Scheduling of Extract, Transform, and Load (ETL) Procedures with Genetic Algorithm

Vedran Vrbanić, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
Damir Kalpić, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

ABSTRACT

A number of ETL procedures are used in the process of loading data to data warehouse systems. Some procedures can be executed concurrently in parallel mode, while for the others there are precedence constraints. Thus, the problem in scheduling procedures for execution is similar to the problem of scheduling of jobs in multiprocessor systems. The solution to this problem has been proposed in the optimum schedule of jobs minimizing the total execution time. When optimizing the schedule for ETL procedures, minimization of the total execution time is not the primary goal. Namely, the ETL procedures provide data required for reports aimed for business users and such reports need to be prepared until the user-defined deadlines. If the deadlines are not breached, the solution is satisfactory, regardless of the total execution time. Also, one cannot assume that all ETL processes are of the same importance – some have higher priorities than the others. That is the reason why prioritization and introduction of explicit bounds to completion time for individual ETL processes is attempted with genetic algorithm (GA). This paper encompasses implementation of the algorithm, experiments with different parameters and testing the quality of obtained solutions.

Keywords: Banking, Data Warehousing, ETL, Genetic Algorithm, Optimization, Scheduling, Timetabling

1. INTRODUCTION

A scenario for optimum execution schedule of ETL procedures is of great significance for functioning of a real-world data warehouse (DWH) system (Vassiliadis, Simitsis, Georgantas, Terrovitis, & Skiadopoulos, 2005). Relational database management system is a powerful solution in storing and processing structured data and generating reports using complex queries, including features like transaction consistency, real-time writing and
multi-table queries (Abdelhafez, 2014), but ETL process development constitutes the most expensive part of a data warehouse project, expressed both in consumption of time and resources. Principal reason for this is that in practice ETL processes have been traditionally designed with a specific vendor tool in mind, from the start of the data warehouse project. Unfortunately, commercial ETL tools are not easy to use, they have a steep learning curve, due to a lack of standard capabilities, e.g. they provide different underlying languages with a wide spectrum of functionality features or complex wizards (Akkaoui, Zimányi, Mazón, & Trujillo, 2013). During the first years of the 21st century, the importance of ETL procedures for data integration tasks has become obvious since they are expensive, labour intensive, mission critical and therefore most important for success of any data warehousing project. The difficulty lies in the combination of data integration and data cleaning tasks where data are transferred from one data store to another, while the details of schema and value mappings play a great role (Anitha & Babu, 2014).

A DWH system is designed in such a way that data processing begins late at night or in early morning hours. In a perfect scenario case, complete data extraction process should be completed before the daily work starts, so the business users can always have the latest data at their disposal. This goal is sometimes reachable, and sometimes not, due to different limitations in the legacy information systems and alike, so that occasionally one has to improvise with imperfect solutions.

The term optimal schedule needs to be defined in the context of this problem. A DWH system consists of a certain number of data marts which are used as a background for the creation of corresponding reports (Bonifati, Cattaneo, Ceri, Fuggetta, & Paraboschi, 2011). Every report must be completed until its deadline. In this perspective, the best schedule is the one in which the minimum number of reports breaches user-defined deadlines. Total execution time does not affect the quality of the solution. The reports differ in importance. E.g. if a bank’s DWH system is in question, the reports on bank portfolio structure which are sent to Central National Bank are of maximum priority because, in case of their delays, the bank would have to pay penalties. That is why the system must urge procedures which are a precondition for the creation of these kind of reports; even if that means that lower priority reports would be delayed. Regardless to the prioritization, explicit time limits must also be introduced. For example, if the Central National Bank requests data every day by 11 AM, then this condition must be specified within the genetic algorithm.

DWH systems fetch data from a number of legacy transaction systems. There are cases when some of these sources fail at the moment when the DWH system must start the data loading process. Data processing cannot be initiated until the data source is enabled. This is a real scenario and in that case the ETL procedure execution schedule is not optimal. But, the solution to this problem is fairly simple, because such cases can be predicted due to previous data source failures and one can create the optimal schedule for this scenarios using GA. If such a scenario happens, one simply should implement the adequate schedule which now does not contain procedures dependent on inoperable data sources. As those procedures are executed using shell scripts or Database Management System (DBMS) jobs, it is not difficult to use different schedules due to data source conditions.

2. PROBLEM PRESENTATION

The job-shop scheduling problem (JSSP) is a common problem in the manufacturing industry. A classical JSSP can be described as a set of N jobs denoted by \( j_i \) where \( i = 1,2,...,N \) which have to be processed on a set of M machines denoted by \( m_k \) where \( k = 1,2,...,M \). Operation of the \( i^{th} \) job on the \( k^{th} \) machine will be denoted by \( o_{ik} \) with the processing time \( p_{ik} \). A schedule is a complete ordered set of operations, required by a job, to be performed on different machines.
Managing Data and Information Quality in Outbound Transportation Systems: A Systematic Approach
www.igi-global.com/article/managing-data-information-quality-outbound/62021?camid=4v1a

Performance Measures and RTB Optimization
Wenxue Huang, Yuanyi Pan and Jianhong Wu (2014). *Encyclopedia of Business Analytics and Optimization* (pp. 1847-1855).
www.igi-global.com/chapter/performance-measures-and-rtb-optimization/107373?camid=4v1a