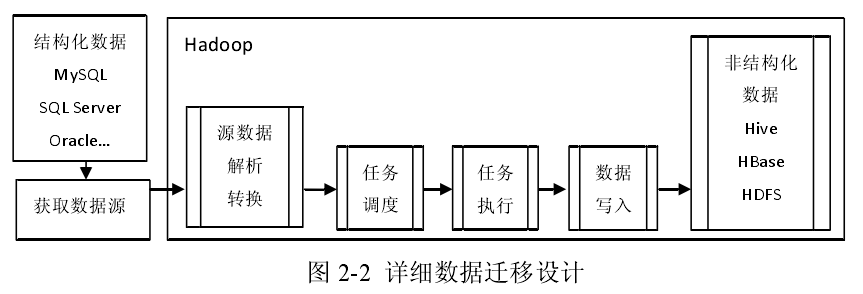
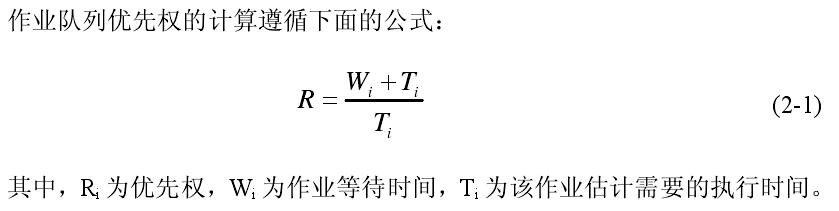
**《基于Hadoop平台的大数据迁移与查询方法研究及应用》**

1. 通常在迁移过程中，自动化和人工处理缺一不可。因为两者并用可以提高数据质量，消除多余或过时的信息，并匹配新系统的要求。在新系统部署之前，对于中等复杂度或高复杂度的应用程序而言，数据迁移的过程（包括设计、抽取、清洗、加载、校验）通常会反复数次。
2. 由于 SQL Server和HBase数据库在很多方面，比如：数据查询、表的结构、触发器及视图等都不尽相同，因此在决定数据转换方案时，需要为数据类型的变化和新旧系统之间的兼容性做好准备。
3. 事实上，将数据从关系型数据库中转移是一种流水线的动作，首先会将数据从SQLServer数据库中导出，随后将数据转成NoSQL的数据格式，最后在新的非关系型数据库中保存。



1. 本系统中将Java的结果集（Result Set）转化为MapReduce处理所需键/值（key/value）对。接下来，需要使用MapReduce调度器进行任务调度，合理安排作业执行顺序及按需分配资源以提高迁移性能。
2. MapReduce是一种可以在大型数据集的系统下，并行工作的分布式编程模型集群，由于MapReduce可以利用 Hadoop的分布式集群实现更优的并发性，因此可用于优化系统在大量数据转换时的处理能力。
3. Map Reduce作业存在的几个阶段中，不论是初始化或是运行环境的准备都只和硬件条件相关。而在执行Map Reduce处理任务的时候，任务调度算法是影响其运行速度和效率的重要因素。
4. Hadoop本身具有几种常用的调度算法可供选择，可以让用户能够根据不同的情况来使用不同调度器以达到效率最优。所以将数据从关系型数据库迁移到非关系型数据库HBase的过程中，即当Job Tracker收到了来自Task Tracker的信息后，如何选择执行任务可以使得迁移效率得到优化需要进行深入探讨。
5. 一般情况下Hadoop拥有四种调度器：先进先出调度器（FIFO，Job Quene Task Scheduler）、改 进 先 进 先 出 调 度 器 （Limit FIFO，Limit Tasks Per Job Task Scheduler）、公平调度器（FS，Fair Scheduler）和计算能力调度器（CTS，Capacity Tesk Scheduler）。
6. Fair Scheduler公平调度器的设计思想，是尽量让所有提交的作业都可以按照比例（基于权重）获得相应的资源，实现了某种意义上的公平。当集群系统只运行单个作业时，该作业可以使用系统全部资源，而当其他作业被提交执行时，系统将会进行资源调整、重新分配。而公平调度器也有它的不足之处，虽然它兼顾到各个队列中的作业，保证了资源的公平共享，但是它在分配资源时候并没有考虑到作业的优先权。
7. 针对作业的优先权来对公平调度器进行略微的改进。
8. 公平调度器主要由作业池管理器（Pool Manager）、负载均衡器（Load Manager）、任务选择器（Task Selector）、权重调整器（Weight Adjuster）、作业调度更新线程（Update Thread）五大部分组成。
9. 基于动态优先权调度算法的核心思想是：单位时间内，定期地动态更新作业的优先权值，从而让调度器在公平亏欠度相等的情况下，对作业的优先权情况进行考虑，选择最合适的作业进行调度执行。
10. 该算法均衡不同作业调度：对于同样的等待时间，优先权随着作业服务时间越短而越高，而对于长作业，等待时间越长优先权可相应的提高，不会长期处于 starving 状态。【31】



1. 公平亏欠度和作业计算资源量会相互作用、相互影响，即公平亏欠度的计算过程与作业计算资源量相关，而基于公平亏欠度的排序又反过来会影响资源分配，从而影响计算资源里的变化。