CCF BDCI CCF BIG DATA & COMPUTING INTELLIGENCE CONTEST

2020 8th

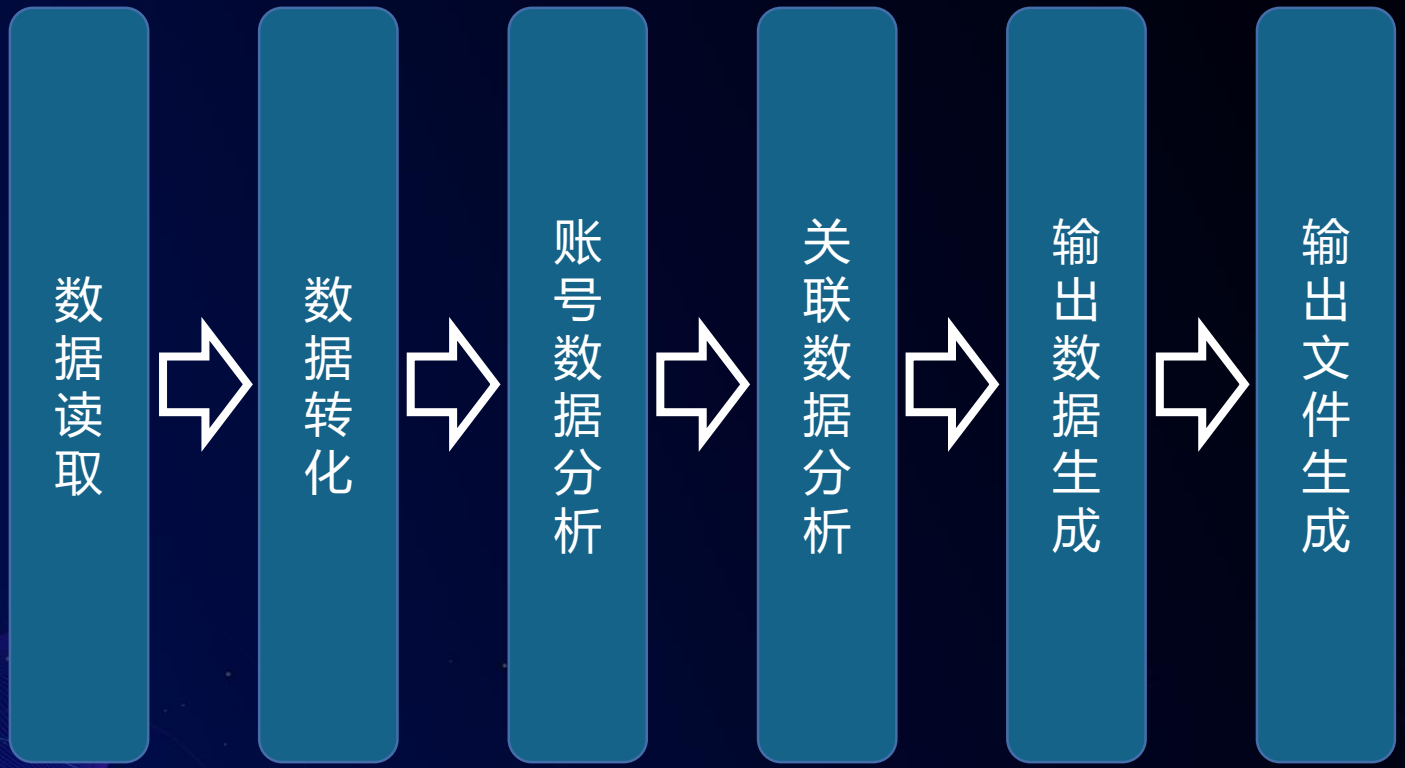
CCF 大数据与计算智能大赛

赛题名：基于大数据的互联网虚拟身份归一处理性能优化

队伍名称：冀数所普拉斯

基本方案（1.0）

程序进程

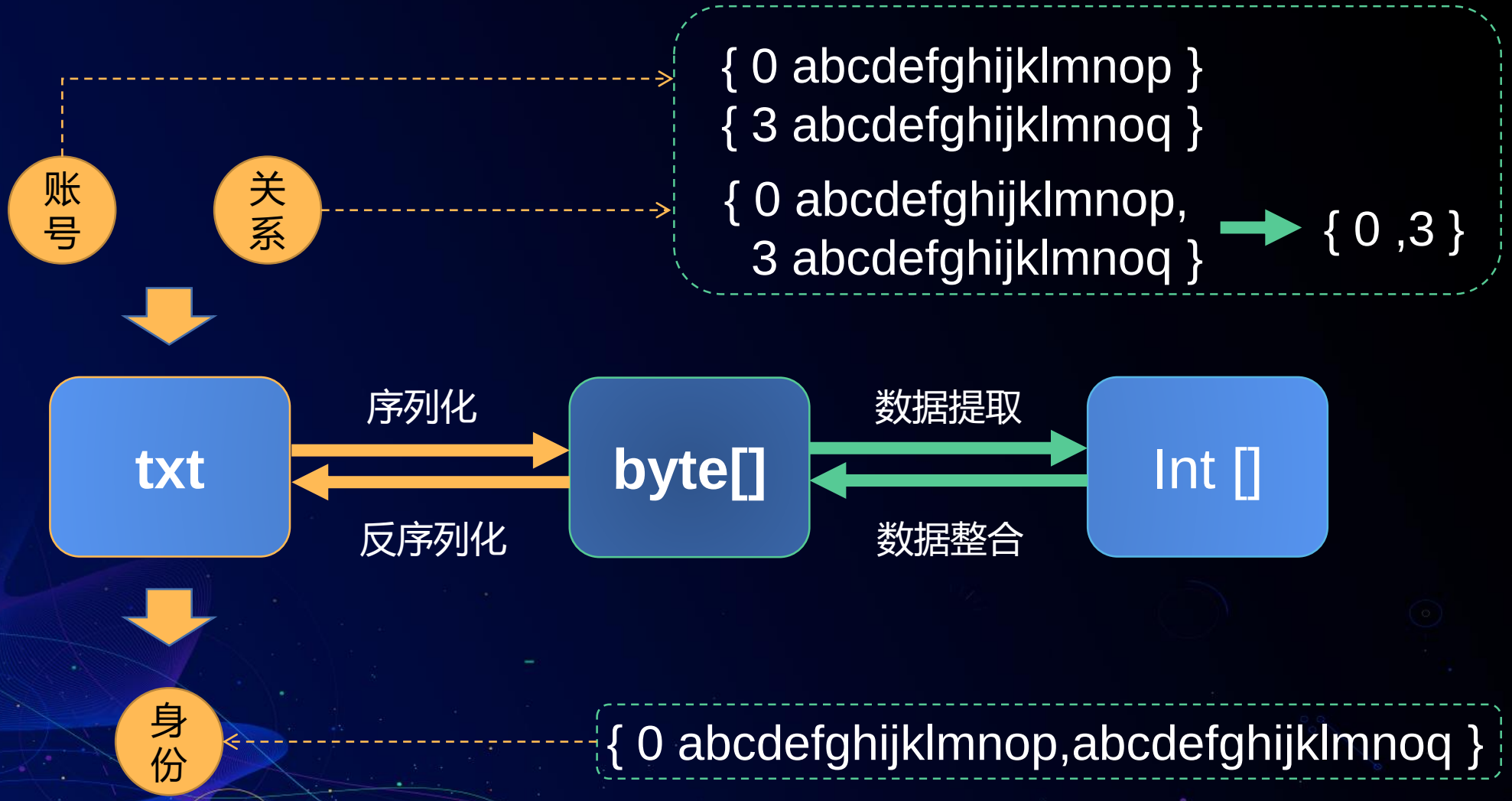


优化策略

数据处理

数据映射：对文本文件进行映射读写

超前预处理：预估账号数量和关系数量



关系计算

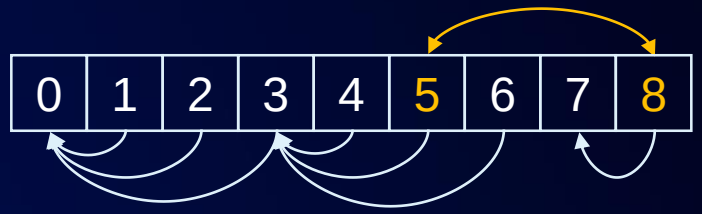
初始化

依赖表
（int[]）

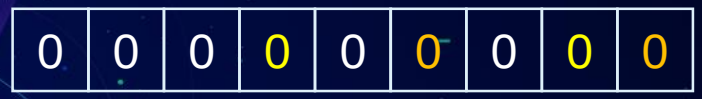
{0,1} {0,2} {0,3} {3,4} {3,5} {3,6} {5,8} {7,8} {8,7}

关联表
（int[]）

关系计算

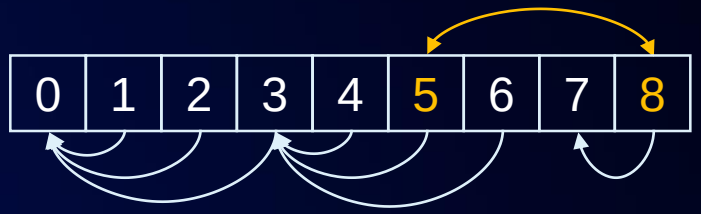


归一化



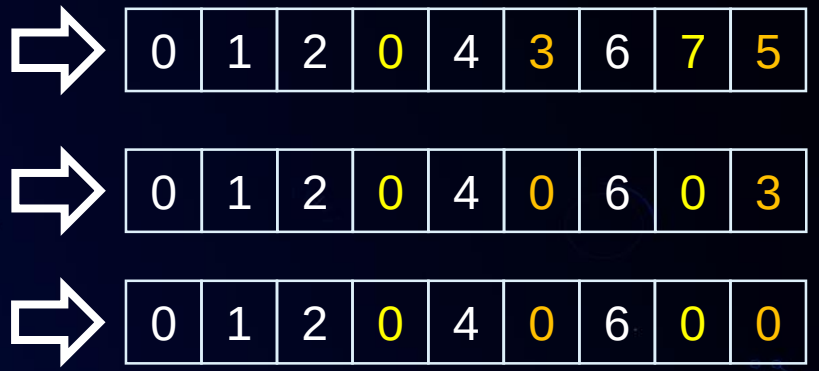
关系计算

初始化



经典并查集

示例 {0,3} {3,5} {5,8}

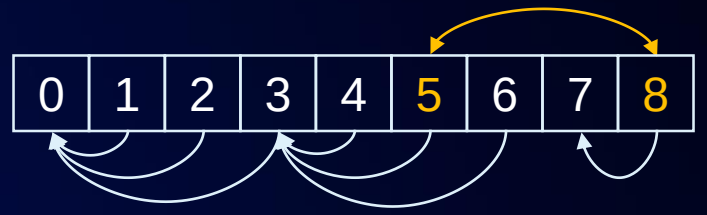


核心：通过递归进行深度并查。

问题：关系越复杂，路径越深，从底部找到根节点会越来越难

关系计算

初始化



局部并查集

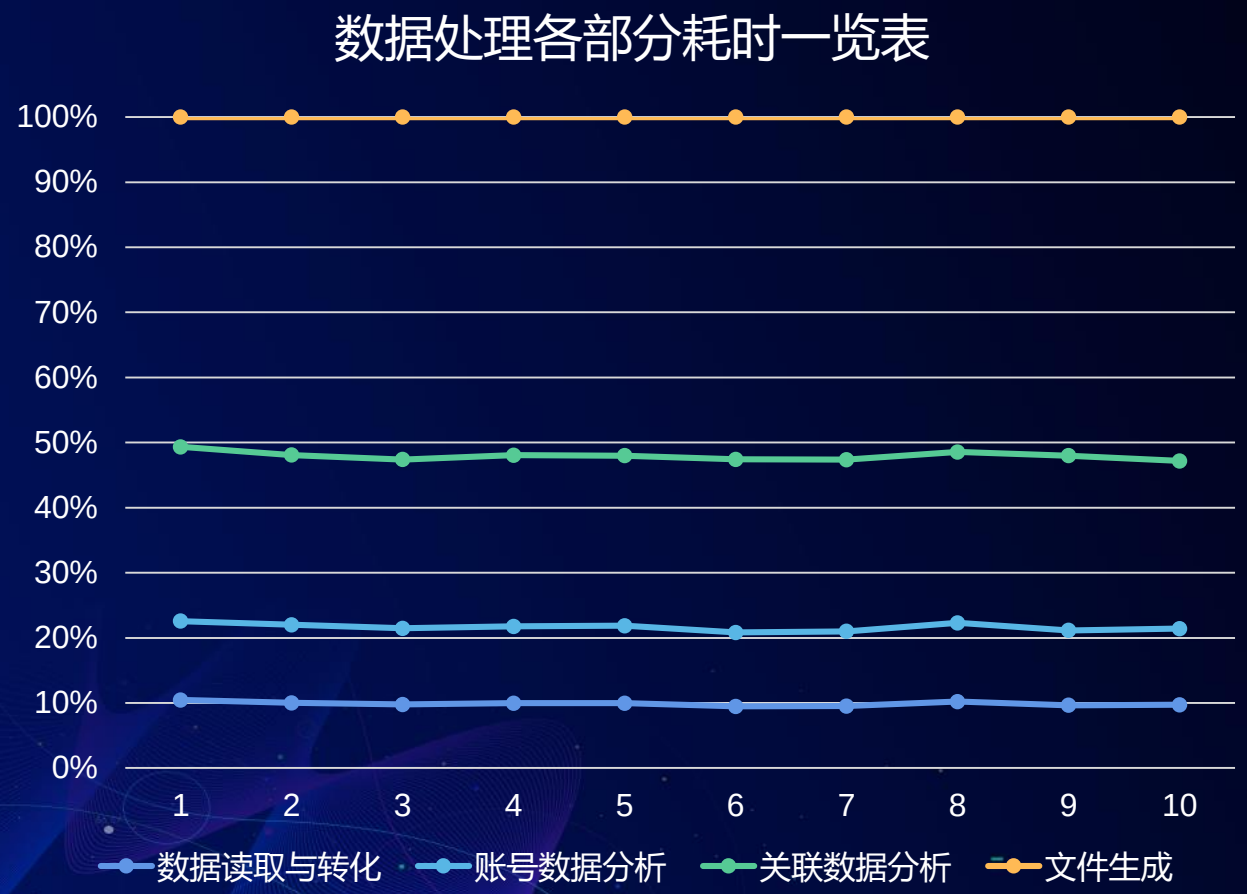
示例

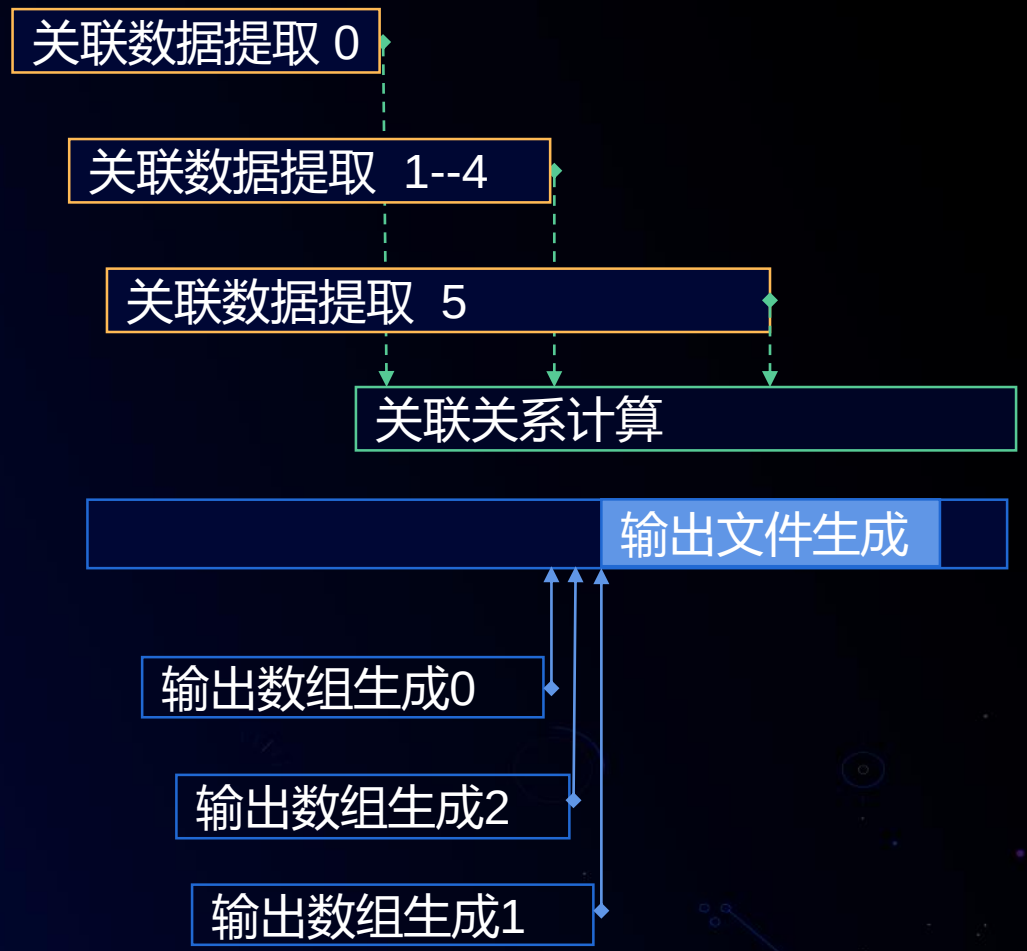
{0,3} {3,5} {5,8} ⇨ 0 1 2 0 4 0 6 7 0

特点：通过判断进行有限深度并查

优势：路径压缩，有限深度，局部并查

线程分配

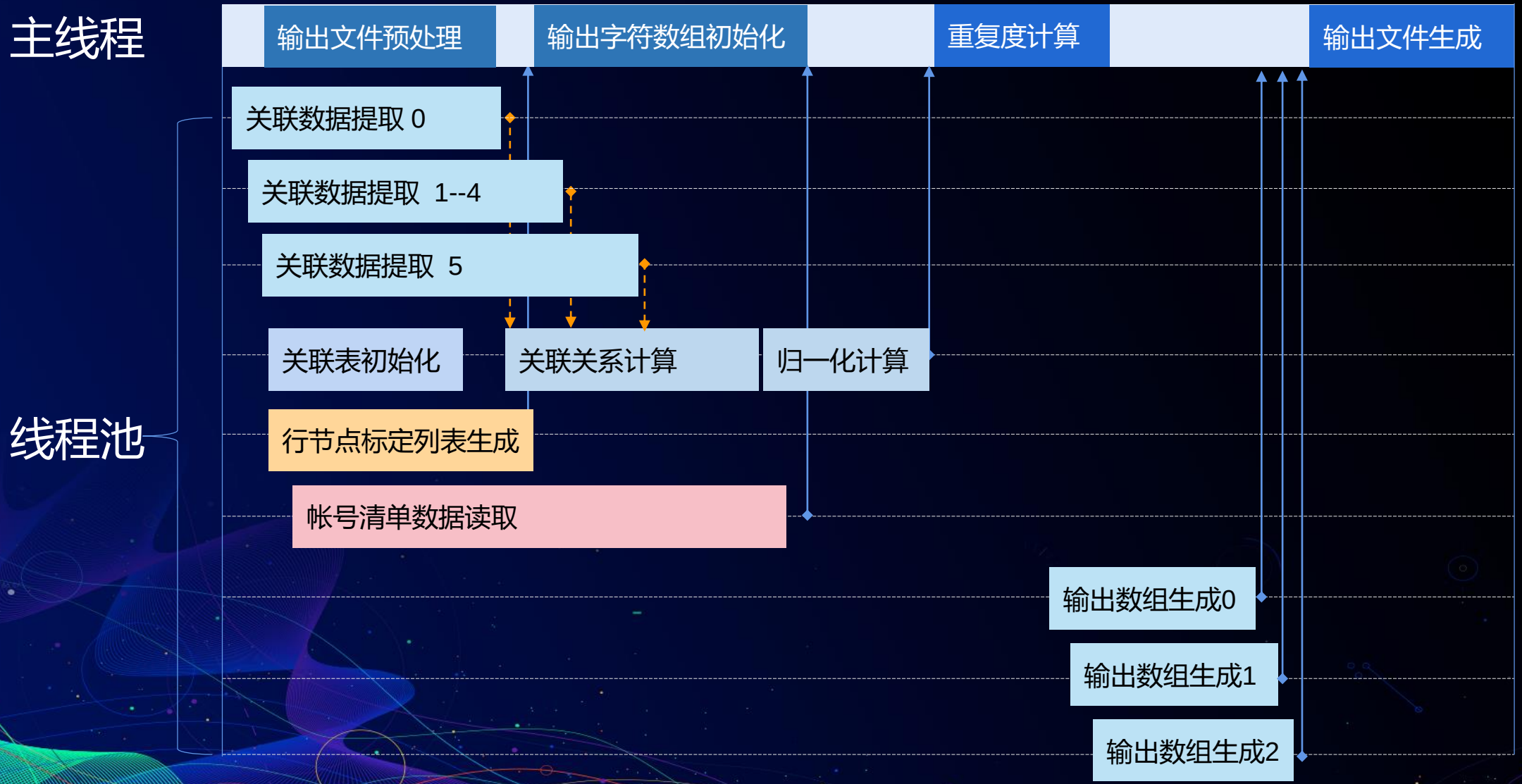




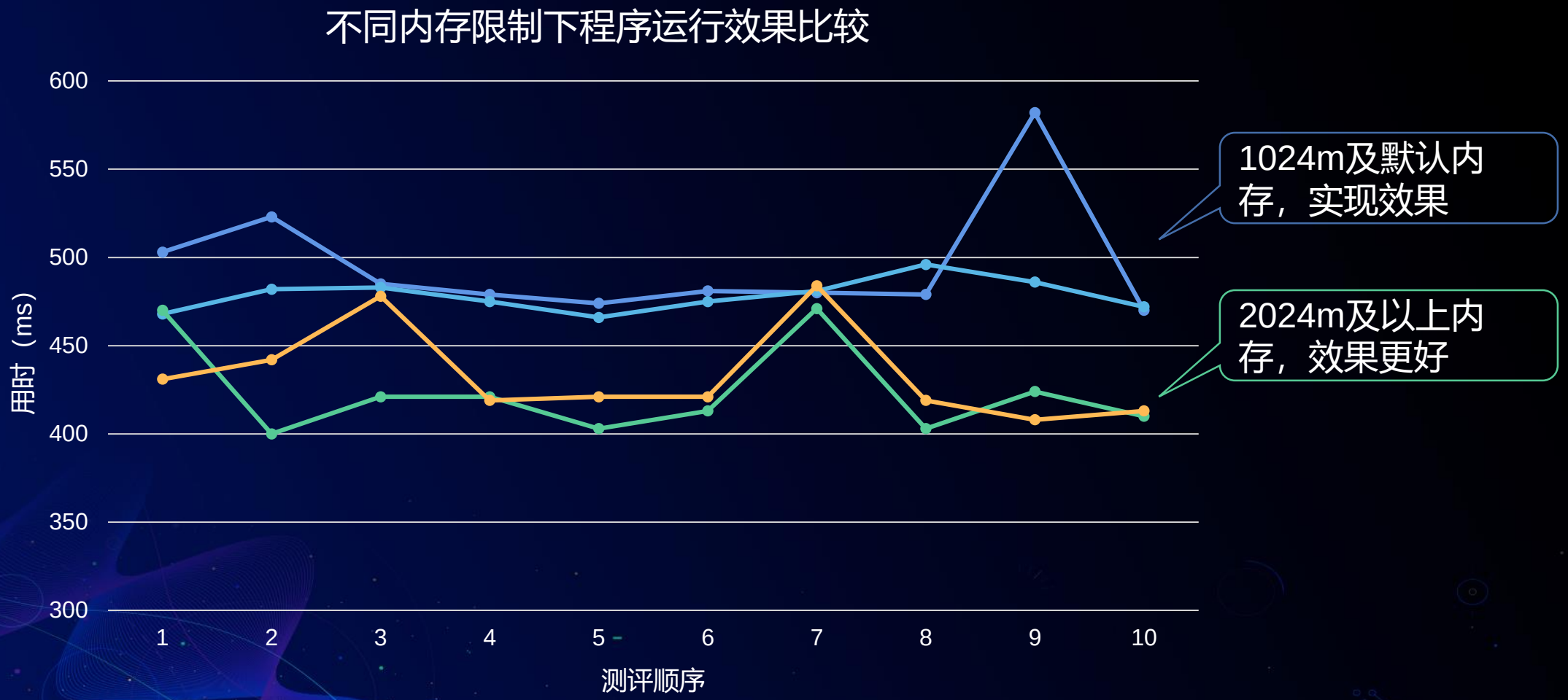
特点：数据读写、关联关系计算耗时较长

处理：分批次递增式多线程读取;依赖关系分析任务即时调用；分批次数据生成

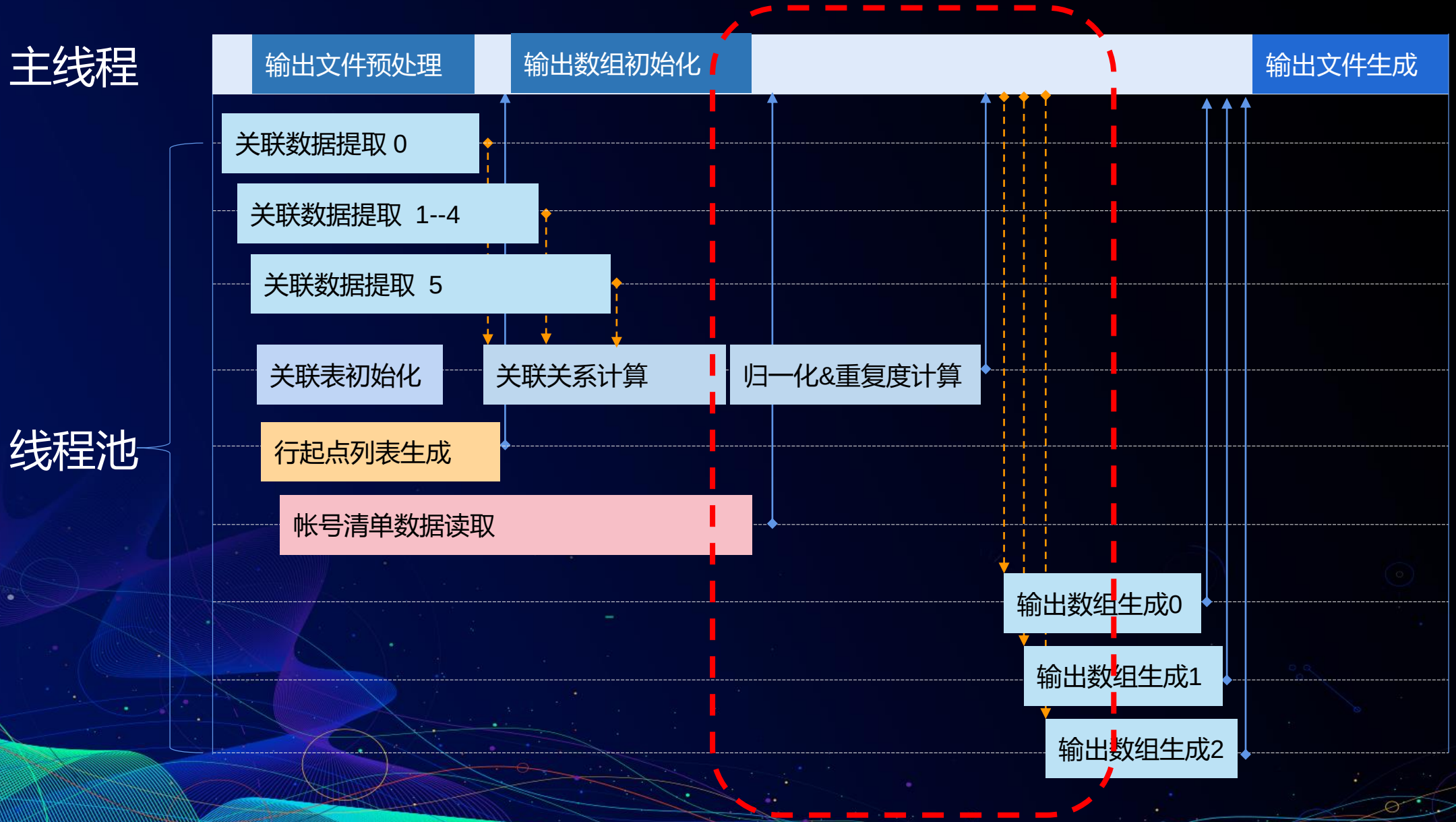
最终方案（2.0）


主线程
输出文件预处理
输出字符数组初始化
重复度计算
输出文件生成
线程池
关联数据提取 0
关联数据提取 1--4
关联数据提取 5
关联表初始化
关联关系计算
归一化计算
行节点标定列表生成
帐号清单数据读取
输出数组生成0
输出数组生成1
输出数组生成2

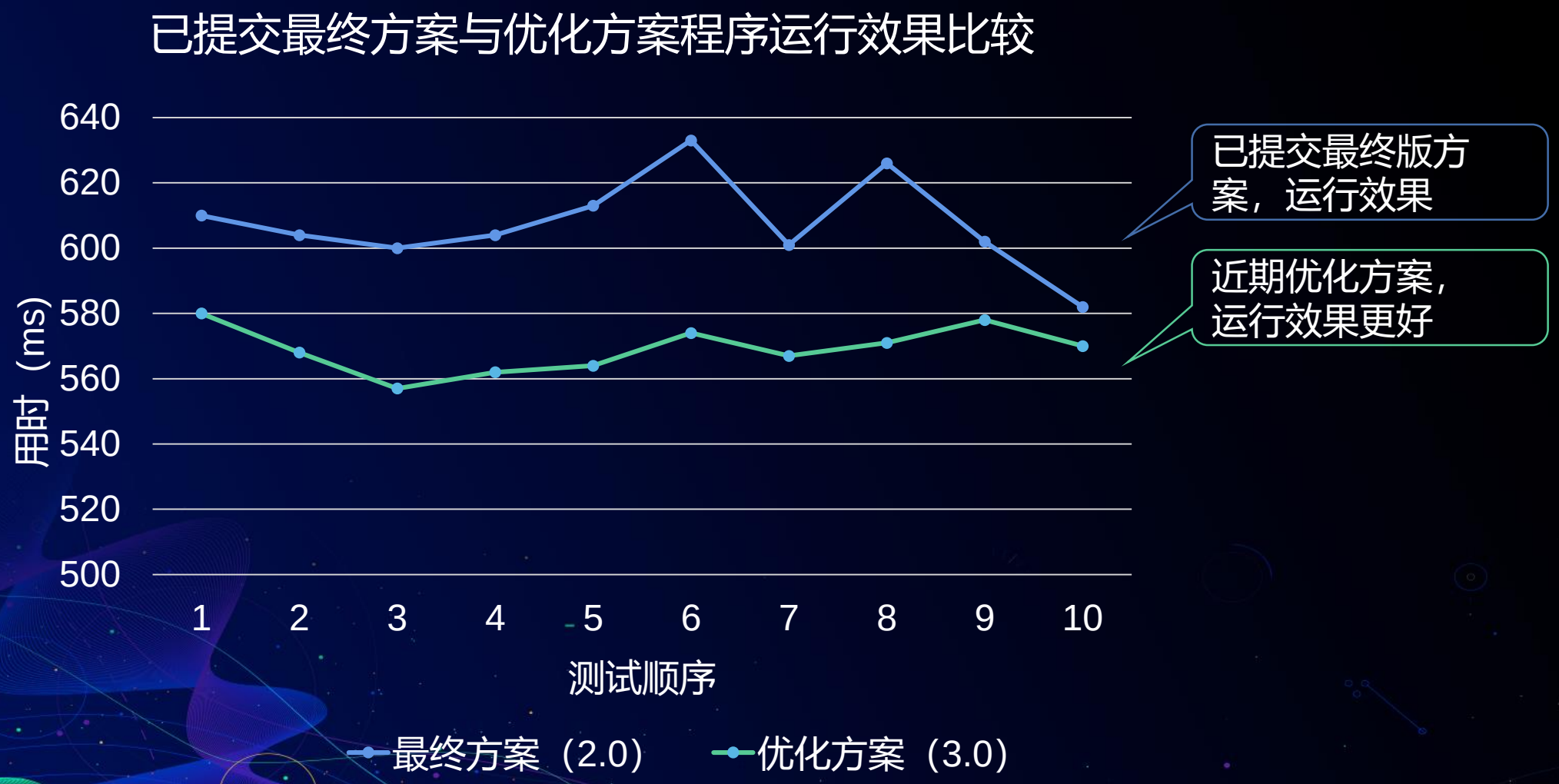
结果分析


不同内存限制下程序运行效果比较
600
550
500
450
400
350
300
用时（ms）
1
2
3
4
5
6
7
8
9
10
测评顺序
默认
–Xms1024m –Xmx1024m –Xmn256m
-Xms2024m -Xmx2024m -Xmn1024m
–Xms3550m –Xmx3550m –Xmn2g
1024m及默认内存，实现效果
2024m及以上内存，效果更好

优化方案（3.0）


主线程
输出文件预处理
输出数组初始化
输出文件生成
线程池
关联数据提取 0
关联数据提取 1--4
关联数据提取 5
关联表初始化
关联关系计算
归一化&重复度计算
行起点列表生成
帐号清单数据读取
输出数组生成0
输出数组生成1
输出数组生成2

优化方案（3.0）


已提交最终方案与优化方案程序运行效果比较
640
620
600
580
560
540
520
500
用时（ms）
1
2
3
4
5
6
7
8
9
10
测试顺序
最终方案（2.0）
优化方案（3.0）
已提交最终版方案，运行效果
近期优化方案，运行效果更好

总结与展望

- ✓ 方案对数据集进行分批次处理，适用于更大规模数据集；
- ✓ 方案尽量避免过多的中间变量生成，能够合理复用已经分配的内存空间，适用更小内存；
- ✓ 方案采用的任务调度、索引优化和查询等算法都对程序整体运行效率产生重大影响，适用于多种大数据应用场景。

谢谢！