

# 《数据仓库》期末大作业

2024 年 10 月 10 日星期四

## 一、选题

根据兴趣，建立某实际行业(如地产、金融、电商等)的数据仓库(离线或实时)并对之进行分析(尽量采用当下主流技术实现，如开源数仓系统 clickHouse,presto,impala,greenplum 等)，题目自拟。

## 二、要求

1. 分组：每组最多 4 人，组内人员分工需明确；
2. 根据数仓的开发流程(可参考 FTP 上 Kimball 的方法，见其书：The Data Warehouse Toolkit, 3rd Edition-Ralph Kimball.pdf, 以及《大数据之路：阿里巴巴大数据实践》)，完成数仓的需求分析、设计(概念设计、逻辑设计和物理设计)、实现和可视化分析；
3. 数据来源应真实可靠，不允许为大作业而人为构造数据；
4. 每组需使用 PPT 介绍其项目的开发过程并演示结果；
5. 大作业汇报结束后应及时整理相关文档形成大作业报告，在规定的时间内(会有通知)内上传到教学平台。

## 三、评分标准参考(共 100 分)

1. 内容的完备性和准确性(70%)

- 项目背景介绍
  - 需求分析
  - 逻辑设计
  - 物理设计
  - 完整的 ETL 过程(包括技术、工具)
  - 多维分析及可视化分析结果展示
2. 工作量及难度(15%)
  3. 答辩时的表现(10%)
    - 态度是否端正
    - 逻辑是否清楚
    - 语言是否简练
    - 小组成员之间答辩衔接配合是否无缝(答辩时需要全体成员上台, 提前安排彼此之间的答辩内容)
    - 是否在规定时间内完成答辩
  4. 答辩后所提交材料的完整性(5%)

#### **四、大作业成绩评定**

- 评分以组为单位, 同组所有成员分数相同
- 每组分数将由主讲教师和两位助教在各自评分的基础上计算平均成绩, 该成绩将作为小组的大作业分数。