。

本文首先概述了 Hadoop 技术的基本原理及其在休闲餐饮业的应用现状；随后，详细列明了基于 Hadoop 的酒店客户数据统计分析应用研究的技术路线，包括数据采集、存储、处理和分析等关键环节；然后针对每一环节进行详细设计；最后，总结了本研究内容对酒店行业发展的积极意义。本文通过对酒店客户数据的统计分析，旨在回答以下问题：

1）酒店房型比例设置是否合理？

2）客户的年龄、性别、职业分布特征？

3）客户的预订时间、预订时长、预订频率的个性化特征？

4）客户酒店忠诚度与酒店入住体验、消费金额是否存在相关性？

2. Hadoop 技术介绍

Hadoop 技术通过算法设定、程序控制，使机器自动抽取、存储、处理和分析大量数据，并给出统计分析结果。它是一个开源的分布式计算框架，主要用于从海量数据中提取有效数据，然后进行模式识别和统计分析，研究数据规律特征，以此推测人或物的行为特征。它允许客户在跨多个计算节点的集群上分布式地存储和处理数据，从而提供高吞吐量、高扩展性和高可靠性的数据处理能力。

Hadoop 的主要技术组件：HDFS 是 Hadoop 的核心组件之一，它提供了高可靠性的分布式文件存储解决方案，可以实现高效大容量的数据存储和读取；MapReduce 核心编程模型，主要用于针对大规模数据集进行分布式计算，可以通过并行计算实现高效的海量数据处理与分析。Hadoop 能够容忍节点故障，确保数据不丢失，可以轻松地通过增加节点来扩展集群的存储和计算能力，支持多

种数据格式。

3. 大数据技术在酒店行业的应用

近年来，大数据技术在酒店行业的应用探索越来越多。旅游酒店大数据分析平台的设计与实现，迟殿委等，针对青岛市的酒店基础数据和客户评论数据进行了探究[2]；基于 Hadoop 的酒店推荐系统，余华咏等，在 Hadoop 分布式架构基础上，研究酒店客户数据行为特征[3]；大数据视角下星级酒店营销管理模式创新的路径选择，茅矛，提出了大数据视角下星级酒店营销管理模式创新的路径选择[4]；大数据背景下经济型酒店营销策略研究，胡恺文，运用 SWOT 分析法对经济型酒店目前的营销环境进行分析，并提出相关策略[5]。同时，通过阅读大量的参考文献，我们可以看出学者们也在纷纷研究 Hadoop 技术在不同领域的细分应用[6,7,8,9,10]，且应用效果良好。因此，本文拟计划基于 Hadoop 大数据技术对当前我国旅游休闲餐饮行业比较关心的热点问题进行研究——了解客户的酒店预订系统情况，研究客户的酒店预订行为特征，分析客户的酒店预订行为数据和酒店忠诚度的关联。

4. 技术路线

基于 Hadoop 的酒店客户数据统计分析应用研究的整体设计思想：系统从前端埋点获取日志文件，从酒店预订系统电脑端、APP、微信小程序、第三方推广系统获取客户业务数据；然后，用 sqoop 将采集到的业务数据迁移到 HDFS 分布式存储，用 flume-Kafka-flume 为传输介质将采集到的日志文件传输到HDFS分布式存储；对这些采集到的数据处理、分析与数据挖掘；最后，生成各种客户活动数据统计分析图表（柱状图、饼状图、曲线图等）。

本研究所采用的技术路线如下图 1 所示。

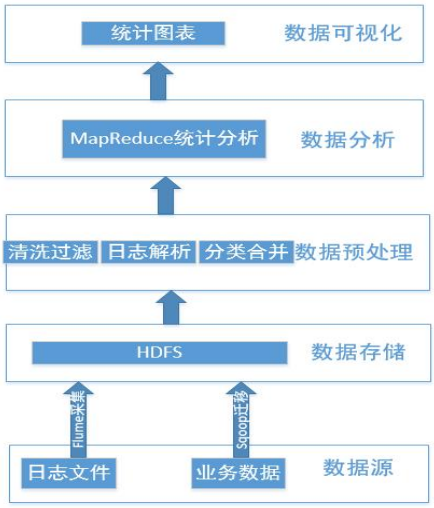

图 1 技术路线

5. 系统设计

5.1 数据采集与存储

本系统中数据采集功能是基础功能，决定了后续数据分析结果的准确性。Hadoop 本身是一个用于存储和处理大规模数据集的分布式系统，并不直接提供埋点数据设置和采集的功能。埋点数据需要在应用程序中预定的位置进行设置，用来追踪客户行为。数据采集过程如下：

（1）数据采集来源：酒店预订系统电脑端、APP、微信小程序、第三方推广系统等。

（2）明确数据需求，编写埋点代码：业务数据，可直接通过 sqoop 传输到 HDFS 存储；日志文件数据，需要在各线上课程资源平台前端预设数据采集埋点，比如客户点击行为、页面浏览时间等。在各线上酒店预订平台中相应的位置（如客户点击事件、页面加载完成等）插入代码，用以收集日志数据。对于 Web 应用，则在前端 JavaScript 中或在后端服务器代码中设置。例如：

```
document.getElementById('button').addEventListener('click', function()
  // 发送埋点数据到后端
  fetch('/log', {
    method: 'POST',
    body: JSON.stringify({ action: 'button_click', timestamp: new Date
    headers: { 'Content-Type': 'application/json' }
  });
});
```

图 2 埋点代码

（3）日志收集：将埋点数据发送到日志收集系统，如 Flume。

（4）数据传输与存储：通过 sqoop 传输业务数据到 HDFS 存储，通过 flume-Kafka-flume 传输埋点数据到 HDFS 存储。通过调整 Sqoop 导入参数，如增加并发数、设置分区方式等，优化数据导入的效率。注意，传输过程中，要确保数据不丢失并能进行缓冲处理。

5.2 数据预处理

对存储的数据进行预处理，如数据清洗、去重、过滤等操作，以减少数据量和提高数据的质量，方便后续进行数据统计分析从而提高数据处理和分析的效率。

主要进行的数据清洗操作如下：

去除重复数据：对于多条重复数据，我们只留取 1 条数据，冗余数据使用 Hive 或 Spark 工具予以去除。例如，多条同一订单编号的酒店订单。

缺失值或异常数据：直接删除。如页面浏览时长缺失、负数的房间预订数量、负数的消费金额等。

数据格式化：统一数据格式，比如日期、数字的格式。对不同平台的日期格式、数字格式进行格式化。

数据过滤：根据特定条件筛选出需要的数据。过滤掉如下数据：那些在极短时间内完成大量客房预订的记录，可能是自动化脚本或其他非正常酒店预订系统行为；那些在极短时间产生大量消费条数的记录和巨额消费的记录；与分析目标无关的列或字段，以简化数据集。

数据整合：将不同来源的数据分类、整合，统一转换为 CSV 格式以便于后续 MapReduce 作业处理。

最后，保留数据预处理日志，以便在必要时能够追溯数据预处理过程。

5.3 统计分析与挖掘

将预处理过的数据导入到不同事实表，这些事实表反应了不同客户的酒店行为特征：时间特征：白天还是夜间，预订的日期和时间（节假日、周末、工作日等），预订的时长（短期住宿、长期住宿）；餐饮消费特征：餐厅选择、食物偏好、饮料选择；客户服务请求：叫醒服务、洗衣服务、房间服务等；与酒店互动特征：关注社交媒体账号、点赞、评论等，住宿喜好：房间位置偏好（高楼层、低楼层、朝向），床型偏好（大床、双床、硬床、软床）；个人信息：年龄、性别、职业等，常住地、旅行目的（商务、休闲、家庭）等；忠诚度：预订次数、房间总数、参与活动情况、积分使用情况。

通过合理地划分任务、减少数据的冗余计算等手段，优化 MapReduce 任务的执行效率和性能。用 MapReduce 对各表数据进行统计，如计算客户的登录终端频次、页面停留时长、消费资源访问次数、房型预订次数、消费品类及次数、消费金额、个人信息特征、服务要求等等。然后，进行描述性分析、聚类分析、相关性分析等，进一步挖掘出客户的酒店预订与消费行为特点。例如：

预订来源分析：统计分析各终端（web 前端、微信小程序、手机）的预订占比，以便了解最有效的获客途径。

客户分布统计：各等级酒店，客户的年龄、性别、职业分布特征；各房型，客户的年龄、性别、职业分布特征。

客户预订频次及房型统计：统计各等级客户的平均预订频次及各房型平均预订次数。

消费品类统计：统计各消费品类的消费次数及总金额，有利于酒店识别最畅销的消费品类和最滞销的消费品类。

个性化推荐：利用协同过滤或基于内容的推荐算法，分析每个客户的酒店预订特点（房型、预订时间与时长、消费品类、服务请求等），为客户提供个性化的酒店资源推荐、房型推荐等。

友情提醒：根据客户的叫醒服务、洗衣服务、房间服务频次特征，结合本地特色文化节等，推测客户的可能日程安排，在微信小程序和 app 端友情推送各种服务和本地文化节的友好提示。

异常检测：识别客户行为异常模式，如只要入住都会要求每天提供叫醒服务或某种服务的客户，突然要求不提供叫醒服务或打扫服务等。这时，需要注意客户是否存在被挟持情况，应采取线上线下相结合措施去确定客户真实原因，必要时报警。

下一次预订预测：综合客户的预订频次、预订时间及房型、要求的服务，预测客户下一次预订时间、房型及可能的服务要求，便于合理规划酒店资源。同时，我们还可以短信、app、小程序推送预订邀请。

调查问卷：统计调查客户的酒店入住体验（优、良、一般、差）、再次入住意愿、良好建议等。

5.4 可视化图表

我们可以使用 Hadoop 生态系统中的可视化工具 Zeppelin 生成各种统计图表(柱状图、饼图、曲线图等)。它提供了多种可视化选项，如图表类型、数据系列和表格设置等，支持多种数据源。通过使用 Zeppelin，我们可以将大量数据转化为直观的图形表示，以便更好地理解和分析。Zeppelin 中能够快速生成美观的数据可视化图形，提供多种视觉元素，如颜色、标签、形状等，可以突出显示重要数据点。因此，Zeppelin 生成的各种统计图表有助于加强酒店对各种客户数据的汇总理解。通过本统计分析研究结果，酒店可以有针对性地完善自己的不足，提高客户体验，从而达到提升营业额的目的。

6. 结语

本文通过Hadoop技术对酒店客户数据从获客来源、房型偏好、客户分布特征、消费品类及金额等多种维度进行了深入分析，深刻揭示了酒店入住客户的多种行为特征，为酒店管理优化方向提供了数据支持和决策依据。依据本研究内容，酒店管理者可以尝试调整酒店房型配比、消费品类供应、服务内容，以提高客户入住体验和增加酒店营业额。

参考文献

[1]赵子晨,杨锋,郭玉辉,陈又新,李钊扬,刘海涛,.基于Hadoop技术的加速器大数据安全存储与高效分析系统设计[J].现代电子技术,2024,(08):9-17.
[2]迟殿委,.旅游酒店大数据分析平台的设计与实现[J].无线互联科技,2022,(07):89-92+98.
[3]余华咏.基于Hadoop的酒店推荐系统[D].南昌大学,2019(02).
[4]茅矛,.大数据视角下星级酒店营销管理模式创新的路径选择[J].企业改革与管理,2023,(23):110-112.
[5]胡恺文,.大数据背景下经济型酒店营销策略研究[J].中国战略新兴产业,2022,(36):92-94.
[6]齐连众,张小凤,.基于Openstack与Hadoop的实验教学大数据系统应用研究[J].现代信息科技,2023,(17):131-135.
[7]汤笛,吴长梦涛,张欣悦,尹茂鹏,张子涵,陈新房,.基于Hadoop平台的灾害大数据处理及可视化[J].电脑与电信,2024,(04):80-84.
[8]张书贵,.基于Hadoop的智慧工作岗位分析大数据平台的设计与实现[J].信息与电脑(理论版),2024,(05):112-114+118.
[9]张鹏飞,江岸,熊念,.Hadoop平台下基于优化X-means算法的大数据聚类研究[J].计算机测量与控制,2023,(12):284-289+309.
[10]李威,邱永峰,.基于Hadoop的电商大数据可视化设计与实现[J].现代信息科技,2023,(17):46-49.

朱叶：（1986.01- ），女，汉族，广西南宁，工程师，硕士。主要研究方向：机器学习与人工智能、数据挖掘与分析。