院系_____专业_____班级_____姓名_____学号_____

……密……封……线……

# 许昌学院 2024-2025 学年第二学期期末考试试题

试题名称：大数据处理全流程实战　　试卷类型：A 卷

适用专业：数据科学与大数据技术 T　　适用年级：　2022 级本科

| 题号 | 一 | 总分 |
|---|---|---|
| 得分 | | |

| 得　分 | |
|---|---|
| 评卷人 | |

## 一、作品题（100 分）

【题目】电商用户行为分析与可视化决策系统

**1、业务场景：**

选择某电商平台的用户行为数据（如点击、加购、下单、支付等行为日志），分析用户行为模式，挖掘潜在商业价值。

**2、项目流程**

- 阶段 1：数据采集与存储
  - ✓ 使用 Flume 从模拟的日志文件或 MySQL 数据库采集数据。
  - ✓ 将原始数据存储至 HDFS，并设计合理的数据目录结构。
  - ✓ 验证 HDFS 的容错机制（如模拟节点故障恢复）。
- 阶段 2：数据预处理
  - ✓ 数据清洗与转换：通过 PySpark 对 HDFS 中的原始数据进行去重、缺失值处理、格式标准化等预处理操作。
  - ✓ 数据聚合与维度设计：构建业务相关的关键维度（如用户行为、商品类别）和指标（如销售额、留存率），生成结构化中间数据集。
  - ✓ Spark SQL 优化：使用 Spark SQL 和 DataFrame API 实现多维数据查询优化，支持后续分析需求。
- 阶段 3：分布式计算与可视化
  - ✓ 基于 Spark Core 完成关键指标计算（如用户留存率、复购率）。
  - ✓ 使用 Fine BI 或 Python（Matplotlib/Seaborn）生成可视化报告，包括：
  - ✓ 用户行为漏斗分析图、热销商品排行榜、用户画像雷达图。
- 阶段 4：综合报告与答辩
  - ✓ 撰写项目技术报告，需包含：

    项目背景、技术选型、流程设计、问题与解决方案、可视化结果分析。
  - ✓ 团队答辩（每组 10 分钟），需展示代码逻辑、可视化成果及团队分工。

**3、考核标准**

| 考核环节 | 具体要求 | 占比 |
|---|---|---|
| 代码与作品 | 完整代码（含数据采集、处理、计算、可视化模块），需注释清晰，运行稳定。 | 25% |

院系_____专业_____班级_____姓名_____学号_____

……密……封……线……

| 考核环节 | 具体要求 | 占比 |
| --- | --- | --- |
| **技术报告（论文）** | 逻辑严谨、结构清晰，包含技术细节与创新点，格式符合学术规范（PDF 提交）。分析结果以可视化图表形式展现。可视化图表需直观、专业，至少包含 3 种图表类型，附带分析结论。 | 40% |
| **答辩表现** | 答辩内容完整，回答问题准确，体现团队协作与项目管理能力。 | 35% |

## 4、数据与工具建议

- 数据来源：
  - ✓ 公开数据集（如阿里天池电商数据集）或自行生成模拟数据。
- 技术栈：
  - ✓ 数据采集：Flume/Sqoop
  - ✓ 存储：HDFS/Hive
  - ✓ 计算：Spark Core/SQL
  - ✓ 可视化：Tableau/FineBI/Python（Matplotlib/Seaborn）