课程大作业：综合练习（30 分）

一、数据源

- 给定 2 路 TCP 数据流，格式为`<topic, timestamp, int_value>`。

  topic：流标识，在整个流过程中不变。

  timestamp：时间戳，递增，但可能与当前时间不一致。

  int_value：整数型数据（小于等于 2^{10}）

二、流处理

- 实现接收 2 路 TCP 数据流的处理程序（processor），要求输出递增文件（或者数据流），格式为：`<topic1, timestamp1, topic2, timestamp2, int_value>`。

  topic1：数据流 1 的标识

  timestamp1：数据流 1 的数据元组的时间戳

  topic2：数据流 2 的标识

  timestamp2：数据流 2 的数据元组的时间戳

  int_value：当数据流 1 与数据流 2 中时间戳 timestamp1 和 timestamp2 距离不超过 30 秒，且两者的 int_value 相同时输出至递增文件

- 给出 processor 并行到多个服务器的方案，包括：
  - 对数据源调整的需求、流的划分和分布化方法。
  - 并行处理时可能的同步处理。

三、Batch 数据处理

- 利用数据流形成 2 个 batch 文件（每个流每小时形成一个 batch，数据量不得小于 6 小时），格式为`<topic, timestamp, int_value>`序列。

- 处理两个流的所有 batch 文件，输出在两个流中共同出现的所有 int_value。
- 给出以上数据处理过程的分布设计和具体实现。
- 本问题可采用开源代码实现。

四、数据流模拟

- 压缩包中 server.py 与 client_demo.py 可用于模拟生成与接收数据流。
- 使用 server.py 时先将 bind_port 修改为其他端口以免发生冲突。
- 该代码文件可供参考，但不强制使用。

作业提交相关

- 将以上作业中的系统设计、命令截图等写入实验报告，然后连同所有代码文件一同打包成压缩文件，上交至网络学堂。
- 迟交作业一周以内，以 50%比例计分；一周以上不再计分。另一经发现抄袭情况，零分处理。
- 助教联系方式：王子寒（wzh16@mails.tsinghua.edu.cn，微信号：chillyprince）