|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | |  | |  |  | | |  |  | |
|  | | |  | |  |  | | |  |  | |
| 毕业论文（设计）题目 | | | 基于Hive的成都市网约车数据离线数仓与可视化 | | | | | | | | |
| 毕业论文（设计）  的主要任务与要求 | | | 任务：  数据获取：运用网络爬虫、数据库连接等技术，从滴滴打车网站数据库、用户行为日志及第三方数据平台采集，也可通过滴滴盖亚数据开放计划  、天池大赛或者Kaggle平台等查找获取数据。预期获取超过上万条数据，且数据要足够全面。  技术栈的调研：使用Hive在于其适合结构化数据的离线分析和长期存储，优势在于稳定性和成本；使用Datax在于其可以支持异构数据源之间的离线数据迁移进行多数据源同步、高性能传输，可以将已经整理好的数据从Hive导出到MySQL中；FineBI则支持多维度数据透视，简单易用。  数据处理：通过数据清洗，去重，异常值过滤等，整合梳理好数据，也可优化添加其他相关因素的数据，为研究做好足够准备。  离线数仓的分层建模：针对于数据建立ODS、DWD、DWS、ADS四层数仓结构，进行原始数据的存储、数据清洗与转换、数据聚合汇总。  数据可视化展示：通过早晚高峰潮汐通勤、热点区域分布、分析空驶率、供需匹配度等多维度进行可视化，将结果以图表形式展现出来。  论文撰写：  绪论（研究背景、现状、目标）  数据处理与数仓构建  数据分析模型构建与应用  数据可视化设计与实现  总结与展望等  要求：  数据方面：获取数据的数据量要足够支撑研究，预期获取超过上万条数据，且数据要足够全面且数据要具有准确性，完整性，一致性；  数仓构建：分层清晰，按照标准的构建方法进行构建；  技术方面：要求多种技术栈进行结合，共同实现研究；  论文方面：符合学术规范，查重率低于 20%，图表编号连续，参考文献不少于 15 篇（含 2 篇外文、2 篇近 2 年文献）。 | | | | | | | | |
| 进度  安排 | 序号 | 工作任务 | | 起止日期 | | | 序号 | 工作任务 | | | 起止日期 |
| 1 | 确定论文题目及指导老师 | | 2025 年 4 月 7 日 - 2025 年 4 月 27 日 | | | 7 | 定稿提交 | | | 2025 年 9 月 29 日 - 2025 年 10 月 12 日 |
| 2 | 收集文献，撰写开题报告、任务书并提交 | | 2025 年 4 月 28 日 - 2025 年 5 月 18 日 | | | 8 | 论文评阅 | | | 2025 年 10 月 13 日 - 2025 年 10 月 19 日 |
| 3 | 开题答辩 | | 2025 年 5 月 19 日 - 2025 年 5 月 25 日 | | | 9 | 论文答辩 | | | 2025 年 10 月 20 日 - 2025 年 10 月 26 日 |
| 4 | 完成一稿撰写 | | 2025 年 5 月 26 日 - 2025 年 6 月 29 日 | | | 10 | 终稿提交 | | | 2025 年 10 月 27 日 - 2025 年 11 月 2 日 |
| 5 | 课题研究与二稿修改 | | 2025 年 6 月 30 日 - 2025 年 9 月 21 日 | | | 11 | 二次答辩及学校审查 | | | 2025 年 11 月 2 日以后 |
| 6 | 中期检查 | | 2025 年 9 月 22 日 - 2025 年 9 月 28 日 | | | 12 |  | | |  |

注：任务下达人为指导教师，任务接受人为学生；本表一式三份，任务接受人、任务下达人及教研室各存一份。