The Warehousing of Application Services Data to Facilitate Identification of Patterns of Network Usage

Ronald Knoetze, Janet Wesson, Charmain Cilliers
Department of Computer Science and Information Systems
University of Port Elizabeth, PO Box 1600, Port Elizabeth, 6000
Tel: (041) 504 2323, Fax: (041) 504 2831
Email: {Ronald.Knoetze, Janet.Wesson, Charmain.Cilliers}@upe.ac.za

Abstract – Techniques that identify patterns of network usage are useful in optimising network performance. The warehousing of application services data can facilitate the implementation of such techniques. The University of Port Elizabeth has an extensive network that supports several application services. The aim of this paper is to discuss the creation of a data warehouse to store application services data collected by PacketShaper, which will be used in the pattern identification process.

Index Terms – data warehousing, application services data, data mining, network performance metrics

INTRODUCTION

THE University of Port Elizabeth (UPE) has an extensive network which supports various application services, such as Integrated Tertiary Software (ITS). These services are used by staff and students at UPE, resulting in thousands of transactions taking place daily.

PacketShaper is a network monitoring tool that was purchased by UPE and placed on the network between the ITS server and the client computers. PacketShaper monitors different aspects of a network, with the data collected being classified into different types of measurement variables [1]. The log files created by PacketShaper are large in size and are stored in three different files. The measurement variables from the log files are then used to create reports for network managers to view, but these reports do not provide an in-depth analysis of patterns of network performance.

The goal of this paper is to discuss the creation of a data warehouse to store the application services data that is collected by PacketShaper. This data warehouse will be used in the data mining process to find patterns and identify future trends in network performance.

I. DATA WAREHOUSE CREATION

A data warehouse is a decision support database that combines multiple heterogeneous data sources and stores these at a single site. A data warehouse is subject-orientated, integrated, time-variant and non-volatile [6]. Data warehouses support information processing by providing a consolidated database that assists with analysis. The data that is incorporated into a data warehouse goes through different phases to ensure that the data is integrated and non-volatile. The pre-processing phases include: data identification, data cleaning, data integration and data transformation.

A. DATA IDENTIFICATION

Network characteristics are the properties of the network that are related to the performance and reliability of the network [2]. Lowecamp et al [2], Leese [3] and Barnford [4] discuss different characteristics that need to be measured. However, they do agree on some common network metrics. These network metrics were used to identify the metrics that best apply to the UPE network. The metrics identified were: delay, throughput, response time, utilization and loss. Once the network metrics were chosen, the variables measured by PacketShaper were mapped onto these network metrics.

B. DATA CLEANING

PacketShaper collects over 100 variables which are classified into 3 different groups, namely *link*, *partition* and *class* variables. Depending on the set-up of the network and how PacketShaper is configured, some of the variables in the different groups are not measured. In the case of UPE, there is only one partition, namely for ITS, and as a result most of the partition variables are not collected. Table 1 indicates the network metrics that were identified as well as examples of PacketShaper variables that were mapped onto the corresponding network metrics. Note that not all the identified PacketShaper variables are shown.

Network Metric	PacketShaper Variables
Delay (ms)	network-delay-avg, normalized-network-delay-avg, server-delay-avg, total-delay-avg
Throughput (kbps)	total-passthru-bytes, total-rx-bytes, total-tx-bytes
Response Time (ms)	avg-round-trip-time, round-trip-time-msecs
Utilization (bps)	avg-bps, kbytes, peak-bps, pkts, total-trans, trans-bytes-avg
Loss (bps)	rx-errors, rx-pkts-dropped, tx-errors, tx-pkts-dropped

Table 1 - Network Metrics and the associated PacketShaper Variables

C. Data Integration and Transformation

Based on the PacketShaper variables that were identified, a snowflake schema was identified to model the data, as shown in Figure 1. The fact table contains the measures that were identified in Table 1. The *class* and *link* tables are subclasses of the fact table. The dimensions *VLAN*, *department* and *time* define how these facts are to be aggregated, with the attributes providing a description for each dimension. The generalised attributes "common variables", "class variables" and "link variables" are shown in place of the actual variables from each group that are to be used.

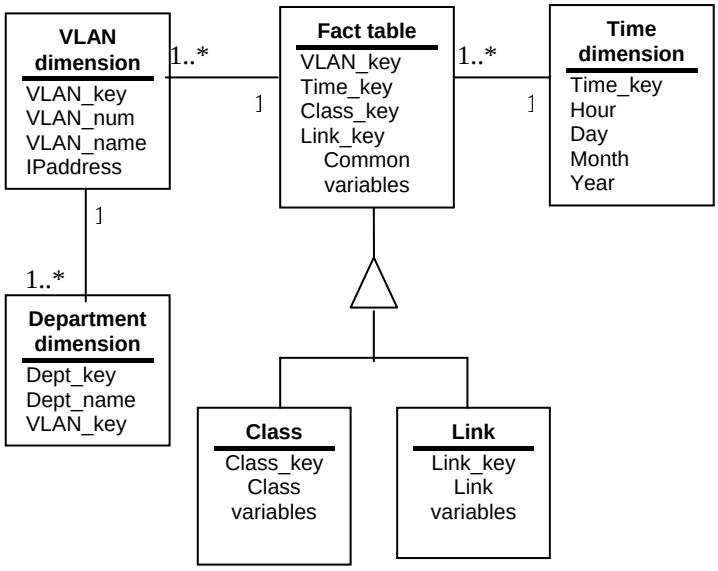

Figure 1 – Snowflake Schema for Data Warehouse

The relationship between the main fact table and the dimensions in this schema is a many-to-one relationship. This technique saves space within the database, but it provides an increase in the number of dimension table joins on foreign keys between these dimensions. This results in more complex queries and reduced query performance [8].

II. Status of Project

At this stage of the project, the data structure has been identified as well as the architecture of the data warehouse. The data warehouse is currently being developed and populated with the application services data collected by PacketShaper. Data pre-processing is being performed on the data from PacketShaper to ensure that the data is consistent, normalised and that any inconsistencies are removed. This pre-processing ensures that all irrelevant data is eliminated.

III. Future Work

Once the data warehouse has been populated, the mining of the data warehouse using various data mining algorithms can take place. Research will be conducted to find out what data mining algorithms exist and how these can be applied to the data in the data warehouse. There are various data mining programs that exist. One such program is WEKA, developed by the University of Waikato, which uses Java to implement various data mining algorithms [9]. Once the data mining algorithms have been identified, a visual representation of the data mining results will be developed.

IV. Conclusion

A data warehouse is needed to store the application services data collected by PacketShaper on the UPE network. The data collected consists of network metrics such as delay and throughput. This data warehouse will facilitate the identification of patterns of network performance and the identification of bottlenecks.

Acknowledgment

We would like to thank the Telkom Centre of Excellence programme and the Department of Computer Science and Information Systems at the University of Port Elizabeth for making this research possible.

References

[1] PACKETEER (2001): Four Steps to Application Performance across the Network. http://support.packeteer.com/documentation/packetguide/5.3.0/documents/4steps.pdf

[2] LOWEKAMP, B., TIERNEY, B., COTTRELL, L., HUGHES-JONES, R., KIELMAN, T. and SWANY, M. (2004): A Hierarchy of Network Performance Characteristics for Grid Applications and Services. http://www-didc.lbl.gov/NMWG/docs/draft-ggf-nmwg-hierarchy-02.pdf

[3] LEESE, M. (2003): GridMon - Grid Network Performance Monitoring for UK e-Science. http://www.gridmon.dl.ac.uk

[4] BARNFORD, P. (2003): Network Performance Measurement and Analysis. www.cs.wisc.edu/~pb/640/perform.ppt

[5] HAN, J. and KAMBER, M. (2001): *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

[6] DATAWAREHOUSING.COM (2003): Star and Snowflake Schemas - Unveiling the Differences. www.datawarehousing.com/techtips/techtip10.asp

[7] WITTEN, I.H. and FRANK, E. (2000): *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann.

Ronald Knoetze received his BSc Hons degree in 2003 from the University of Port Elizabeth. He is presently doing his MSc in Computer Science at the University of Port Elizabeth. His current field of research involves using data mining algorithms to obtain patterns from application services data.