**课程名称大数据实训**

**实验5-Lambda 架构和实时处理**

课程名称：大数据实训

实验学时： 40 学时

综合性、设计性实验：□🗹是

学生人数：15

任务编制人：课程负责人 刘喆

# 实验目的

* 1. 了解 lambda 架构存在的意义和解决的问题
  2. 可以用 flume/hadoop/kafka/spark/flink 等技术实现 lambda 架构
  3. 掌握常用的 java 字节码工具 btrace/greys-atom/byteman/jvm-sandbox 可以从源码级定位和解决问题
  4. 可以使用火焰图技术来分析程序的瓶颈在哪里，并优化

# 评分办法

1. 分小组在指定的时间之前完成任务，按任务的完成情况，给小组打分
2. 按每个人对于实验的贡献，包括但不限于 代码贡献/文档贡献/issue 贡献，给每个人打分
3. 每个人本次实验的最终成绩，为上面两个数字的乘积

# 实验要求

- PM

- 各种问题及解决的 issue

- RD

- 完成 spark streaming 的 uv 统计，数据写入hdfs， 项目名 steper

- spark streaming 运行的 stage 截图 stage.png 或 flink 的有 stage 的截图

- 找到 streaming 处理的上限，即每秒最多处理多少条数据，这个上限是如何测试出来的 limit.md

- 想办法把这个上限提高 20%， 或者给出分析，这个事情作不到， improve.md

- 某一个 worker 上处理 uv 时的火焰图 flame.svg

- 对于这个火焰图的分析 ana.md 以及改进意见，改进代码，给出改进后的火焰图 new\_flame.svg

- 把实时 uv 用 websocket 或其它数据可视化技术展示到 web 上, 截图 real.png （可选,加分项）

- 把完整的 lambda 架构组织完整，特别是最终的数据合并部分 lambda.md （可选，加分项）

# 实验环境

1. 硬件环境
2. 软件环境

# 实验内容

1. 自己贡献的源码或文档的 url
2. 自己提的 issue，附 github 连接即可
3. 其它没有记录的问题（应该都在 issue 里了，如果有落下的，写这里）