大数据系统软件 Lab2 实验报告

组员

- 陈荣钊 2018010863
- 伍冠宇 2018010683
- 张后斌 2018010858

实验内容

在 Lab1 的基础上，实现分布式系统中对表的自然连接功能。

代码架构

Cluster.Join

Cluster.Join 依据表的输入顺序，自左向右依次进行 Join。

Node.JoinTableRPC

Node.JoinTableRPC 需要传入两个参数：需要 Join 的本地表的名称；远端传来的数据集。

Node.JoinTableRPC 首先提取 传入Schema 和 本地Schema 的 ForeignKey，然后会使用 ForeignKey 完成 Join 操作。

附加功能

join over union

每对发生 Join 的 Slave Node 会产生一个数据集，Cluster.Join 不会在两张表的 Join 操作后马上合并这些数据集，而是先将结果缓存在内存，等待所有 Join 操作完成后才执行合并。

多表连接

Cluster.Join 执行时，首先 Join 第一张表和第二张表，得到一系列结果集。接着使用得到的每个结果集依次 Join 每个 Slave Node 中保存的下一张表。

多表连接的测试代码见文件 `lab2_multi_table_test.go` 。