Aligning the Warehouse and the Web

Hadrian Peter
University of the West Indies, Barbados

Charles Greenidge
University of the West Indies, Barbados

INTRODUCTION

Data warehouses have established themselves as necessary components of an effective IT strategy for large businesses. To augment the streams of data being siphoned from transactional/operational databases warehouses must also integrate increasing amounts of external data to assist in decision support. Modern warehouses can be expected to handle up to 100 Terabytes or more of data. (Berson and Smith, 1997; Devlin, 1998; Inmon 2002; Imhoff et al, 2003; Schwartz, 2003; Day 2004; Peter and Greenidge, 2005; Winter and Burns 2006; Ladley, 2007).

The arrival of newer generations of tools and database vendor support has smoothed the way for current warehouses to meet the needs of the challenging global business environment (Kimball and Ross, 2002; Imhoff et al, 2003; Ross, 2006).

We cannot ignore the role of the Internet in modern business and the impact on data warehouse strategies. The web represents the richest source of external data known to man (Zhenyu et al, 2002; Chakrabarti, 2002; Laender et al, 2002) but we must be able to couple raw text or poorly structured data on the web with descriptions, annotations and other forms of summary meta-data (Crescenzi et al, 2001).

In recent years the Semantic Web initiative has focussed on the production of "smarter data". The basic idea is that instead of making programs with near human intelligence, we rather carefully add meta-data to existing stores so that the data becomes "marked up" with all the information necessary to allow not-so-intelligent software to perform analysis with minimal human intervention. (Kalfoglou et al, 2004)

The Semantic Web builds on established building block technologies such as Unicode, URIs(Uniform Resource Indicators) and XML (Extensible Markup Language) (Dumbill, 2000; Daconta et al, 2003; Decker et al, 2000). The modern data warehouse must embrace these emerging web initiatives. In this paper we propose a model which provides mechanisms for sourcing external data resources for analysts in the warehouse.

BACKGROUND

Data Warehousing

Data warehousing is an evolving IT strategy in which data is periodically siphoned off from multiple heterogeneous operational databases and composed in a specialized database environment for business analysts posing queries. Traditional data warehouses tended to focus on historical/archival data but modern warehouses are required to be more nimble, utilizing data which becomes available within days of creation in the operational environments (Schwartz , 2003; Imhoff et al, 2003; Strand and Wangler, 2004; Ladley, 2007). Data warehouses must provide different views of the data, allowing users the options to "drill down" to highly granular data or to produce highly summarized data for business reporting. This flexibility is supported by the use of robust tools in the warehouse environment (Berson and Smith, 1997; Kimball and Ross, 2002).

Data Warehousing accomplishes the following:

- Facilitates ad hoc end-user querying
- Facilitates the collection and merging of large volumes of data
- Seeks to reconcile the inconsistencies and fix the errors that may be discovered among data records
- Utilizes meta-data in an intensive way.
- Relies on an implicit acceptance that external data is readily available

Some major issues in data warehousing design are:

- Ability to handle vast quantities of data
- Ability to view data at differing levels of granularity
- Query Performance versus ease of query construction by business analysts
- Ensuring Purity, Consistency and Integrity of data entering warehouse
- Impact of changes in the business IT environments supplying the warehouse
- Costs and Return-on-Investment (ROI)

External Data and Search Engines

External data is an often ignored but essential ingredient in the decision support analysis performed in the data warehouse environment. Relevant sources such as trade journals, news reports and stock quotes are required by warehouse decision support personnel when reaching valid conclusions based on internal data (Inmon, 2002; Imhoff et al, 2003).

External data, if added to the warehouse, may be used to put into context data originating from operational systems. The web has long provided a rich source of external data, but robust Search Engine (SE) technologies must be used to retrieve this data (Chakrabarti, 2002; Sullivan, 2000). In our model we envisage a cooperative nexus between the data warehouse and search engines. We introduce a special intermediate and independent data staging layer called the meta-data engine (M-DE).

Search Engines are widely recognized as imperfect yet practical tools to access global data via the Internet. Search Engines continue to mature with new regions, such as the Deep Web, once inaccessible, now becoming accessible (Bergman, 2001; Wang and Lochovsky, 2003; Zillman, 2005). The potential of current and future generations of SEs for harvesting huge tracts of external data cannot be underestimated.

Our model allows a naïve (business) user to pose a query which can be modified to target the domain(s) of interest associated with the user. The SE acts on the modified query to produce results. Once results are retrieved from the SE there is a further processing stage to format the results data for the requirements of the data warehouse.

MAIN THRUST

Detailed Model

We now examine the contribution of our model. In particular we highlight the Query Modifying Filter (QMF), Search Engines submission and retrieval phases, and meta-data engine components. The approaches taken in our model aims to enhance the user experience while maximizing the efficiency in the search process.

A query modification process is desirable due to the intractable nature of composing queries. We also wish to target several different search engines with our queries. We note that search engines may independently provide special operators and/or programming tools (e.g. Google API) to allow for tweaking of the default operations of the engine. Thus the Query Modifying Filter (labeled filter in the diagram) may be used to fine tune a generic query to meet the unique search features of a particular search engine. We may need to enhance terms supplied by a user to better target the domain(s) of a user. Feedback from the meta-data engine can be used to guide the development of the Query Modifying Filter.

The use of popular search engines in our suite guarantees the widest possible coverage by our engine. The basic steps in the querying process is:

1. Get user's (naïve) query
2. Apply QMF to produce several modified, search engine specific queries
3. Submit modified queries to their respective search engines
4. Retrieve results and form seed links
5. Use seed links and perform depth/breadth first traversals using seed links
6. Store results from step. 5 to disk

Architecture

For effective functioning, our proposed system must address a number of areas pertaining to both the data warehouse and SE environments, namely:

1. Relevance of retrieved data to a chosen domain
2. Unstructured/semi-structured nature of data on the web
3. Analysis & Generation of meta-data