WEB-BASED DATA MINING SERVICES
A Solution Proposal

Serban Ghenea and Cornelia Oprean
Omnis Group, 6 Barbu Vacarescu, Bl. 2, Apt. 1, Bucharest 020281, Romania

Keywords: Data mining, web utilities, SQL Server 2005.

Abstract: The paper presents the results obtained in building a web-based solution that provides to registered users, accessing a portal on the Internet, the possibility to perform complex business analysis tasks, using data mining algorithms and services implemented by Microsoft SQL Server 2005 Analysis Services. The database platform sustains the web operation of a complete ERP system, offering support for back-office management and establishing a B2B environment that automates collaborative business processes.

1 INTRODUCTION

The large-scale evolution of the Internet and of web-related technologies made possible the creation and deployment of innovative tools that turn once sophisticated and extremely complex software solutions into ubiquitous resources, accessible as traditional pay-per-use utilities.

The vast majority of B2B portals and e-marketplaces available today on the Internet are configured to span every segment of one specific industry (vertical eMarkets) or are able to connect buyers and sellers across many industries (horizontal eMarkets). Existing solutions give users the possibility to publish general information and trade leads, facilitating business contacts that may generate commercial transactions, further to be sustained and managed using external IT system support.

The paper describes the capabilities of a web-based platform that advances this concept, with the purpose to establish a unique solution that supports marketing operations and commercial transactioning, providing also integration with the ERP back-office support. Users may access, over the web, solutions for the management of the following enterprise activities: financial-accounting, procurement, sales and invoicing, fixed assets, human resources, documents workflow, legal reporting to authorities, on top of standardized catalogs of partners, activities and products.

ERP operational data can further be analysed using SQL Server 2005 data mining algorithms and services, implemented under the same concept of web utilities.

The solution is able to draw businesses away from managing their own IT infrastructures and enables them to subscribe to information services, under a concept similar to clasic utilities consumption (water, electricity, gas, cable TV), saving time and money, reducing administrative overhead and answering to the lack of IT specialists.

2 SOLUTION ARCHITECTURE

The platform, hosted in a Data Center, may be accessed on the web by any Internet user. Following the registration process, according to the profile of each registered company, a company database is created and specific configuration settings are automatically generated and applied.

Figure 1 shows a possible configuration of the solution platform, that contains several SQL Server 2005 database servers and one or more Windows 2003 Web Servers, working together in a failover clustering architecture.

The Participants database holds index information on all registered companies, whereas the Configuration database stores the personalized profiles that are created by the participants in the portal, after registration. One client ISIS .NET database is created for each company that subscribes to the ERP services.

The solution benefits from a service-oriented architecture, implementing services that can further

Ghenea S. and Oprean C. (2007).
WEB-BASED DATA MINING SERVICES - A Solution Proposal.
In *Proceedings of the Second International Conference on Software and Data Technologies - Volume ISDM/WsEHST/DC*, pages 203-208

be used by other services or other applications. The solution includes units of service-oriented processing logic, such as products-catalog services, marketplace offers services, purchase order services.

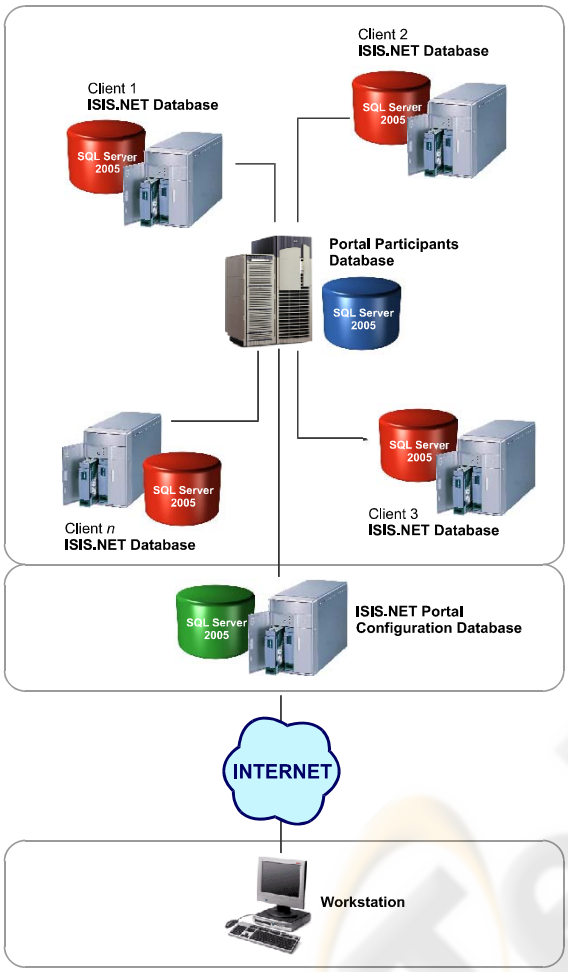

Figure 1: Platform Infrastructure Configuration.

2.1 Available Services

An anonymous user on the Internet is able only to browse the general information stored in the portal – company profiles, product and service offers.

A registered company may use the B2B portal marketing services, may act as a buyer or seller on the eMarket, may employ workflow services and manage the flow of commercial documents, may access ERP specific functionalities, may carry on business analysis tasks or may operate any combinations of the above.

The web interface allows access to the entire set of available services. The platform core is represented by the ERP engine, that sustains all operations. There are separate solution sections in the interface for each service category, as shown in Fig. 2. The data mining services work directly with the database, while all the other services make use of the business layer built around ERP functionalities.

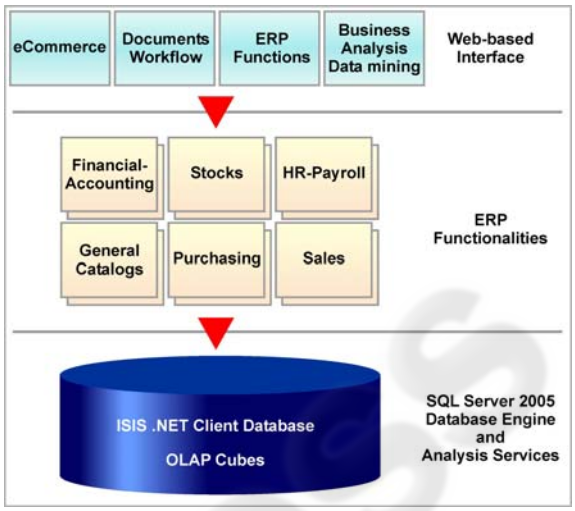

Figure 2: Solution Configuration and Available Services.

2.2 Solution Implementation

The equipment used in experiments consisted of:

- client computer Intel Pentium III at 1.13 GHz with 512 MB RAM, running Windows XP Professional with SP2;
- server running SQL Server 2005 on 2 X Xeon 3GHz, with 4GB RAM and 2 X 160 GB mirrored hard disks, under Windows 2003 Server.

The solution was developed using the Microsoft latest technologies: Windows Server 2003, Windows Vista, SQL Server 2005 Database Engine and Analysis Services (SSAS), Visual Studio 2005, ASP.NET 2.0, Net Framework 3.0 (Windows Workflow Foundation, Windows Communication Foundation, Windows CardSpace).

The user interface benefits from the user experience innovations introduced by Windows Vista, such as:

- Aero Theme and System Font (Segoe UI);
- Windows SideBar gadgets;
- Windows Vista tone in all UI text.

On the B2B section of the portal, the workflow services ensure the management of the documents flow between the portal participants, taking

advantage of the following benefits offered by the Windows Workflow Foundation framework:

- The workflow scheme is graphically designed in a workflow designer, the administrator being able to dynamically configure the workflow process;
- The workflow is created as a state machine workflow, based on defined events;
- The solution allows usage of rule conditions in workflows and ensures dynamic update of these rule conditions;
- The solution benefits from using transactions and fault handling for workflow management.

In order to implement the service-oriented architecture (SOA) concepts, the solution uses the Windows Communication Foundation (WCF) framework. The solution implements services for products, offers, requests-for-quotes, order and tender documents, using a layered service architecture.

The Service Interface Layer defines the operations provided by the service, the messages required to interact with each operation, and the patterns by which these messages interact.

The Business Layer incorporates components that implement the business logic of the service.

The Data Access Layer contains the logic necessary to access data, as well as specific service agents.

To secure SOAP messages that flow between the service and the client, the solution uses Windows CardSpace technology, in order to replace the password-based authentication, thus eliminating the "phishing" threat, specific to Internet communication.

3 DATA MINING SERVICES

The solution offers access to ERP services (commercial, stocks, financial, accounting, payroll, transportation management) for the registered companies that subscribe to these services. After the registration is validated by a general administrator, a client ISIS .NET database is created for the accepted company. The data mining services provide advanced business analysis capabilities, based on data mining algorithms implemented in SQL Server 2005. The data subject to analysis resides in the ISIS .NET client databases.

3.1 Data Source

The examples shown in detail below are obtained from the client databases of two companies that operate in the hospitality industry. Users having access to the Data Mining Services of the portal are able to select the type of data analysis to be applied and the appropriate viewer solution.

3.2 Data Preparation

During the testing phase, appropriate data cleansing and modification processes were applied, in order to eliminate records with zero quantities, to solve data inconsistencies due to the text representation of the Country field (duplicates and misspelling) and to decompose the Date field in 3 separate fields – year, month, day. In the production phase, the data preparation process will be automatic.

New Analysis Services projects were created using Business Intelligence Development Studio and appropriate data source were added in order to connect to the ISIS .NET client databases. The data mining algorithms were run on views created for testing purposes. The views include only the tables that store sale documents, customers and profit center data. The analysis was limited to the period 2004-2007. In total, 43,050 records were processed, among which 23,519 records in the documents table of one database and 19,434 records in the documents table of the second database. An OLAP cube was created in each database, defined on the dimensions *Customer_Country* and *Date*, using the measure *Amount*, representing the value specified on the sales document.

3.3 Data Mining Results

The SQL Server 2005 Analysis Services generated a set of results that were displayed using Microsoft Excel and the viewers that are provided by SQL Server 2005 Analysis Services for each distinct data mining method that was applied: OLAP cube processing, time series, clusters, association rules.

3.3.1 OLAP Cube

The cube views shown in Fig. 3-6 were generated in Microsoft Excel, using a pivot table built based on the cube definition.

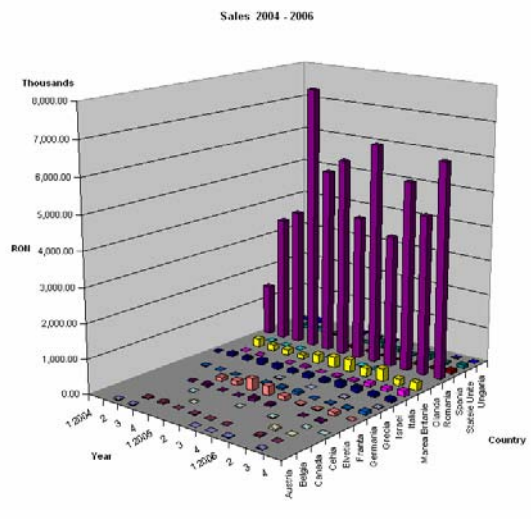

Figure 3: Sales Chart per Country, Client1.

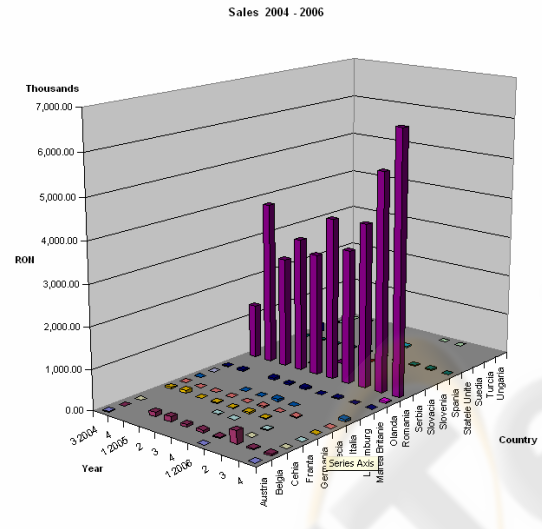

Figure 4: Sales Chart per Country, Client2.

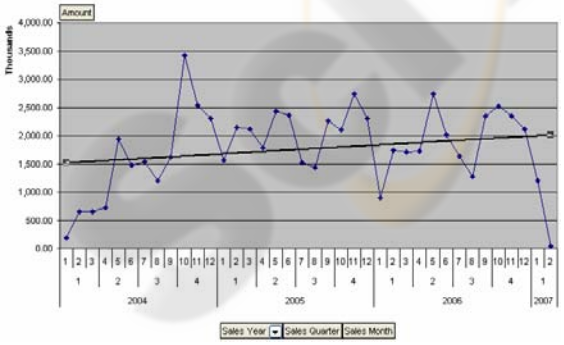

Figure 5: Sales Trend Line, Client1.

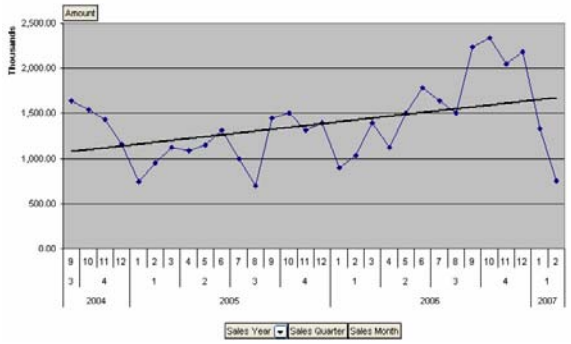

Figure 6: Sales Trend Line, Client2.

3.3.2 Time Series

The SSAS Time Series Viewer displays models that are built with the Time Series algorithm. This is a a regression algorithm for use in creating data mining models to predict continuous columns, such as product sales, in a forecasting scenario. The results obtained on the two databases are shown below. The chart in Fig. 7 display the behavior of the time series, together with the predicted values for the future. Fig. 8 shows the decision tree built based on the predictable attributes.

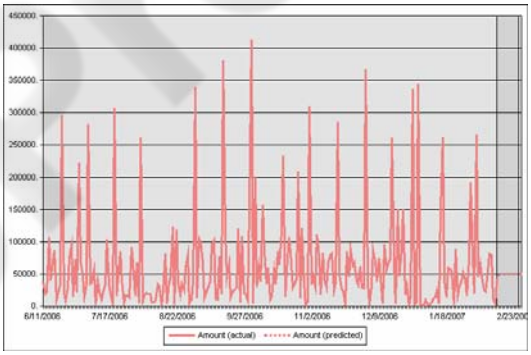

Figure 7: Sales Time Series Chart.

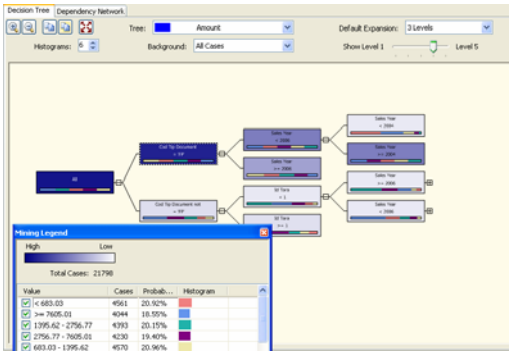

Figure 8: Decision Tree.

3.3.3 Clusters

The Sequence Clustering algorithm is a sequence analysis algorithm for use in exploring data that contains events that can be linked by following paths, or sequences. The Microsoft Sequence Cluster Viewer displays all the clusters identified in a mining model. The shading of the line that connects one cluster to another represents the strength of the similarity of the clusters. If the shading is light or nonexistent, the clusters are not very similar. As the line becomes darker, the similarity of the links becomes stronger.

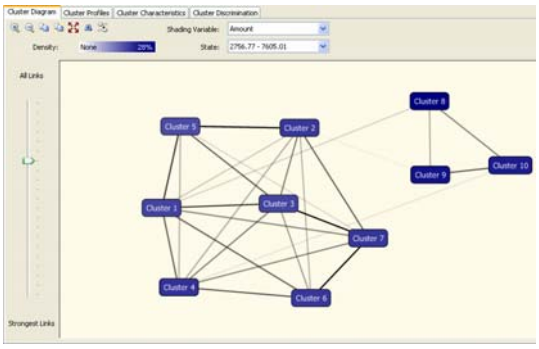

Figure 9: Sales Clusters Diagram.

The viewer provides, in Fig. 10, an overall view of the clusters built in the model. Each column that follows the *Population* column in the grid represents a cluster discovered by the algorithm. The rows show the composition of the clusters for each attribute in the model. Distinct ranges of attribute values are represented in distinct colors.

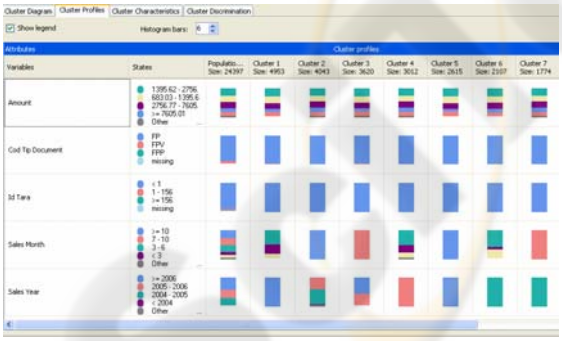

Figure 10: Sales Clusters Profile.

3.3.4 Association Rules

The Microsoft Association algorithm is an association algorithm for use in creating data mining models that can be used for market basket analysis. The algorithm finds itemsets that describe items that are typically found together in a transaction and discovers rules that predict the presence of other items in a transaction, based on existing items. The viewer displays the list of itemsets that the model identified as frequently found together. The tab displays a grid with the following columns: *Support*, *Size*, and *Itemset*. In Fig. 11, the largest itemsets include the records indicating sales to Romanian customers (Country_ID=1) with support value 22,124 and records corresponding to sales invoices, with support value 21,798.

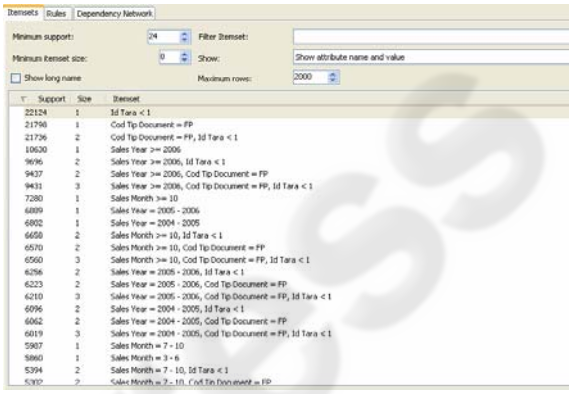

Figure 11: Association Itemsets.

The dependency network viewer displays nodes, each node representing an item. The arrow between nodes represents the association between items. The direction of the arrow dictates the association between the items according to the rules that the algorithm discovered. For example, in Fig. 12, an arrow points from the node *SalesYear=2005-2006* towards the node *Amount>=7,600*. The slider at the left of the viewer acts as a filter that is tied to the probability of the rules. Lowering the slider shows only the strongest links.

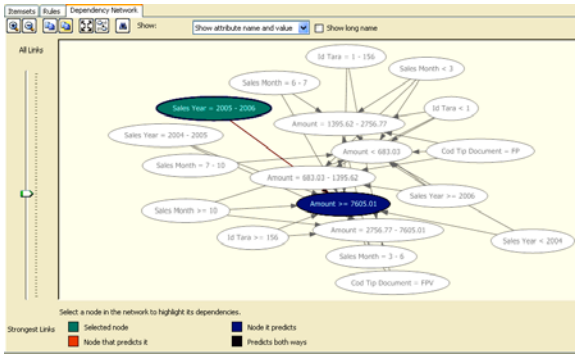

Figure 12: Association Dependencies.

4 CONCLUSION

The ISIS .NET platform provides a web-based solution that includes a B2B portal for collaborative business processes and an ERP system for specific back-office operations. Installed in a data center, the platform offers to registered users the possibility to administrate their complete business environment using advanced technology, with minimum costs. The operation of an unique ERP solution ensures that all data are consistently stored in the same database structures. The solution developers are therefore able to create an unique set of business analysis tools, based on Microsoft SQL Server 2005 Analysis Services, that can further be operated similarly by all system users.

The data mining platform must further be developed to provide distinct sections for the analysis of data subject to the operation of various ERP modules: financial-accounting, stock management, transportation, human resources, etc. No matter the business specifics, companies that administrate stocks using the ISIS .NET Stock Management module - that stores data in the ISIS .NET client database - will be able to share the same business analysis tools, included in the solution platform. As the databases grow and the models develop in time, it will become necessary to embed in the platform industry-strength data mining facilities. For better efficiency, at clients' requests, the system will generate separate data warehouses, starting from the ISIS .NET client databases, that will support the data mining tasks. The platform will include all the tools for the back-office management of the data warehouses: periodic data loading, data preparation, back-up and archiving, administration and interpretation of data mining results.

The described solution architecture complies to the current trends of establishing powerful nodes that provide modern utilities and services using standardized delivery networks, similar to the situation of traditional utilities. Water, electricity, gas, cable TV, Internet access, games, music and movies, office and application software, data and knowledge will follow similar distribution flows, built using consistent standards. Concentration and consolidation will be improved, to allow sharing out at a global scale.

REFERENCES

Cornelia Oprean, Serban Ghenea, 2007. Delivering IT Services as Web Utilities. In *Proceedings of 16th International Conference on Control Systems and Computer Science*, vol. 2, pag. 169-174.

Serban Ghenea, 2007. Application Tool for Experiments on SQL Server 2005 Transactions. In *WSEAS Transactions on Computers*, issue 2, vol. 6, pag. 223-228.

Ian H. Witten, Eibe Frank, 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Second Edition.

Margaret H. Dunham, 2003. *Data Mining, Introductory and Advanced Topics*, Prentice Hall.

Thomas Erl, 2005. *Service-Oriented Architecture – Concepts, Technology and Design*, Prentice Hall.

Thomas Erl, 2004. *Service-Oriented Architecture – A Field Guide to Integrating XML and Web Services*, Prentice Hall.

ZhaoHui Tang, Jamie MacLennan, 2005. *Data Mining with SQL Server 2005*, Wiley.

Laurence Moroney, *"The Windows Communication Foundation: A Primer"*, http://www.devx.com/dotnet/Article/29414/0

Juval Lowy, *"What You Need To Know About One-Way Calls, Callbacks, And Events"*, http://msdn.microsoft.com/msdnmag/issues/06/10/ WCFEssentials /default.aspx.

Aaron Skonnard, *"Serialization in Windows Communication Foundation"*, http://msdn.microsoft.com/msdnmag/issues/06/08/ ServiceStation/default.aspx.