Data Integration and Warehousing in Telecom Italia

Stefano M. Trisolini
Telecom Italia
Data Administration,
Data Warehouse, Data Mining
Stefano.Trisolini@telecomitalia.it

Maurizio Lenzerini, Daniele Nardi
Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
Via Salaria 113, 00198 Roma (Italy)
lenzerini@dis.uniroma1.it, nardi@dis.uniroma1.it

Abstract

We discuss the main methodological and technological issues arosen in the last years in the development of the enterprise integrated database of Telecom Italia and, subsequently in the management of the primary data store for Telecom Italia data warehouse applications.

The two efforts, although driven by different needs and requirements can be regarded as a continous development of an integrated view of the enterprise data. We review the experience accumulated in the integration of over 50 internal databases, highlighting the benefits and drawbacks of this scenario for data warehousing and discuss the development of a large dedicated data store to support the analysis of data about customers and phone traffic.

1 Data Integration

Telecom Italia is the main Italian provider of national and international telecommunication services, and the 5th world-wide company in its field. It is active in more than 40 countries, and it has 90 thousands employees. The goals of Data Administration, Data Warehouse, Data Mining are: enterprise data modeling, providing infrastructure for data access, physical design and data distribution, data warehousing, data quality and security, and data mining.

In 1993, Telecom Italia has launched a strategic project called IBDA, with the following main goals:

- the definition of the Enterprise Data Model and the migration/evolution of existing data;
- the design and implementation of databases covering the main sections of the Enterprise Data Model (Customers, Suppliers, Network, Administration, etc.);
- the design and implementation of the client/server architecture and the communication middleware;
- the design and implementation of data access services.

The driving motivations of the IBDA strategic project are typical of a scenario where there is a proliferation of legacy databases with a large overhead in the design and maintenance of software for interfacing applications and provide access to the data. IBDA was based on a staged implementation of services that form a layer separating data from the application processes. More specifically, the IBDA service for a database is the exclusive agent that provides access to the data. The access is actually accomplished through contracts that enforce the semantic and referencial policies for the database.

In 1996, the project was formally completed with the integration of 48 operational databases, while in the subsequent years new databases have been continuosly added to IBDA. At present, 52 databases are included in IBDA and there are undergoing projects for adding more.

2 Data Warehousing

In Telecom Italia, data warehousing has been a natural evolution of data integration. Starting from the standpoint of a large integrated enterprise database, and given the size of the data to be stored in the data warehouse, the architecture of the Telecom Data Warehouse includes a Primary Enterprise Data Warehouse, which feeds the data to several layers of aggregation that are typically present before the data become available to the final user (Secondary Data Marts).

An integrated data store provides a solid basis for the design of Data Warehouse Applications: many of the problems that arise for data warehousing are anticipated (data extraction, cleaning, and reconciliation), thus providing good quality data for analytical processing. Moreover, Data Warehouse Applications can be developed in a more coherent way, because of the existence of the integrated data store.

SIGMOD '99 Philadelphia PA

Nonetheless, there are several important questions that are not addressed by the work on integration. More specifically, integration does anticipate semantic problems with data, but does not address efficiency issues, that are critical for data warehousing. Thus, to guarantee a proper performance of the data warehouse, a major re-organization of the data store may be required, with additional costs.

As a consequence of the layered approach to data warehousing taken in Telecom Italia, a strong emphasis is given to methodologies. Telecom Italia has devised proprietary methodologies and tools for incremental schema and data integration, which provide a basis for data acquisition in Data Warehouse Applications. Telecom Italia is also devising proprietary methodologies, for the development of applications, that are based on the cooperation with the top management, iterative refinement and rapid prototyping. Such methodologies need to take into accout both the complexity and the internal organization of roles and responsibilities of Telecom Italia. In the design of methodologies, and their supporting tools, Telecom Italia has benefited from the experience gained by the Dipartimento di Informatica e Sistemistica of the Università di Roma "La Sapienza" within the Esprit Research Project on Data Warehouse Quality [CDGL+98, JLVV99].

Currently, our Primary Enterprise Data Warehouse includes the following subsystems:

- *Interconnection traffic*, containing call detail records (CDRS), whose main purpose is to analyse network usage patterns between Telecom Italia Network Nodes and other service providers. The daily loading is 40 millions CDRS and the store contains 6 months history data.
- *Customer*, containing information on Telecom Italia products and services by customer, to analyse customer data. The daily loading is about 150 GB and the average size is 1.5 TB.
- *Voice traffic*, containing additional information on CDRS records from PSTN switches, to support various analysis tasks for Marketing, Administration and Customer Management. The daily loading is 130 millions of CDRS and the store contains 12 months history data.

3 Summary and Lessons

We have sketched the data integration and data warehousing activities in Telecom Italia. We summarize the experence gained in the developments of these projects as follows.

1) *Integration is an incremental process*, both because, from the organizational point of view, it is unlikely to be able to merge all the data sources in one step, and because the Data Administrator needs to incorporate new data sources that continuously become available.

2) The *construction of a data warehouse is an incremental process* as well, especially in a scenario where a data warehouse structure, including several levels of organization, is adopted, as it is the case in enterprises of the size of Telecom Italia.

3) The incremental nature of the approach to both data integration and data warehousing deeply influences the design of *methodologies and tools* supporting these tasks.

4) All the *Information Technology partners are potentially involved* in data warehouse projects (Operations, Development, Data Management), and this impacts on the complexity of such projects.

5) A *dedicated structure to manage data warehouse implementation* is fundamental, prior to any implementation.

6) *Scheduling of extraction/transformation jobs* is critical and custom development must be carefully planned. Moreover, 10-15% of the transformation tasks require ad hoc development.

7) *Clustering technology* gives good results on data warehouses up to 2TB; beyond this threshold, a specific feasibility study is needed.

Acknowledgments

We wish to thank Massimo Biondi, Diego Calvanese, Patrizia Cannuli, Giuseppe De Giacomo, Riccardo Rosati, Daniela Segreto, Valeria Stancati, for their contribution to the activities on Data Integration and Data Warehousing at both the Dipartimento di Informatica e Sistemistica of the Università di Roma "La Sapienza", and Telecom Italia. We also thank all the partners of the Esprit Project DWQ [JLVV99].

References

[CDGL+98] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi, and Riccardo Rosati. Information integration: Conceptual modeling and reasoning support. In *Proc. of the 6th Int. Conf. on Cooperative Information Systems (CoopIS-98)*, pages 280–291, 1998.

[JLVV99] Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis. *Fundamentals of Data Warehouses.* Springer-Verlag, 1999. See also *Foundations of Data Warehouse Quality*, ESPRIT LTR Project No. 22469 DWQ, `http://www.dbnet.ece.ntua.gr/~dwq`.