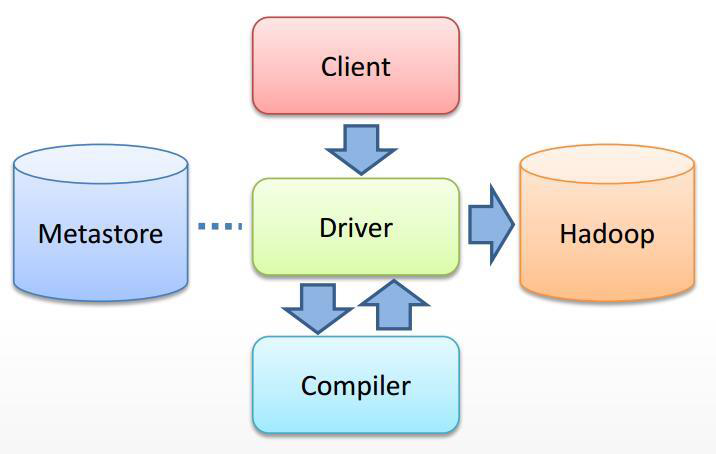
一：hive执行过程概述

1. 概述：
   1. hive将hql转换成一组操作符(operator),比如groupbyOperator,joinOperator等
   2. 操作符Operatator是hive的最小执行单元
   3. 每个操作符代表一个hdfs操作或者mr程序
   4. hive通过execMapper和ExecReducer执行MapReduce程序，执行模式分为本地模式和分布式模式



1. hive操作符列表



1. hive编译器的工作职责
   1. parser:将hql语句转换成抽象语法树（ast）
   2. Semantic Analyzer:将抽象语法树转换成查询块
   3. Logic Plan Generator:将查询块转换成逻辑查询计划
   4. Logic Optimizer:重写逻辑查询计划，优化逻辑执行计划
   5. Physical Paln Generator:将逻辑计划转换成物理计划(Mapreduce jobs)
   6. Physical Optimizer:选择最佳的join策略，优化物理执行计划
2. 优化器类型



上表中带①的，优化目的都是尽量将任务合并到一个job中，以减少job数量，带②的目的是减少shuffle数据量

二：join

1.对于join操作：

|  |
| --- |
| SELECT pv.pageid, u.age FROM page\_view pv JOIN user u ON pv.userid = u.userid; |

1. 实现过程

Map：

以join on 条件中的列作为Key,如果有多个列，则key是多个列的组合

以 JOIN 之后所关心的列作为 Value，当有多个列时，Value 是这些列的组合。在 Value 中还会包含表的 Tag 信息，用于标明此 Value 对应于哪个表

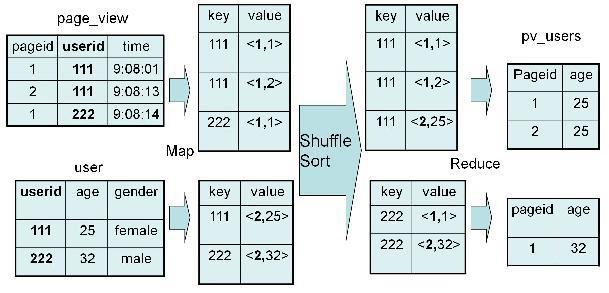
3、按照 Key 进行排序

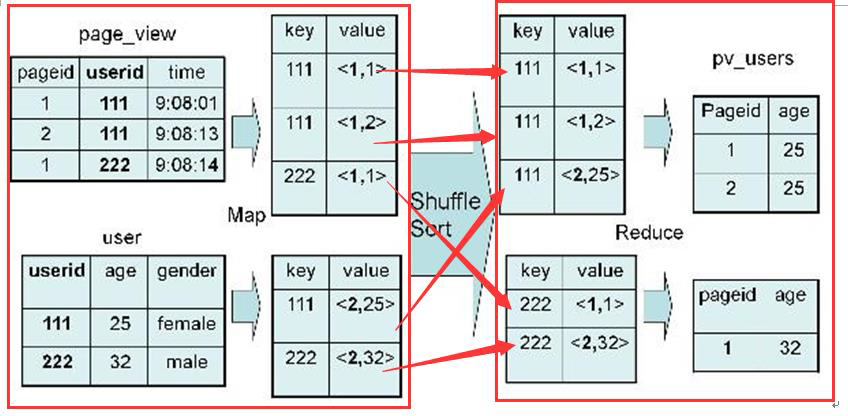
**Shuffle**：

　　1、根据 Key 的值进行 Hash，并将 Key/Value 对按照 Hash 值推至不同对 Reduce 中

**Reduce**：

1. Reducer 根据 Key 值进行 Join 操作，并且通过 Tag 来识别不同的表中的数据
2. 具体实现过程

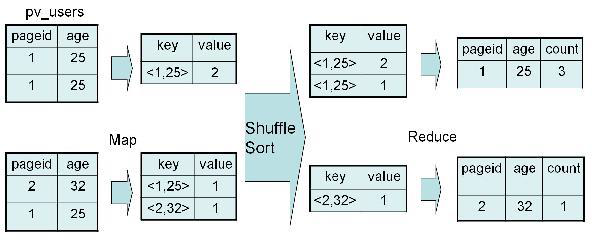


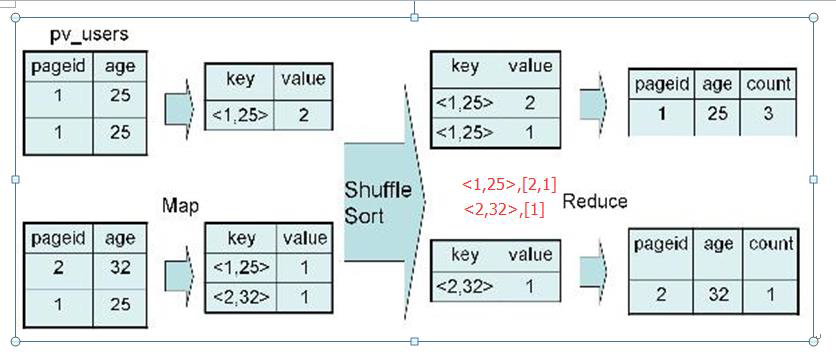


1. group by
2. 对于group by操作

|  |
| --- |
| SELECT pageid, age, count(**1**) FROM pv\_users GROUP BY pageid, age; |

2.实现过程



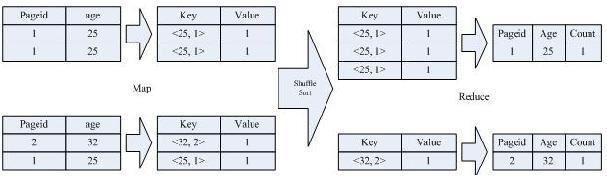


四：disinct

1. 对于distinct的操作
2. 按照 age 分组，然后统计每个分组里面的不重复的 pageid 有多少个

|  |
| --- |
| SELECT age, count(distinct pageid) FROM pv\_users GROUP BY age; |

2.实现过程



3.详细过程解释

该 SQL 语句会按照 age 和 pageid 预先分组，进行 distinct 操作。然后会再按 照 age 进行分组，再进行一次 distinct 操作