Universidade do Minho
Escola de Engenharia

Semana da Escola de Engenharia
October 24 - 27, 2011

DESIGNING ETL TASKS FOR GRID ENVIRONMENT EXECUTION

Vasco Santos
Department of Informatics
E-mail: vsantos@estgf.ipp.pt

KEYWORDS
Data Warehousing, ETL, GRID, Parallel Processing, Relational Algebra.

ABSTRACT

Data Warehouses store integrated and consistent data in a subject-oriented data repository dedicated especially to support business intelligence processes. Nevertheless, in order to maintain a data warehouse up-to-date, data intensive tasks retrieve regularly specialized information from specific preselected information sources, transforming and conforming it accordingly to some specific business requirements provided by decision-makers (Kimball et al. 1998). Such tasks, commonly named as Extract-Transform-Load (ETL) processes, have a limited time frame window to be executed over an ever increasing amount of data with extremely complex operations. The common approach to deal with the need of more computational power is the acquisition of new and more powerful hardware. This expensive approach disregards the unused computational resources available in desktop computers already present at most enterprises' computational environments. This work intends to define a different approach to deal with ETL processes, taking advantage of parallel processing over a GRID environment using XML data as an effective support to data storage and communication, demonstrating that GRID environments could be a real alternative for the implementation of low cost data warehouses.

The approaches to deal with distribution and parallelization of tasks have evolved significantly during the last few years. New terms have arisen. Today, GRID computing (Foster et al. 2001) and Cloud computing (Vouk 2008) are quite common to appear when we think of doing some job in a distributed and parallel way. Organizations have started to test grid middleware software to take advantage of the inactivity of their computing devices to help them in processing intensive tasks, in order to take advantage of the processing power available but underused. This approach maximizes the investments already made and postpones expensive computer acquisitions (Demiya et al. 2008).

Business Intelligence is a component of an Enterprise Information System that deals with large amounts of data and provides analysis that support management's decision-making process. Since the amount of data normally increases through time, the processing power needed to analyze it in due time also increases undermining the infrastructure available. A GRID environment might be a suitable solution to this kind of problem due to the easiness and inexpensiveness of adding new processing nodes to the infrastructure (Pascal 2005; Poess and Nambiar 2005).

The scientific community has concentrated efforts in studying the application of grid environments in everyday business operations, with the purpose of designing and providing low cost computational power to data intensive operations. More recently, researchers shifted a little bit their focus of attention, relegating for a second place the grid itself (architecture, models, functionalities, etc.) to look for the complex problem of task distribution and management (Mury et al. 2010), particularly in the distribution and management of workflows (Blythe et al. 2005): one field of study that is becoming more popular in Data Warehouse Systems due to their inherent distributed characteristics. These characteristics blend well with the possibilities that any grid environment provides. However, not all activities in a DWS environment have been extensively studied in conjunction with the advantages that a grid environment provides by nature. The ETL processes are data intensive tasks that prepare transactional data to be loaded into a DW (Kimball and Caserta 2004; Albrecht and Naumann 2008). They have been studied widely mainly in their modeling phase (Simitsis 2003) and in the optimization of the workflows that are generated (Vassiliadis et al. 2005; Simitsis et al. 2010). However, it is recognized that there is a lack of research in the combination of grid environments and ETL processes. Our proposal focuses in the advantages that can be obtained by using a grid environment as the main infrastructure to support the execution of ETL tasks. Nevertheless, in order to take advantage of the potentialities of a grid in ETL, a very structured workflow of operations must be defined. Therefore, we must decompose the ETL logical model in operations that can be distributed over the grid. Our choice was based on the relational algebra operators, more specifically, on extended relational algebra operators (Albert 1991; Grefen and de By 1994). With extended relational algebra is possible to represent the most common ETL operations, and so distribute the workflow of operations over the grid. Data is stored in XML format for better compatibility in a heterogeneous environment being relational algebra operations coded in JAVA.

The next step of our work deals with the scheduling of the grid ETL workflow with particular focus on availability, performance prediction and proximity due mainly to bandwidth bottlenecks.

REFERENCES

Albert, J. 1991. *Algebraic Properties of Bag Data Types*. Proceedings of the 17th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc.

Albrecht, A. and F. Naumann 2008. *Managing ETL Processes*. Proceedings of the International Workshop on New Trends in Information Integration, NTII 2008, Auckland, New Zealand.

Blythe, J., S. Jain, et al. 2005. *Task scheduling strategies for workflow-based applications in grids*. IEEE International Symposium on Cluster Computing and the Grid, 2005.

Demiya, T., T. Yoshihisa, et al. 2008. "Compact grid : a grid computing system using low resource compact computers." *Int. J. Commun. Netw. Distrib. Syst.* 1,No. 2: 17.

Foster, I., C. Kesselman, et al. 2001. "The Anatomy of the Grid : Enabling Scalable Virtual Organizations." *International Journal of High Performance Computing Applications* 15,No. 3: 200-222.

Grefen, P. W. P. J. and R. A. de By 1994. *A multi-set extended relational algebra: a formal approach to a practical issue*. Data Engineering, 1994. Proceedings.10th International Conference.

Kimball, R. and J. Caserta 2004. *The Data Warehouse ETL Toolkit - Pratical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. , Wiley Publishing, Inc.

Kimball, R., L. Reeves, et al. 1998. *The Data Warehouse Lifecycle Toolkit - Expert Methods for Designing, Developing, and Deploying Data Warehouses*. , John Wiley & Sons, Inc.

Mury, A. R., B. Schulze, et al. 2010. "Task distribution models in grids: towards a profile-based approach." *Concurrency and Computation: Practice and Experience* 22,No. 3: 358-374.

Pascal, W. 2005. *A Model for Distributing and Querying a Data Warehouse on a Computing Grid*.

Poess, M. and R. O. Nambiar 2005. Large scale data warehouses on grid: Oracle database 10*g* and HP proliant servers. *Proceedings of the 31st international conference on Very large data bases*. Trondheim, Norway, VLDB Endowment. No.**:** 1055-1066.

Simitsis, A. 2003. *Modeling and managing ETL processes*. 29th International Conference on Very Large Data Bases, Berlin.

Simitsis, A., K. Wilkinson, et al. 2010. *Optimizing ETL workflows for fault-tolerance*. Proceedings of the 26th International Conference on Data Engineering, Long Beach, California.

Vassiliadis, P., A. Simitsis, et al. 2005. Blueprints and Measures for ETL Workflows. No.**:** 385-400.

Vouk, M. A. 2008. "Cloud Computing – Issues, Research and Implementations." *Journal of Computing and Information Technology* 16,No. 4: 235-246.

AUTHORS' BIOGRAPHIES

VASCO SANTOS was born in Oporto, Portugal and went to the University of Minho, where he studied Systems and Informatics Engineering and obtained his degree in 1997. He worked for a three years in the textile industry before starting his academic career at ESTGF.IPP. He obtained his Master degree in 2004 and is currently working on his PhD in the Department of Informatics at University of Minho. His e-mail address is : vsantos@estgf.ipp.pt.