数据处理思路—论文《MVCom: Scheduling Most Valuable Committees for the Large-Scale Sharded Blockchain》

数据来源：http://xblock.pro/tx/

数据处理思路：此处讲述论文中省略掉的 dataset 生成思路。其他细节请参照论文实验环节。

为了便于分析，比特币部分交易数据集包含了比特币交易数据的三个快照:dataset1_2014_11_1500000、dataset2_2015_6_1500000 和 dataset3_2016_1_1500000。实验中使用的是 dataset3_2016_1_1500000。它包含 1378 个 block 的的数据。每个数有 Block ID，Block hash (identifier in the blockchain)，Creation time of block，Number of transactions。

- 对于 offline 的情况，我们对 1378 条数据进行 t 次洗牌（shuffle），洗牌后的第 t 组数据对应第 t 个 epoch。取前 shard=500，800，1000 个 Block 对应的数据作为输入。
- 每个 shard 到来的时间延迟 Latency 由两部分组成，即 formation Latency + consensus Latency：
    1. committee formation 的 latency 是候选节点成功解决 PoW 难题，加入到分片，最后形成 100 个成员的委员会的时间，其期望为 600s。
    2. consensus latency 为分片内委员会成员参与 PBFT 共识所花的时间，通过计算分片对交易数量进行三阶段投票的时间得到。当分片内委员会成员数量为 100 时，该部分延迟的期望为 54.5s。
- 对于 online 的情况，在得到第 t 个 epoch 的数据后，我们基于 total latency 对分片进行排序，设定 Capacity = 800*shard，当前 i 个分片的交易数量总数超过 Capacity 且分片的数量 i 大于总数量的 50%，即 i > shard*Nmin(Nmin=0.5)，开始执行 SE 算法，当 i= shard*Nmax%（Nmax=0.8）时停止执行 SE 算法。即我们对 i、i+1, …, shard*Nmax% 分别执行一次算法。