

一周实训：数据分析系统

5天 · 起点镜像 P00 · 每组4—5人

交付一个**能运行、能核对、能复现**的小型系统。

实训成果与系统核对

成果	核对动作
HDFS输入	打开一条原始事件
离线报表	与独立小样本核对
流式统计	新文件更新窗口和告警
网页展示	读取本次计算输出
恢复与复现	原查询重启、干净P00重算

每个成员保留操作日志和一项代码修改。

三个自选主题

题目	数据分类	数值单位
电商销售	books、food、tools	分
校园食堂	rice、noodles、drinks	分
环境监测	zone-A、zone-B、zone-C	摄氏度

同一骨架、同一难度、同一评分标准。

选题与统计问题

主题	要回答的问题
电商销售	哪类订单最多，销售额及平均订单额是多少？
校园食堂	哪个品类消费最多，哪些窗口均额超阈值？
环境监测	哪个区域均温较高，哪些窗口触发温度告警？

选择一个主题，数据字典和页面标题采用相同单位。

数据、计算、状态与展示

层次	位置	产物
输入	HDFS stream目录	JSON事件文件
离线计算	Spark/YARN	batch.json
流式计算	Spark本地模式	stream.json、checkpoint
展示	Streamlit	表格和图表
操作入口	Windows SSH	命令与网页隧道

每组画出上述对象在本机中的实际路径。

系统的数据链路

生成带业务ID和时间的JSON事件

完整文件发布到HDFS

PySpark完成离线分类统计

Structured Streaming更新窗口和告警

Streamlit读取结果并展示

哪些已经提供，哪些由你完成

已提供

软件环境与启动工具

生成器与schema

结果写入与页面骨架

报告、分工与核对模板

学生完成

统计口径、数据核对

离线清洗和汇总

窗口聚合与告警条件

恢复证据和运行README

五天的交付节点

日期	当天必须看到的成果
1	HDFS输入与字段口径
2	可手算核对的离线结果
3	窗口更新与告警
4	展示、异常处理、恢复
5	复现与个人操作记录

五天的角色轮换表

成员	第1天	第2天	第3天	第4天
A	环境	离线	流计算	核验展示
B	离线	流计算	核验展示	环境
C	流计算	核验展示	环境	离线
D	核验展示	环境	离线	流计算

第5人轮换测试复现职责；第5天全员填写个人操作记录。

每人都要实际操作

每天轮换：

环境与数据采集

离线分析

流计算

核验与展示

第5人增加测试与复现职责，也必须修改代码。

第一天的检查清单

起点采用实训镜像P00

`labctl doctor` 确认Hadoop、Spark和Streamlit

填写主题、成员、字段与单位

上传固定seed输入并查看一条记录

每人用SSH完成一次HDFS读取

当天提交：TEAM.md、数据字典与输入目录记录。

第一天：启动并投放数据

```
labctl start practicum  
cd ~/work/practicum  
python generate.py --theme shop --seed 42 \  
    --files 6 --rows 20 --interval 0 \  
    --hdfs /course/practicum/stream
```

选其他题目时替换theme。

一条完整业务事件

```
{ "event_id": "shop-42-0-0",  
  "event_time": "2026-10-27T08:00:00Z",  
  "category": "books", "value": 2500 }
```

上例用于说明字段与单位：2500分 = 25元。

生成器的实际ID格式与值，以生成文件为准。

固定seed与输入重现

参数	作用
theme	选择分类及值域
seed	固定伪随机序列
files	文件数量
rows	每文件记录数
interval	文件发布间隔

默认 $6 \times 20 = 120$ 条。记录全部参数并保留源码版本。

HDFS输入的发布与检查

```
hdfs dfs -ls /course/practicum/stream  
hdfs dfs -cat '/course/practicum/stream/*.json' |  
head -n 3
```

生成器先上传隐藏临时文件，再改名发布

完整JSON文件进入正式输入目录

同seed同目录重复发布会拒绝覆盖

新一轮数据采用新seed或独立输入目录

明确数据接口

字段	含义
event_id	业务事件唯一标识
event_time	带时区的事件时间
category	分类或区域
value	金额分数或温度整数

记录seed、文件数、每文件行数，才能复现输入。

第二天： 离线统计

在batch.py补全：

- 按event_id去重

- 过滤空值或负value

- 按category汇总count、total、avg

结果保存为batch.json， 供给定页面读取。

第二天的离线计算接口

原始事件

按event_id去重

过滤value为NULL或负值

按category分组

计算count、total、avg

写出batch.json

计数单位为有效事件条数，金额主题的total单位为分。

六条小样本的手算

ID	类别	value
e1	A	100
e2	A	300
e3	B	200
e4	B	400
e5	C	500
e6	C	700

A: count=2、total=400、avg=200。

B: 2、600、300; C: 2、1200、600。

异常记录的预期处理

在独立测试目录中加入：

记录变化	离线期望
复制已有event_id且内容相同	去重后总数保持一致
value为NULL	排除该事件
value为-1	按基础规则排除
新ID、合法value	对应类别数量增加1

运行前写出count与total的预测。

用小样本证明结果正确

选至少6条记录，保留输入原文

手算每类数量、总值和均值

用独立Python或SQL对照

再扩大到全部合成数据

手算与独立核对采用原始输入逐项推演。

批流结果的独立核对

在业务ID唯一、清洗口径相同的输入上：

$$\sum_W count_{W,c} = count_c$$

$$\sum_W total_{W,c} = total_c$$

类别均值采用 `total/count` 重新计算。

窗口均值需按count加权，才能合成全局均值。

第三天：窗口与告警

在stream.py补全：

过滤无效value

按10秒事件时间窗口和category分组

统计count、total、avg

比较均值与本题阈值

先 `--once` 核对，再持续等待新文件。

第三天的运行顺序

```
spark-submit stream.py --once  
cat stream.json  
spark-submit stream.py
```

先核对有限输入，再持续等待新文件。

另一个终端使用新seed发布，观察已有事件时间窗口更新。

告警规则的边界测试

规则：窗口均值 `avg > threshold`。

avg	threshold	告警
2999	3000	否
3000	3000	否
3001	3000	是

温度主题采用30℃阈值，比较符号保持一致。

告警阈值与单位

主题	基础阈值
电商/食堂	窗口平均值大于3000分
环境监测	窗口均温大于30℃

把阈值改大之前，先预测告警窗口数怎样变化。

第四天：接通结果展示

```
streamlit run app.py \  
  --server.address 127.0.0.1 --server.port 8501
```

Windows访问SSH隧道: <http://127.0.0.1:18501>

刷新页面读取最新JSON，计算任务在终端运行。

第四天的服务启动与网页连接

服务	启动位置	查看方式
HDFS/YARN	labctl start practicum	19870 / 18088隧道
流查询	项目目录终端	日志与stream.json
Streamlit	项目目录另一个终端	18501隧道

页面显示batch.json、stream.json及对应图表。

页面数据为空的排查顺序

确认 `batch.json` 与 `stream.json` 存在且可解析

查看Spark任务是否完成或持续运行

确认Streamlit在项目目录启动

查看8501监听与SSH隧道

浏览器刷新页面读取最新结果

将计算、文件、服务与网络分别检查。

恢复实验的操作记录

时刻	操作与记录
停止前	保存结果、输入清单和checkpoint路径
停止	Ctrl+C结束流查询
重启	使用相同代码、输入与checkpoint
核对	无新文件时结果保持一致
继续	新seed输入带来预期更新

另做一次虚拟机关机后的服务恢复。

三项必做验证

独立测试NULL、负值和重复event_id

保留checkpoint，停止并重启流查询

虚拟机关机后，按README恢复服务

将命令、输入和关键输出粘贴到报告，按模板插入规定截图。

区分两类重复

同一个已处理文件：进度记录用于避免原样重读。

业务内容相同、文件名不同：仍可能再次计数。

离线按event_id去重；基础流程序按输入记录计数。

系统拓扑与状态策略

单VM多进程，共享物理主机资源

complete模式适用于本次有限样例

长时间运行需要考虑watermark和状态清理

新查询结构采用独立checkpoint

README记录拓扑、状态与重启条件。

README的可执行结构

小节	应包含内容
环境	P00编号、账号、版本
输入	theme、seed、files、rows、目录
运行	按顺序列出批流与页面命令
核对	小样本期望、批流总数与总值
恢复	输入、状态、结果路径与重启顺序
分工	TEAM.md与个人修改位置

干净复现的输入输出清单

复制到新P00：实际运行代码、README和所用小样本。

重新产生：

- 固定seed的输入文件

- HDFS正式输入目录

- batch.json与stream.json

- 本轮查询checkpoint

对比预期结果，并记录新一轮运行日志。

第五天：干净环境复现

从新的P00镜像开始：

- 复制本组代码和README

- 用规定seed重新生成数据

- 执行离线与实时任务

- 核对结果并启动页面

保留本轮输入、命令、日志和重新计算的JSON。

每人填写D5.3个人记录

姓名学号与本人任务

预测、修改前后代码片段

运行命令与实际输出

对本人修改的书面解释

每位成员插入一张本人操作截图。

个人操作记录示例

可选操作	记录内容
改阈值	预测、代码变化、实际告警
查HDFS事件	路径、命令、读取结果
改清洗条件	受影响记录与前后结果
修路径错误	错误、诊断、修复与验证
重启原查询	输入、状态目录与恢复结果

在报告D5.3中记录本人实际完成的一项。

选做扩展的核对方法

扩展	要回答的问题
MapReduce核对	独立类别计数是否与Spark一致?
ZooKeeper协调	当前采集者退出后谁接替?
watermark	给定迟到范围如何影响状态与结果?

基础链路通过后选择一项，保留新增输入和运行证据。

评分与选做

内容	分值
系统功能 / 数据正确性	35 / 20
恢复与复现	15
个人操作与解释	25
文档及AI记录	5

基础通过后，再选MapReduce核对或ZooKeeper协调。

提交Word报告

“源码与运行说明”： batch.py全文、stream.py全文、verify.py或verify.sql全文、app.py修改片段或说明使用给定页面、README运行说明全文

完成各步骤实际结果、规定截图和已有书面栏目

AI使用情况： 简短说明或一张截图； 未使用填“未使用”

保存后重新打开， 核对文本和截图清晰可读

正式提交一份填写完成的Word报告。