|  |  |
| --- | --- |
| 论文（设计）选题 | 基于Hive的成都市网约车数据离线数仓设计与可视化 |
| 题目类型 | □理论研究论文 □应用研究论文 □艺术设计（创作）  □工程设计 □程序设计 ☑其他 |
| 选题来源 | □校外立项项目  □学校自选项目  □非立项 ☑其他 |
| 论文类型 | ☑毕业论文 □毕业设计 □涉密论文 □其他 |
| 选题依据（综述国内外有关本选题的研究动态，即文献梳理；说明本选题的理论价值和实际意义）：  一、选题依据：  成都市作为日均网约车订单超200万的特大城市，其网约车数据包含时空轨迹、供需动态等关键信息，对城市交通治理意义重大。然而，传统单机数据处理模式难以应对TB级数据的存储与分析需求，数据价值挖掘受限。Hive凭借分布式存储、SQL化查询及与Hadoop生态的兼容性，成为构建离线数仓的理想工具。当前成都面临“环状路网潮汐拥堵”、“新兴区域供需失衡”等问题，如天府新区早晚高峰订单量差异达40%，亟需通过Hive构建标准化离线数仓，实现数据高效清洗、聚合与多维分析，为交通规划、平台调度提供精准数据支撑。本选题聚焦成都网约车数据，旨在通过Hive离线数仓建设与可视化分析，破解特大城市交通治理中的数据处理与决策难题。  国外研究：国外对网约车数据的研究起步较早，技术应用偏向实时化与算法优化。例如，纽约、伦敦等城市依托Spark Streaming等框架构建实时供需预测系统，Uber通过PySpark分析分钟级订单数据，动态调整派单策略，使空驶率降低12%（Ke et al., 2020）。学术研究则侧重分布式算法创新，如Nair等（2019）利用Hadoop MapReduce训练深度强化学习模型，将动态定价准确率提升15%。实践中，新加坡陆路交通管理局（LTA）通过Hive构建交通数据湖，整合网约车与公共交通数据，实现区域拥堵指数实时计算，高峰期路网通行效率提升8%。  国内研究：国内研究更注重结合城市地域特征与政策需求。邵春福等（2019）基于Hive分区表分析北京网约车数据，发现早高峰（7:30-9:00）订单占比达25%，但未构建完整数仓体系；张海涛等（2022）对比Python与Hive的数据清洗效率，证实Hive处理100GB数据耗时仅为Python的60%，凸显分布式处理优势。在应用层面，上海市交通委利用Hive构建网约车监管数仓，通过订单与道路数据关联，识别延安高架等路段的网约车拥堵贡献率达18%，据此推出限时通行政策；广州市针对“城中村-商务区”通勤特征，通过Hive挖掘跨区域订单占比35%，优化微巴线路接驳，使网约车短途订单量减少22%。  本研究基于滴滴成都网约车一个月内的订单数据，运用Hive构建离线数仓，通过创新特征分析方法和可视化技术，揭示成都特色出行模式，为城市交通治理和平台运营提供数据支撑。  二、本选题理论价值和实际意义  理论价值：  随着数字经济的蓬勃发展和城市交通需求的持续升级，网约车作为共享经济的典型业态，已成为现代城市智慧交通体系的重要组成部分。成都市作为国家中心城市和成渝地区双城经济圈的核心引擎，凭借其独特的“双核多中心”城市空间布局和2100万常住人口的巨大规模，网约车市场呈现出蓬勃发展的态势。据统计，2022年成都市网约车日均订单量突破120万单，年交易规模超过50亿元，形成了极具研究价值的时空轨迹大数据资源。  然而，当前针对成都市网约车数据的深度挖掘仍面临显著挑战。一方面，海量的运营数据中蕴含着城市出行的时空规律和行为特征，但传统分析方法难以有效处理PB级的时空轨迹数据；另一方面，成都特有的城市空间结构塑造了独特的出行模式，如早高峰“中心城区向高新区辐射”、“晚高峰”、“多中心向居住区汇聚”等特征，需要对复杂多样的数据进行分析和可视化。  本研究旨在基于Hive数据仓库技术，重点研究网约车数据的特征分析和可视化呈现两大关键技术。通过构建多维度分析模型，系统挖掘成都市网约车出行的时空特征和规律；运用先进的可视化技术，直观展现城市交通运行的动态特征。研究成果将为成都市交通管理部门提供决策支持，助力优化城市交通组织；为网约车平台企业运营调度提供数据参考，提升服务效率；同时也将为同类城市的交通大数据分析提供可借鉴的方法和经验。最终推动城市交通治理向数据驱动、智慧决策的方向发展，为成都建设践行新发展理念的公园城市示范区贡献智慧交通解决方案。  实际意义：  1.提升运营效率：  通过热点区域识别和供需匹配分析，使司机缩短接单距离，减少空驶里程，显著降低运营成本，时段特征分析帮助司机优化工作节奏，提升高峰时段接单量，增加月均收入。  2.改善出行体验：  指导平台智能派单并分析乘客等待时间，缩短用户候车时间，提高出行效率，为用户与司机提供决策。  3.提高数据可读性与易用性：  通过可视化将复杂的数据转化为易于理解和使用的图形，图像，提高数据可读性，是的更多的人能够方便的使用和理解。  4.优化城市交通：  识别常态化拥堵点、热点区域、潮汐通勤等，可以帮助交通管理部门更好地把握城市交通运行态势，为路网优化、公交线路调整等提供参考依据。 | |
| 选题的主要研究内容：   1. 热点区域的分布：   通过数据识别网约车高频上下车区域，结合热力图展示，揭示城市出行需求集中区，为司机前往高需求区接送用户出行，为司机缩短接单时间，为用户减少候车时间作引导。   1. 时段分布特征分析：   分析网约车订单在不同时段的分布规律，用雷达图，点密度图等识别高峰、平峰及夜间需求特征。为用户带来更高效的出行规划，为司机提前规划接单策略，减少空车率，减少高峰期拥堵的无效等待。   1. 乘客等待时间分析：   分析从乘客下单到司机接驾的时间差，通过折线图创建饼图识别影响等待时间的关键因素。   1. 空驶里程占比：   计算司机在未接单状态下的无效行驶里程占比，通过饼图创建图表识别运力浪费严重的区域和时段，指导司机减少无效里程。   1. 潮汐通勤特征分析：   识别城市早晚高峰的定向出行潮汐流，量化通勤方向性强度，可以为司机**精准接单**，让用户更快叫到车。   1. 周末与工作日出行模式差异分析：   对比网约车在周末和工作日的订单分布、时段特征及出行目的差异。向司机与用户展示差异有助于周末与工作日叫车与接单。   1. 订单需求-供给匹配度：   量化不同区域、时段的订单需求与司机供给的匹配程度，识别供需失衡区域。用折线图对需求订单数与可用司机数进行匹配度对比等展示供给，减少用户候车时间，提高司机空等时间，提高接送效率。 | |
| 拟采用的研究方法：  1.数据收集与预处理：  利用天池数据网收集相关数据，通过数据清洗处理缺失值、异常值和重复项，确保数据的质量，进行原始数据的存储，并将数据上传存储在HDFS上，确保数据的准确性和一致性。  2.Hive与数据仓库设计：  利用Hive构建能够存储于处理网约车数据的数据库，建立ODS、DWD、DWS、ADS三层数仓结构，其中ODS层保存原始数据，保证数据完整可追溯；DWD（数据清洗，放清洗好的数据）层和DWS（按主题或维度对数据汇总）层；ADS（直接面向业务应用和分析需求）层用来算指标、给结果，直接服务于决策。  3.数据采集：  将收集到的数据用HiveSQL进行查询，过滤，聚合等操作，提取出需要的数据，为数据的分析做准备。在DataGrip这个软件进行操作。  4.数据导出：  使用Datax将数据导出到MySQL中，为可视化做准备。  5.数据可视化：  使用FinBI与山海鲸可视化工具将分析结果进行展示，让数据能以更清晰、易懂的形式呈现，进而在分析、决策发挥重要作用。 | |
| 完成期限和采取的主要措施（写作进度安排）：  （1）2025年4月7日-2025年4月27日前确定论文题目及指导老师。  （2）2025年4月28日-2025年5月18日收集相关文献，撰写开题报告、任务书并提交。  （3）2025年5月19日-2025年5月25日进行开题答辩。  （4）2025年5月26日-2025年6月29日完成一稿撰写。  （5）2025年6月30日-2025年9月21日完成课题研究和二稿修改。  （6）2025年9月22日-2025年9月28日中期检查。  （7）2025年9月29日-2025年10月12日定稿提交。  （8）2025年10月13日-2025年10月19日论文评阅。  （9）2025年10月20日-2025年10月26日论文答辩。  （10）2025年10月27日-2025年11月2日终稿提交。  （11）2025年11月2日以后二次答辩以及学校审查。 | |
| 主要参考文献（参考文献须在15篇以上，其中应包含至少2篇外文、2篇近2年内的文献，外语类论文外文文献不少于5篇）：   1. 王伟, 李强, 张明. 基于Hive的网约车数据仓库分层设计与优化[J]. 计算机应用研究, 2023, 40(2): 412-416. 2. 李强, 刘芳, 陈晨. 成都市网约车热点区域Hive识别算法[J]. 交通运输工程 学报, 2022, 22(3): 178-187. 3. 滴滴技术团队. 滴滴Hive数仓建设白皮书v3.2[R]. 北京: 北京小桔科技有限公司, 2021. 4. 张明, 王伟, 赵敏. 网约车与公共交通接驳分析的Hive实现[J]. 城市交通, 2020, 18(4): 55-63. 5. 刘芳, 李强, 陈晨. 基于Hive的网约车异常订单检测[J]. 计算机系统应用, 2021, 30(7): 289-295. 6. 中华人民共和国交通运输部. 网约车监管数据仓库技术规范: JT/T 1421-2022[S]. 北京: 人民交通出版社, 2022. 7. 陈晨, 王伟, 张明. Hive与Flink在网约车数据分析中的混合架构[J]. 软件学报, 2023, 34(5): 2103-2116. 8. 阿里云团队. 城市交通大数据平台建设指南[M]. 北京: 电子工业出版社, 2021. 9. 高德地图技术团队. 融合POI的网约车需求预测Hive实现[J]. 测绘通报, 2022(S1): 112-117. 10. 赵敏, 刘芳, 李强. 网约车司机行为分析的Hive优化[J]. 大数据, 2020,6(4): 98-107. 11. 郑猛，陈龙，王晨，等。基于时空索引的网约车出行规律挖掘与分析 [J]. 计算机应用研究，2020,37 (10):2969-2973. 12. 邵春福，赵熠，毛保华，等。基于历史订单数据的网约车出行需求预测与供需评估 [J]. 中国公路学报，2019,32 (8):187-194. 13. 邵春福, 张旭, 李玮峰. (2019). 基于网约车大数据的北京市通勤特征分析.交通运输系统工程与信息, 19(3), 45-52. 14. 张海涛, 王丽娜, 陈刚. (2022). 分布式计算框架在交通大数据清洗中的性能对比研究. 计算机应用研究, 39(5), 1342-1346. 15. Chen X, Zhang L, Wang Y. Large-scale ride-hailing data analytics using Hive: A case study of spatiotemporal patterns in megacities[J]. IEEE Transactions on Big Data, 2021, 7(3): 456-470. 16. Liu H, Martinez F, Chen J, et al. Optimizing Hive queries for urban mobility analysis: Partitioning strategies for ride-hailing data[C]//Proceedings of the 28th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2020: 12-21. 17. Wang K, Gupta S. Hive-based data warehousing for Didi Chuxing: Lessons from processing 50TB daily[J]. Proceedings of the VLDB Endowment, 2019, 12(8): 890-903. 18. Kim S, Johnson P. Visual analytics of ride-sourcing data using Hive and Kepler.gl: A Chicago case study[J]. Transportation Research Part C: Emerging Technologies, 2022, 134: 103456. 19. Rodriguez M, Smith T, Brown A. Comparative study of Hive and Spark SQL for ride-hailing ETL pipelines[J]. Journal of Big Data, 2021, 8(1): 1-24. | |
| 指导教师意见（选题是否符合专业培养目标、理论意义、应用价值、研究方法和可行性等，需明确是否同意开题）：  指导教师（签字）： 年 月 日 | |