Harvesting the Web to Procure Secure Information for Enterprise

Ritu Punhani, Alpana Kakkar, D. Jain

International Research Journal
ISSN 1839-6518
Vol. 1, No. 1, 2011
82800101201105

Harvesting the Web to Procure Secure Information for Enterprise

Ritu Punhani

Asstt. Professor, Institute of Information Technology,
Amity University, Noida, U.P, India

Alpana Kakkar

Deputy Director, Institute of Information Technology,
Amity University, Noida, U.P, India

Dr. D. Jain

HCL Technologies Ltd., India

International Research Journal
ISSN 1839-6518
Vol. 1, No. 1, 2011
82800101201105

Harvesting the Web to Procure Secure Information for Enterprise

Abstract

The potential that web resources have is yet unharnessed. The colossal information resources of the Web are largely untapped by the Enterprises. Enterprises has not yet seriously considered use of web content as a potential input to the data warehouse, even when the web has been proven to be the universal delivery mechanism for global data.

Many researchers have agreed that the paradigm of the Web is radically different than that of the data warehouse. One might also say that web content is highly volatile and diverse and hence harvesting the web to procure the desired information of real business value for an enterprise is like challenging its imagination.

As a chaotic and unmanageable influence, the Web can be perceived as a threat to the security and tranquility of the warehouse environment. A stable infrastructure is required. The issue of Integrity and reliability of web content is critical to the requirement of refining information through the processes of discovery, Validation, Acquisition, Design and Formatting, Dissemination, and Information Security. This paper proposes a process model along with architecture to support this methodology within a data centre.

Keywords

Mining, Harvesting the Web, Enterprise, Information Security, Web Content

Introduction

The potential that web resources have is yet unharnessed. The colossal information resources of the Web are largely untapped by the Enterprises. Enterprises has not yet seriously considered using Web content as input to the data warehouse, even when the Web is becoming the universal delivery mechanism for global data.

Doing business as usual is now a risky strategy in such competitive business environment. The conventional wisdom about any company is no longer valid. The objective of harvesting and procuring information from the Web is to enable a company to adapt and even thrive within these massive changes. Amid the chaos of the Web is a diversity of ever-changing information, some of which is critical to your business. The challenge is to wade with big boots through the Web, discovering and acquiring those information resources that have value to enterprise. One must move from an information refining process that is haphazard and wasteful to one that is systematic and productive.

Data warehousing systems are widely used successfully for systematic business intelligence. However, those systems only deal with data from internal operational systems. In the majority of data warehousing efforts, enterprises focus inward. As markets become turbulent, data from internal systems becomes less relevant to managing the business and planning for its future.

The business should be keenly alert to outside information. As management guru Peter Drucker argues, the challenge is "to organize outside data because *change occurs from the outside*". Drucker predicted that the obsession with internal data leads organizations to be blindsided by external forces [RIC99-1].

Discover, harvest, refine and procure information from sources external to your organization and utilize this information to impact your key business processes. Valuable information about external business factors is readily available on the Web and is becoming more so each day. [MID11]

Web harvesting is not surfing the Web haphazardly, wandering from one intriguing item to another nor is it a one-time search of the Web. On a continuous and systematic basis, this process model of harvesting the web must deliver the right information to the right people at the right time in the enterprise. The business model of Information Security must aid the KPA managers to use the information harvested and acquired from the web, designed and formatted for input to the data warehouse and disseminated among enterprise KPAs with complete Confidentiality, Integrity and Availability norms.

Harvest Reliable Information

Business analyst requires advanced search indexing tools for harvesting the information from the Web. The specific objectives of this architecture are:

- To discover web content that is highly relevant to the enterprise [MID11].
- To validate the discovered content and ensure that data inserted into an application satisfies defined formats and other input criteria.
- To acquire the content so that it is properly validated within a historical context [BOL03].
- To Design and formatting the content into a useful form that's compatible with the data warehouse. The objective of Data Warehousing is to bring together information from disparate sources and put the information into a format that is conducive to making business decisions. Data Warehousing requires both business and technical expertise and involves the following activities:

- Accurately identifying the business information that must be contained in the Warehouse
- Identifying and prioritizing subject areas to be included in the Data Warehouse
- Managing the scope of each subject area which will be implemented into the Warehouse on an iterative basis.
- Developing a scalable architecture to serve as the Warehouse's technical and application foundation, and identifying and selecting the hardware/software/middleware components to implement it
- Extracting, cleansing, aggregating, transforming and validating the data to ensure accuracy and consistency
- Defining the correct level of summarization to support business decision making
- Establishing a refresh program that is consistent with business needs, timing and cycles
- Providing user-friendly, powerful tools at the desktop to access the data in the Warehouse
- Educating the business community about the realm of possibilities that are available to them through Data Warehousing
- Establishing a Data Warehouse Help Desk and training users to effectively utilize the desktop tools
- Establishing processes for maintaining, enhancing, and ensuring the ongoing success and applicability of the Warehouse

- To disseminate the content to the proper people so it has direct and positive impacts on specific business processes [BOL03]
- To manage the previous steps in a systematic manner as part of the production operations of a data center environment [BOL03]
- To provide Information Security, an ongoing process of exercising due care and due diligence to protect information, and information systems, from unauthorized access, use, disclosure, destruction, modification, or disruption or distribution. The never ending process of information security involves ongoing training, assessment, protection, monitoring & detection, incident

response & repair, documentation, and review. This makes information security an indispensable part of all the business operations across different domains to provide reliability, confidentiality, integrity and availability of information [KNI11].

The Data Warehouse Concept and Web Farming

First of all a formal definition of the term data warehouse seems appropriate: "... a data warehouse is a subject oriented, integrated, non-volatile and time variant collection of data in support of management's decisions" [INM02]. The core of a data warehouse is a data base, in which data from different operational systems are historically saved in different levels of aggregation. Due to the fact that, as a rule, analysts make complex inquiries and demand intuitive working with the database, a multidimensional data model seems appropriate. [FEL03] Now combining the concept of data warehousing and web farming, the main objective is to refine web content in a systematic manner. To do so requires a discipline to transform raw data into validated information -- much like a farmer transforms seed into a harvest. With web farming, refining the content involves the processes of discovering, validation, acquisition, design & format, dissemination and information security.

Discovery is the exploration of available Web resources to find those items that relate to specific topics. Discovery involves considerable "detective" work far beyond searching generic directory services (such as Yahoo) or indexing services (such as AltaVista). Furthermore, the discovery activity must be a continuous process because data sources are continually appearing (and disappearing) from the Web. A business analyst is the central figure in this activity and requires advanced search and indexing tools to be productive [RIC99-1].

Validation is to validate the data for functionality, quality, integrity, availability, scalability and security based on the defined business requirements by an organization. It may be necessary to ensure that the data is completely valid. If the data is not valid, the integrity of the business analysis relying on the data may be compromised. For example, a value representing a monetary transfer between banks in different countries must be in the correct currency. Data should be validated at the source by business analysts who understand what the data represents. Validating data can be a time-consuming process. The validation process can be automated by writing stored procedures that check the data for domain integrity. However, it may be necessary to validate data manually. If any invalid data is discovered, determine where the fault originated and correct any processes contributing to the error [RIC99-1].

Acquisition is the collection and maintenance of content identified by its source. The main goal of acquisition is to maintain the historical context so you can analyze

content in the context of past changes. Acquisition requires a secured server platform with large storage capacity [RIC99-1].

Design & Format is the analysis, validation, and transformation of content into a more useful format and into a more meaningful structure. The formats can be Web pages, spreadsheets, word processing documents, and database tables. As we move toward loading data into a warehouse, the structures must be compatible with the star-schema design and with key identifier values.

Dissemination is the packaging and delivery of information to the appropriate consumers, either directly or through a data warehouse. It requires a range of dissemination mechanisms from predetermined schedules to ad hoc queries. Newer technologies such as information brokering and preference matching may be desirable [RIC99-1].

Information Security requirements are specified in the requirements specification. They should specify who should have access to the information and who should be allowed to use it. The term *web farming* doesn't imply that valuable information exists somewhere on the Web, just waiting to be found and immediately used. Instead, it implies that hard work is involved to prepare the field, seed the crops, cultivate the soil, and then finally harvest the crop. The value of web farming comes from applying effort over time and with patience to the information resources of the Web. Cultivating a few seeds of data will eventually produce a harvest of information. If you are farming the Web's information resources, what specific information from that huge vastness should you farm? Obviously, just as for any other aspect of your enterprise information architecture, you'll want to concentrate on those information clusters that currently do or potentially can have the most bang for your business's buck.

Putting It All Together

As Figure 2 shows, the data warehouse occupies a central position in the information flow of a web farming system. Like operational systems, the web farming system provides input to the data warehouse. The result is to disseminate the refined information about specific business subjects to the enterprise.

The primary source of content for a web farming system is the global Web. This source can be supplemented (but not replaced) by an enterprise's intranet. Intranet content is limited to internal information about the enterprise, such as internal web sites, word processing documents, spreadsheets, and email messages [RIC99-2]. Regardless of its source, most information acquired by the web farming system will not be in a form immediately suitable for incorporation into the data warehouse. It

will either be unstructured hypertext or unverified tabular values. In either case, you must perform the refining process before loading the information into the warehouse.

As Figure 3 shows, a robust web farming system spans a variety of roles:

1. One or more information analysts control the activities of web farming through the discovery, acquisition, structuring, and dissemination processes. The information analysts will occasionally probe the Web and distribute information to specific individuals; however, their primary focus is on activities within the data centre, which is where most processing is performed. The databases for control metadata, content (in various stages of refining), a staging area, and the data warehouse are all managed environments.

2. The person programming agents creates algorithms for searching and structuring web content based on the control information accumulated by the analysts.

3. The data administrator designs and supervises the web content's flow into the data warehouse.

4. The system administrator ensures the security and reliability of the overall systems.

Web Farming results in practical management and technical skills for implementing effective business intelligence systems within your company.

What to Farm

In contrast to *web mining*, the term *web farming* does not imply that valuable information exists somewhere on the Web, just waiting to be found and immediately used. Instead, it implies that hard work is involved to prepare the field, seed the crops, cultivate the soil, and then finally harvest the crop. The value of web farming comes from applying effort over time and with patience to the information resources of the Web. Cultivating a few seeds of data will eventually produce a harvest of information. If you are farming the Web's information resources, what specific information from that huge vastness should you farm? Obviously, just as for any other aspect of your enterprise information architecture, you'll want to concentrate on those information clusters that currently do or potentially can have the most bang for your business's buck.

To understand the concept of Web Farming [BOL03-2], consider the old story about the farmer. Two metropolitan folks are surveying the farmer's beautiful farm. Several times they say to the farmer, "God and you sure did a great job with this place."

Finally, the farmer pauses in thought and replies, "Yes, but you should have seen it when God had it all to Himself."

The term web farming is cute, almost unprofessional in its tone. Not to be confused with personal experiences of surfing the Web (i.e., mostly frustration with a touch of elation), web farming has a serious side that draws on its metaphor to agriculture [BOL03-1]

Many people falsely think that web farming is like planting a small garden in the backyard. In contrast, Web farming is like managing a large agricultural concern that involves many people and several thousand acres of farmland. Further, this business is to be managed to financial and productivity objectives, so that the results generated are of proven value.

For a successful implementation of a web farming system, a six-stage methodology is recommended. Each stage builds upon the previous, with the goal of blending the web farming activity within the data warehouse and eventually within the knowledge management system for the enterprise.

Select the Seeds by instituting the business case based on the objectives and business environment of the enterprise

Move to Official Version by legitimizing the web farming activity within the organization and by building infrastructure for production operations

Connecting Smartly by exploiting technology especially for discovery and structuring of information to build information-pipes

Check-In for Validation by validating the information before passing to the information pipelines reaching data warehouse

Hitting the Bull's Eye by structuring information for the data warehouse by revisiting the business objectives in light of the warehouse schema

Harvest the Fruits by publishing the content of value to the organization in lieu of desired objectives.

The proper architecture is critical to the success of a web farming system. The most difficult part of web farming is the rendezvous with the data warehousing system. Many people pursue data warehousing systems for simplistic reasons and with unrealistic expectations. These systems often become 'black holes' into which data is poured never to be seen again.

Both the Web and data warehousing are hot technologies receiving considerable attention within the IT industry. In many areas, the combination has proven highly

successful. However, no one has seriously considered extracting content from the Web and using it as input to the data warehouse. Reactions to using web content tend to be negative. Web content is supposed to be too unreliable and unstable for business decisions. The interactions with web sites are usually too messy. Transformation of hypertext into a structured database is often impossible. Images and sound contain a lot of hidden content but are not discernible to a machine.

Now let's assume you are an IT Analyst at BigMart (a hypothetical retailer) who has decided to build a sales data mart as the first step in rolling out comprehensive analytics. In discussions with the sales department IT Analyst figured that the no. of units sold, dollar amount of sales, and the number of unique customers in a segment are the main metrics they look at. Going deeper you figure that the sales guys are likely to want few analysis categories that are analysis by product, product category/class, brand, store location (city, state, region, country etc.), customer demographics, analysis by brand and also by individual promotions and promotion categories.

Web farming would be valuable by enhancing the brand dimension about products giving mode specific reflection brand choice of the customer. By adding information on brand dimension, you can perform more enhanced analysis and deliver the information on analysis by brand. By knowing what types of customers buy what types of product brand at which stores, we can promote specific brand sales and anticipate demand. Now that you have figured out the sales of BigMart metrics to be measured, gives you the facts you would need in the data mart 'fact table' for calculating them. The 'fact table' linked to the 'dimension tables' makes up the 'star-schema' (because of the star-like structure)

Correspondingly there are 6 dimension tables joined to the fact table through foreign keys in the star-schema.

The dimension tables in turn have detailed data that can now be used for defining ad-hoc queries for analysis segments. For example, we can put demographic filters on the customer dimension (say status - married/ unmarried, age<=29, working/ non-working, college-educated), choose specific product class(es) in the product table (say Dairy Products), specify brand in brand dimension (say Amul Dahi, Amul Milk, Mother Dairy Dhai, Mother Dairy Milk….), specify a limited time period, and then get our metrics calculated for the ad-hoc segment [AKS11].

The Figure 7 shows the detailed information available in the dimension tables for defining ad-hoc segments.

Conclusion

Web farming can be successfully done, reaping tremendous value for the business and bypassing the data warehouse entirely. However, establishing the Web farming function is much easier for an enterprise if it has a mature understanding of data warehousing. In many ways, the current benefits from data warehousing are "low-lying fruit" -- easy accomplishments (relatively speaking) of purging the sins of monolithic legacy systems.

Web farming would challenge enterprises with deeper issues concerning information refinement and knowledge management. Web farming will be an agent of change (even of a disruptive sort) to the controlled and structured world of data warehousing. This is a necessary change -- a maturing of the basic objectives of data warehousing into a practical step toward knowledge management for the enterprise.

Figures and Tables

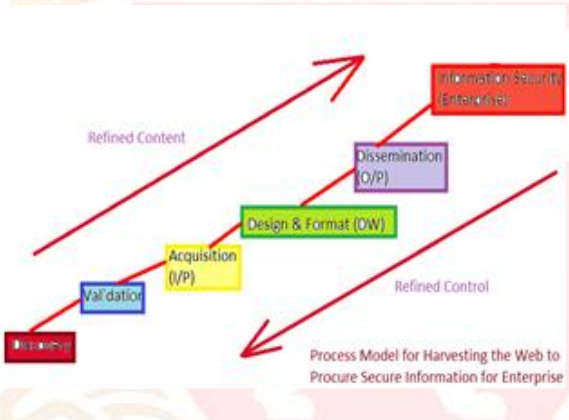

Figure 1: Refining of Web Content

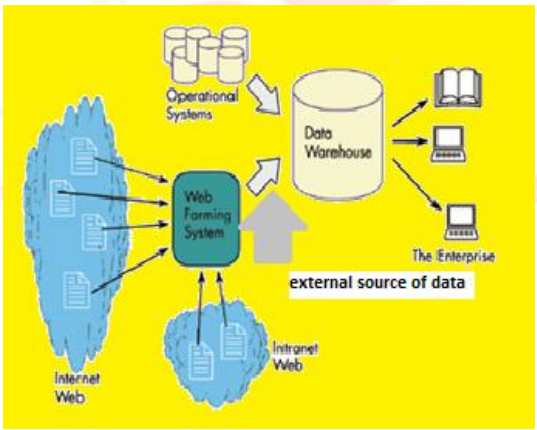

Figure 2: Web Farming System

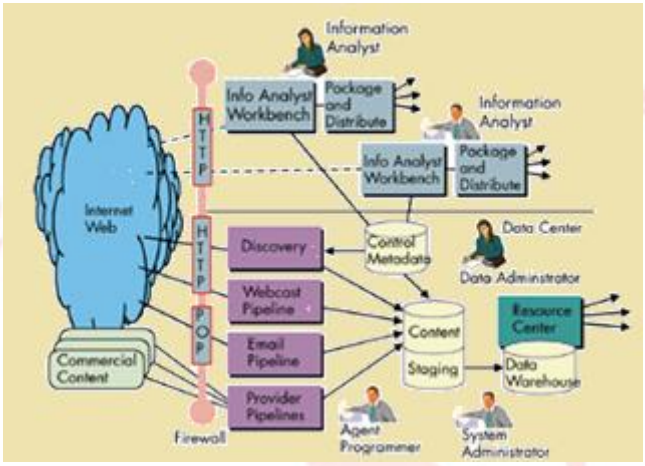

Figure 3: Robust Web Farming System

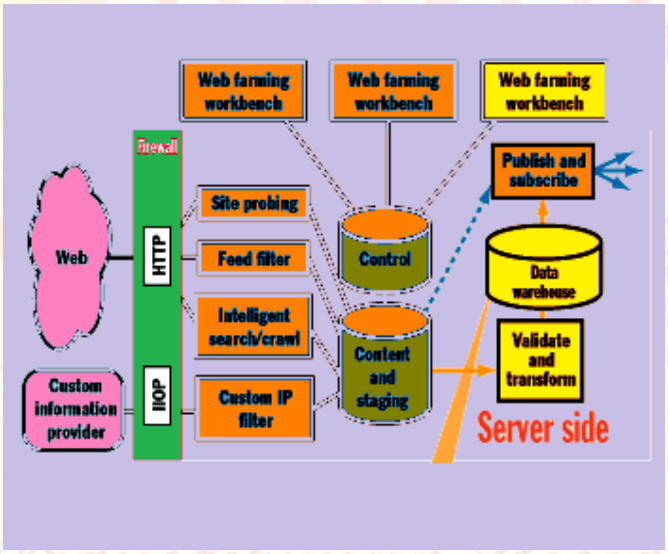

Figure 4: Web farming System Architecture

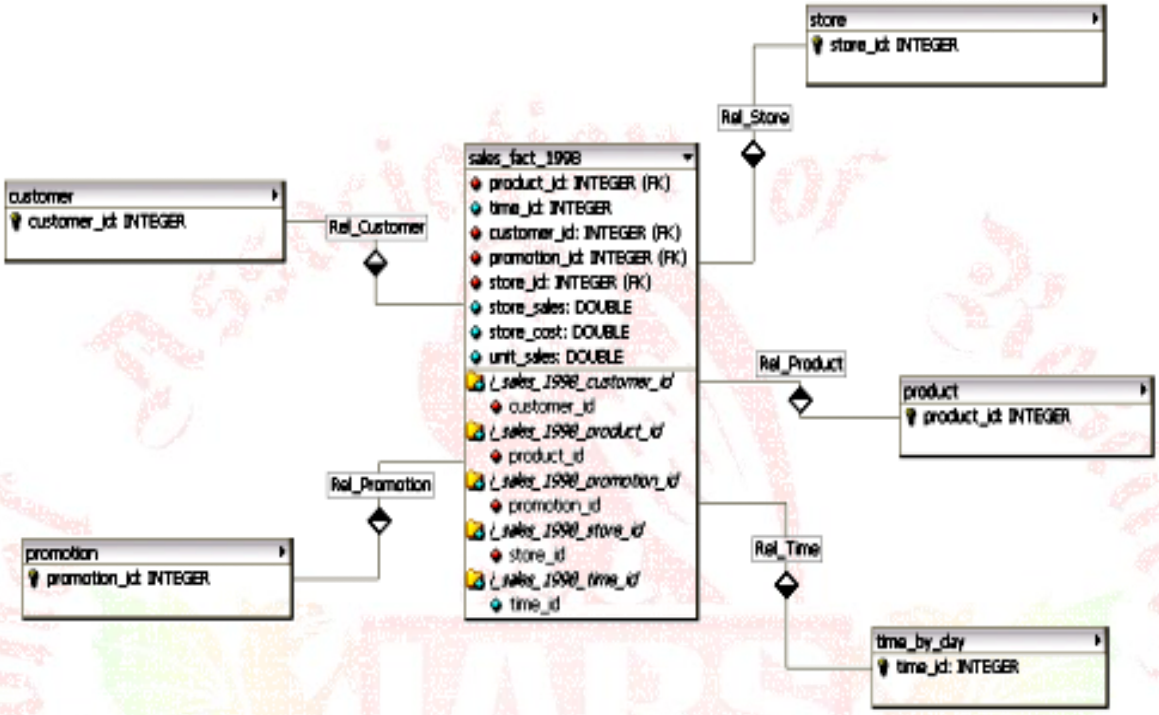

Figure 5: Star Schema for BigMart

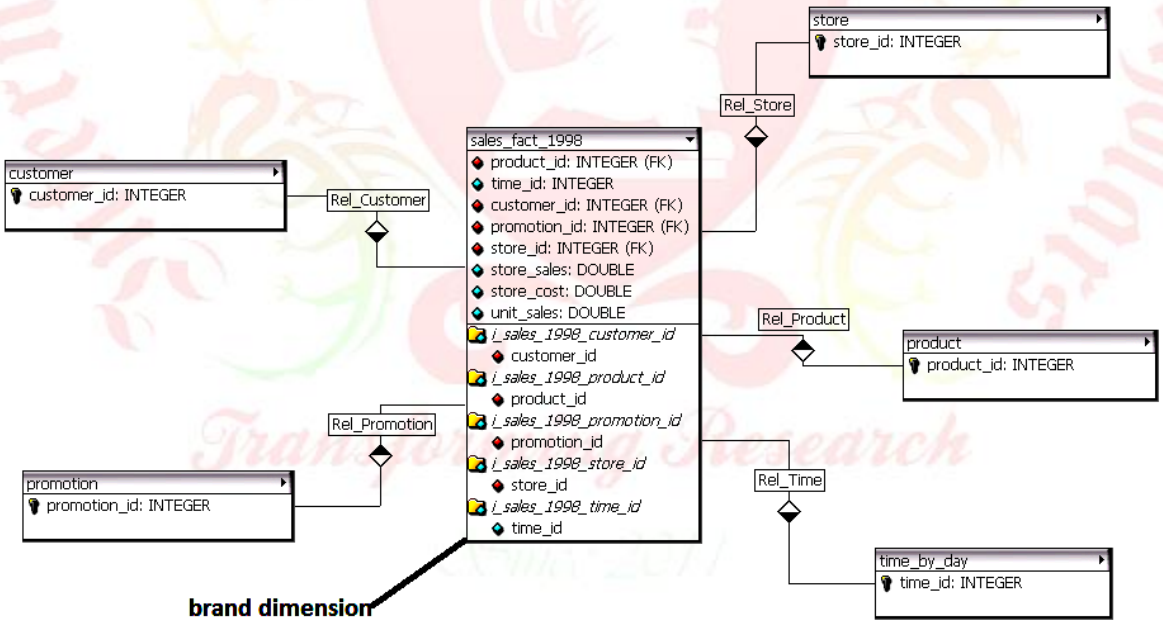

Figure 6: High level Dimensions model for sales in BigMart

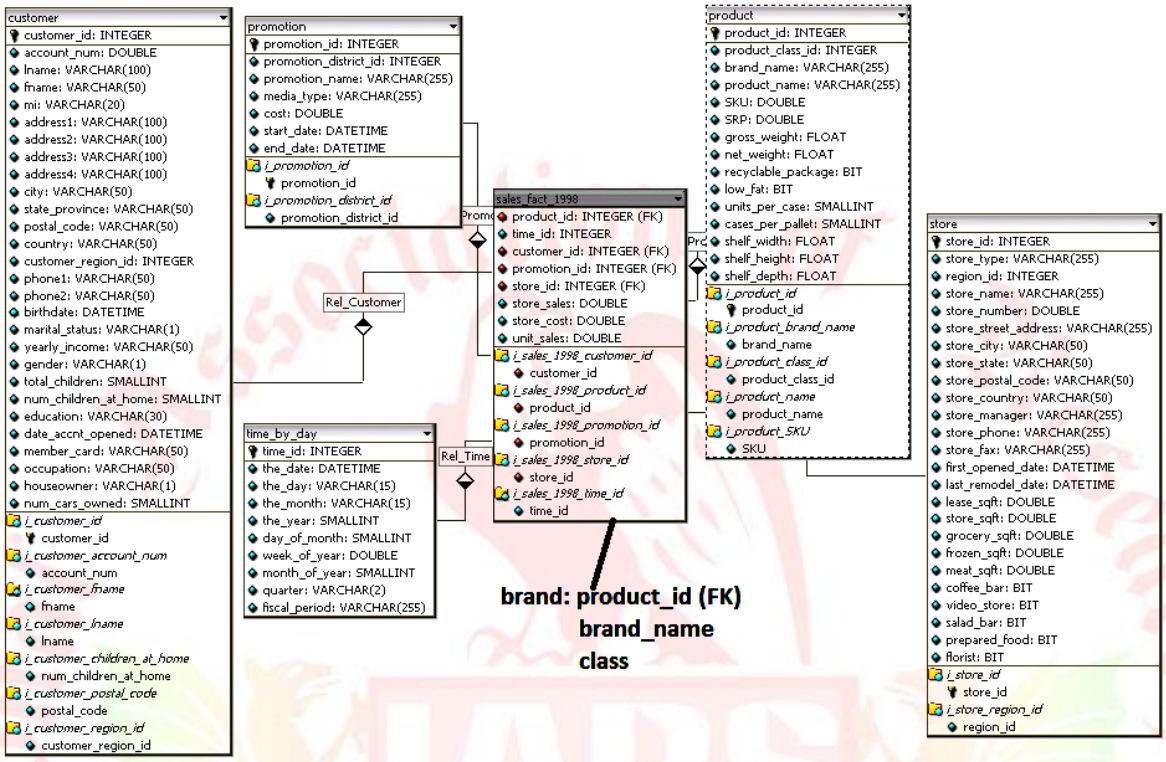

Figure 7: Defining ad-hoc analysis system

References

1. Bolder Technology Inc. DBA, “Its Farming, Not Mining”, 2003, , WebFarming.com
   DOI: http://webfarming.com/intro/intro03.html [BOL03-2]

2. Bolder Technology Inc. DBA, The Agriculture Metaphor, WebFarming.com, 2003, DOI: http://webfarming.com/intro/intro04.html [BOL03-1]

3. Bolder Technology Inc. DBA, WebFarming.com, 2003, DOI: http://webfarming.com/intro/intro02.html [BOL03]

4. Felden, Carsten, Peter Chamoni, “Web Farming and Data Warehousing for Energy Tradefloors”, WI '03 Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence, IEEE Computer Society Washington, DC, USA ISBN:0-7695-1932-6, 2003 [FEL03]

5. Hackathorn, Richard, "Farming the Web - The Web's content can be harvested for information that's crucial to making strategic decisions", BYTE Magazine – Core Technologies, 1997. [RIC97]

6. Hackathorn, Richard, "Farming Web Resources for the Data Warehouse", Information Management, 1999,
DOI: http://www.information-management.com/issues/19990601/1001-1.html [RIC99-1]

7. Hackathorn, Richard, "Web Farming", DB2 magazine online, 1997
DOI: http://www.webfarming.com/intro/DB2mag.pdf [RIC99-2]

8. Hackathorn, Richard, DBMS - Reaping the Web for Your Data Warehouse, 1998. [RIC98]

9. Inmon, W. H., Building the Data Warehouse, 3rd Edition, New York, 2002 [INM02]

10. K-Ninety East Africa Limited, Information Security, 2011, DOI: http://www.k-90ea.com/consultancy.html [KNI11]

11. Lamont, Judith, "Innovative applications make government more responsive", KMWorld Volume 12, Issue 6, 2003 [JUD03]

12. Markov, Zdravko, Daniel T. Larsoe, "Data Mining the Web – Uncovering Patterns in Web Content, Structure and Usage", John Wiley & Sons, INC. Publication, page-59,143,156, 2007. Midriff Net Solutions, "Web Farming", 2011 DOI: http://midriff.in/webfarm.asp, [MID11]

14. Vyas, Akshay, "Sales Data Mart – Dimensional Model for Retail", 2011 DOI: http://aadityainc.blogspot.in/2011/10/sales-data-mart-dimensional-model-for.html [AKS11]

15. World Wide Web consortium, DOI: www.w3.org/TR/REC-rdf-syntax

– END –

International Association of Research Scholars

IARS

Transforming Research

www.iars.info

Certificate of Recognition

This certificate is awarded to

Ritu Punhani

in recognition of his / her contribution

"Harvesting the Web to Procure Secure Information for Enterprise"

to Vol. 01, No. 01, 2011 *of*

IARS International Research Journal

An International Refereed Research Journal

ISSN 1839-6518 (Australian ISSN Agency)

08 FEB 2011

Editor in Chief

International Association of Research Scholars

IARS

Transforming Research

Since 2011

www.iars.info

Certificate of Recognition

This certificate is awarded to

Alpana Kakkar

in recognition of his / her contribution

"Harvesting the Web to Procure Secure Information for Enterprise"

to Vol. 01, No. 01, 2011 *of*

IARS International Research Journal

An International Refereed Research Journal

ISSN 1839-6518 (Australian ISSN Agency)

Editor in Chief

IARS International Research Journal 08 FEB 2011