Railway Data Warehousing System for Booking Data Analysis

Carlo DELL'AQUILA, Ezio LEFONS, and Filippo TANGORRA
Dipartimento di Informatica
Università di Bari
via Orabona 4, 70126 Bari
ITALY

Abstract: - The analysis, design, and implementation of the data warehouse system for the decisional process based on the Italian train booking data are presented. Trenitalia, the Italian main train service company, is the customer, and TSF (railway telesystems company) the IT solution provider. In particular, the feasibility requirements, functionality, technical architecture, and product technology are described. Moreover, the guidelines about operational environments interfacing with the data warehouse for data acquisition and elaboration, and related problems are dealt with. With our contribution and the aim of software reuse, the provider has released the prototype system, and, in order to satisfy all the customer's requests, the entire warehouse's data marts concerning the project will be completely re-engineered.

Key Words: - Data mart, data warehouse, train booking, railway transportation analysis.

1 Introduction

Data warehouses provide historical, summarized data directly for decision support system applications. They can be used also as source for building *data marts* which contain information for specific departmental sectors. The metadata repository stores information on data sources, access mechanism, user login and data mart schemas [2,3,9,10,12].

The warehouse can be implemented divided into two layers: Operational Data Store (ODS) and Enterprise-wide Data Warehouse (EDW).

The ODS stores tactical data from production systems that are subject-oriented and integrated to address operational needs. The detailed, current information in the ODS is transactional in nature, updated frequently (at least daily) and only held for a short period of time. It integrates information from the production systems, relieves the production systems of reporting and analysis demands and provides access to current data. The goal of ODS is to provide a tactically-structured, efficient information processing environment to satisfy the analysis and reporting capabilities required for the day-to-day operations of the business.

The EDW stores data from all subject areas within the business for analysis by end users. Its scope is the entire business and all operational aspects within the business. An EDW is normally created through a series of incrementally developed solutions. It provides a single source of corporate enterprisewide data, a single source of synchronized data for each subject area, and a single point for distribution of data.

Most data warehouses use a Staging Area (SA) to process and clean the data coming from outer data sources. The SA simplifies the summary building and the warehouse management. In order to customize the warehouse architecture for the different needs of end users, Data Marts (DMT) can be used.

A Data Mart is a subset of data warehouse facts and summary data that provides users with information specific to their requirements. DMTs are designed for a single line of business: whereas data warehouse typically assembles data from multiple source systems, Data Marts assemble data from fewer sources. Therefore, they are often smaller, less complex and easier to build and maintain than data warehouses. Data Marts are "read-only" objects, they need to be periodically updated from the source data, and they can be implemented as objects of the warehouse itself (dependent DMTs) or as outer systems (independent DMTs).

In this paper, the analysis, design, and implementation of a case study will be presented. Trenitalia, the Italian main train service company, is the customer and TSF (*TeleSistemi Ferroviari*, railway telesystems company) the IT solutions provider. The customer makes his requests through the Passengers Division Marketing Structure. In particular, the technical offer for the re-engineering of the Transaction Processing Facility booking data component of the Data Warehousing System "DWHPAX" (Passengers Division) is discussed.

2 Decision support system architecture

We present the feasibility requirements, functionality, architecture, and the product technology description.

Moreover, the guidelines about the specific operational environments interfacing with the data warehouse for data acquisition and elaboration are dealt with [4]. With the aim of software reuse, the provider has realized the first project release, but in order to satisfy all the customer's requests, all the warehouse's Data Marts concerning the project will be completely re-engineered.
The architecture of the decision support system consists of three main components: (1) extract, transform, and load tools (ETL tools), (2) data warehouse, and (3) analysis tools.
In Figure 1, the architecture of the Italian railway system for the booking process is shown.

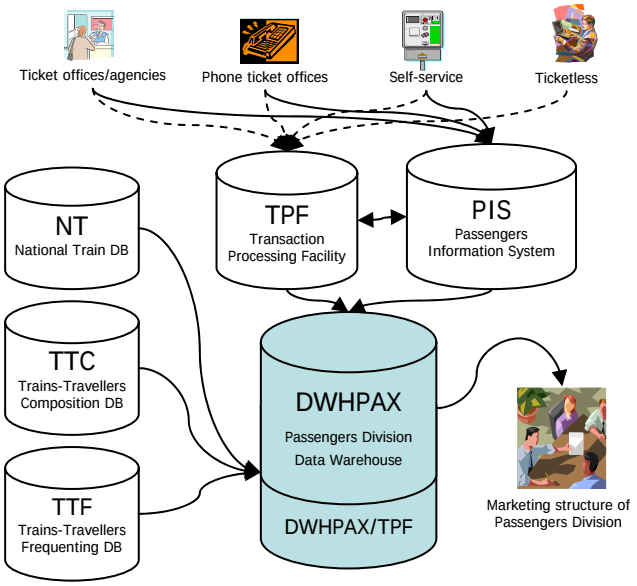

Fig. 1. The Italian railway system architecture.

Such a structure accords with the management of the usual, various levels of data flow corresponding to the sequence of steps for adapting the data to the decision-maker needs. That is to say, the source level, the refresh level, the warehouse level, and the analytic level [6-9].

3 The source level

Since data warehouses contain data consolidated from several operational databases, the integration of data coming from different heterogeneous sources is necessary. The source data are stored on relational or legacy databases that are components of information systems owned by the decisional organization or external organizations.
The data warehouse contains data deriving from many sources, typically other systems or flat files. The metadata and raw data of the Online Transaction Processing (OLTP) systems are present, and there is an additional kind of data, the so-called Summary Data. Metadata are data about data; they contain, for example, information about tables and fields names. Summaries, called Materialized Views in Oracle, are very useful because they pre-compute long operations in advance.
The source data flows dealt with by the Transaction Processing Facility (TPF, see Figure 1) come from the Sale System of the operational Passengers Information System (PIS) and from the set of information concurring in the booking analysis realization. Besides PIS/TPF, the other operational systems involved are: the National Train database (NT), which contains the train route and railway kilometres; the Trains-Travellers Composition database (TTC), which contains saleable delivered services, associated trains and antenna trains (coupling/release); and the Trains-Travellers Frequenting database (TTF), which contains the train registry.
Booking data come from the Technologies and Systems Structure & Sale Area Platform, into *two* daily sequential datasets containing booking for the present day and booking for the next 120 days (four months), respectively. Other useful data are directly extracted from TTC and TTF databases, and from the NT component, or indirectly from the data warehouse itself.
Some problems, reported below, can arise on source data flows because of both the impossibility of comparing data coming from heterogeneous sources and systems, and possible data gaps due to incomplete data. Some of the problems appeared in the customer requirement analysis phase, while others have been detected during the technical analysis.

a) The more relevant problem descends from the customer's request to bring the associated train bookings to the master train one. In fact, for all those master trains with associated leaving from midnight, the daily TPF data are incomplete, because the booking data will be consolidated only when integrating them with the next day data. Thus, both daily booking data and booking/offer data of the data mart will be accessible one day later.
b) The same problem as the precedent (a), occurs for the TTC/TPF mapping. Mappings will be complete for all those programmed trains managed by TTC and TTF whose classification, product type, and source/target are known; extraordinary, charter, religious, and other nonprogrammable trains are excluded.

c) There is dissimilarity between delivered services extracted from TTC and those coming from TPF files. TPF data and TPF data till 120 days (four months) about completely foreign routes will not be considered, because they are not assignable to Italian trains. Similarly, "mixed" routes, i.e., those with source or target in a foreign station, will be limited to the national territory. However, original data will be maintained on the warehouse for possible special analyses. The TPF 120 flow contains booking data for each leaving train date (from the present to 120 days), but there is no information about the booking issuing date. Day by day data are related to bookings made for train/date, train/start and train/end booking stations.

d) The determination of the medium charge and load factor are exact only for trains with booking required. In fact, for trains with optional booking, these measurements give conservative estimates.

e) Information about train recurrence and composition (of cars) is stored into TTC in textual form, which is automatically unmanageable. Thus, the corresponding information interpretation is impossible to be extracted.

f) The birth of one new travellers-stop on the train route will not determine the recalculation of the old elementary (now compound) route values.

4 The refresh level

At this level, the ETL tools allow the decisional data administrator to extract, to transform, to load, to refresh, and to integrate data from the several source data described by different data schemas in the data warehouse. From a technological viewpoint, the solutions of distributed services for information systems are adopted, such as, for example, the management of the data inconsistency and the structure incompatibility.

The ETL process is fundamental to the creation of quality information in the data warehouse. Data are taken from different source systems, then they are cleansed, verified, validated, converted into a consistent state, and then moved into the warehouse. The Extraction phase is the process of selecting specific operational attributes from the various operational systems. The Transformation phase is the process of integrating, verifying, validating, cleaning and time-stamping the selected data into a consistent and uniform format for the target databases. Rejected data are returned to the data owner for correction and reprocessing. The Transportation phase is the process of moving data from an intermediate storage area into the target warehouse database.

We now present the ETL process on DWHPAX (Data Warehouse Passengers Division), the data warehouse of Trenitalia, the major Italian railway industry, whose database is implemented with Oracle 9i [1,5,11,13].

Figure 2a exemplifies the first source of the ETL process, that is, a flat file.

```
2004-08-12104   81 2851 81 7331 51 0    2  0   0   0   0   0   N
2004-08-12104   81 2851 81 5707654 0    3  0   0   0   0   0   N
2004-08-121092  80 1106880 3036 80 0    5  0   0   0   0   0   N
2004-08-12115   80 1548580 2034780 0    1  0   0   0   0   0   N
2004-08-12118   81 1187 81 2903480 0    1  0   0   0   0   0   N
2004-08-12119   80 1545880 1187 81 0    1  0   0   0   0   0   N
2004-08-12119   80 1902280 1187 81 0    1  0   0   0   0   0   N
2004-08-12120   80 1031680 58   84 0    1  0   0   0   0   0   N
2004-08-121235  81 2851 81 1700 83 0    0  12  8   0   2   0   N
2004-08-12126   80 1031680 58   84 0    2  0   0   0   0   0   N
2004-08-121288  80 5071 83 2034780 0    2  0   0   0   0   0   N
..............  .. .... .. ....... .   .  .   .   .   .   .   .
```

Fig. 2a. Example of flat file coming from the TTC database.

Flat files come into DWHPAX typically from other systems belonging to Trenitalia, like PIS/TPF (booking data), TTC (trains composition data), TTF (trains-travellers data), and TN (train route and rail kilometres data). The Oracle utility that loads data from flat files to the database tables is SQL*Loader. It loads data according to a control file positional method, like that shown in Figure 2b. Each row of the flat file is divided into 15 columns, the first going from the 1st to the 10th character, the second from the 11th to the 15th, and so on. Then, the columns are loaded into the fields of the table `CMM_ODS_TPF_TPF_G60` as described in the control file.

```
LOAD DATA
TRUNCATE
INTO TABLE CMM_ODS_TPF_TPF_G60
(
DAT_RGT         "trunc(sysdate)"
DAT_TRE         POSITION(1:10) "TO_DATE(:DAT_TRE,'YYYY-MM-DD')",
NUM_TRE         POSITION(11:15) ,
COD_RET_TRE     POSITION(17:18) ,
COD_STA_INI     POSITTON(20:24) ,
COD_RET_INI     POSITION(25:26) ,
COD_STA_FTN     POSTTTON(28:32) ,
COD-RET_FIN     POSITION(33:34) ,
PSI_PRI_PRE     POSITION(36:39) ,
PSI_SEC_PRE     POSITION(40:43) ,
CUC_PRI_PRE     POSITION(44:47) ,
CUC_SEC_PRE     POSITION(48:51) ,
CUC_CMF_PRE     POSITION(52:55) ,
PSI_LTT_PRE     POSITION(56.59) ,
PSI_TAA_PRE     POSITION(60:63) ,
TIP_PRE         POSITION(64:64)
)
```

Fig. 2b. Example of control file against the flat file of Figure 2a.

5 The data warehouse level

Data warehouse implementation requires tools for end user access. The choice of the tools depend upon the

user's requirements for information. The tool complexity varies from simple reporting tools to complex OLAP tools, to highly advanced data mining tools.
Acquisition and population functions work on several distinct elaboration environments:

- Client/server environments, where the TTC and NT databases are resident on; and
- Data Warehouse server environment running on a UNIX machine, interacting with other systems, where the Passengers Division warehouse database resides on and where the Frequentation database, from which the Train Registry will taken, is currently resident on.

The TTC environment produces, through SQL-Server procedures, the flat files that will be loaded in the Data Warehouse server environment through FTP procedures. Similar procedures load the NT data into the Data Warehouse server. These procedures are designed and realized with the double purpose of (a) retrieving all the previous data about associated trains which lie in the interested period and (b) gradually reducing and discharging of the Passenger's Division host, and they replace the current host procedures.
At the same time, all information about Train Registry will be moved from FTV database to the Data Warehouse server. These data, together with those coming from other source systems, will be managed using procedures written in:

- SQL*Loader, for loading on the operational data store layer, and
- PL/SQL, for loading on the Enterprisewide Data Warehouse and Data Mart layers.

Other data this component needs will be directly taken from the Data Warehouse database.
Information about Delivered Services, Associated Trains and Antenna Trains is loaded from TTC source system. Train Routes, with respective validity dates and kilometre Number, are taken from the NT source system. Information about Train Registry is present on DWHPAX in the TTF instances.
TPF flow data (coming from PIS/TPF) include: train leaving date, train number, railway the train belongs to, source and destination station codes of the booking, source and destination railway codes of the booking, and total number of booked travels grouped by date, train, and seats/couchettes/beds each relative to the first, the second, and the comfort classes.
Since each TTC and each TPF record is associated with one generally composite source/target, and composition train and delivered services do not change, for each composite source/target, delivered services (TTC) and booking (TPF) will be assigned to every elementary route. This calculation, preparatory to the mapping, occurs in the TTC/NT environment and it is the input for the first loading in the operational data store layer.
The Train Registry is loaded from the TTC one, where special trains are defined, too: it will be merged with the PIS/TPF files and not known by TTF, thus it is not classifiable. This registry is daily updated in order to consider variations occurred (train number change, new special trains, registry modifications, etc). In this way, trains not classified before will be completed with the correct information (TTF classification, relation, source/target, etc) as soon as they are accessible.
TPF booking and TPF 120 booking files will be recalculated in order to bring associated train booking to the corresponding master train, with the exception that if the booking refers to an associated train leaving after midnight, then it will be brought to the master train of the previous day. Thus, booking data of the generic date "X" will be enforced on the "X+1" date. The structure is the same as the files containing a DAT-RGT (Registration Date) field, except for the TPF 120 table.
Moreover, in the TPF 120 table, there is the *Daily Delta* attribute which contains the count difference between booking data registered for each service in the date "R+1" and those in the date "R" when P and N are equal (R is the record registration date, P the train leaving date, and N the train number in question). This algorithm enables us to estimate the bookings of the date "R".
The booking pre-calculation previously described, will be done during the ODS to EDW loading phase (cf., Introduction). In the ODS layer, the structure is the same as the source system's one, and the data calculated will be mixed, for each elementary route, with those deducted from the Delivered Services contained in the TTC database. Booking will be again re-calculated in order to consider the Direct Services as well.
Also the seat types on TTC, the delivered services on TTF and the bookings on TPF, are not homogeneous. So, in order to integrate them, a new unique registry, whose occurrences will coincide with those of TPF, has been created, where TTF and TTC Ordinary Couchettes correspond to the Second Class Couchettes of TPF.
When the new schedule will come in effect, a portion of the database will be remade recalculating all TPF 120 data of the four months preceding the schedule change, both for the Enterprisewide Data Warehouse and the Data Mart layers. Analogous procedures and

recalculations will be done if a current schedule variation occurs.

6 The analytic level

At this level, sophisticated data analyses with OLAP and data mining techniques are used to accomplish reports, responses to complex queries, and results of simulation of new market scenarios. From the technological view-point, services are required for aggregate data management, query optimization, indexing structures and user-friendly interfaces.

Customer's needs about the TPF booking and TPF 120 booking components and about reporting are exemplified in the outputs summarized in the sub-sequent figures provided by the Passengers Division Marketing Structure.

Figure 3 shows the bookings daily trend graph for the issuing date, while Figure 4 refers to the leaving date, both relative to the period 1-15 April 2003 (dotted line) and 1-15 April 2004 (continuous line).

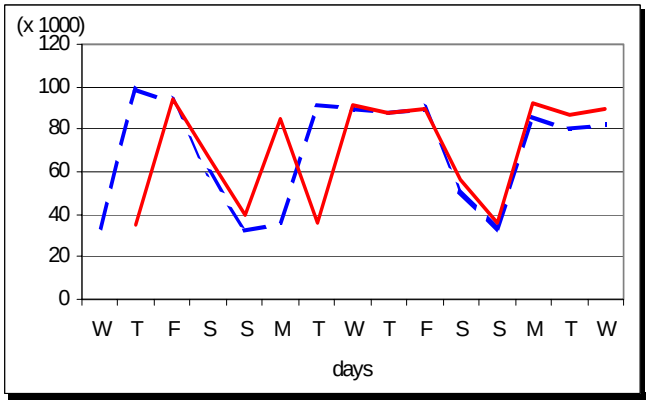

Fig. 3. Bookings daily trend graph (issuing date).

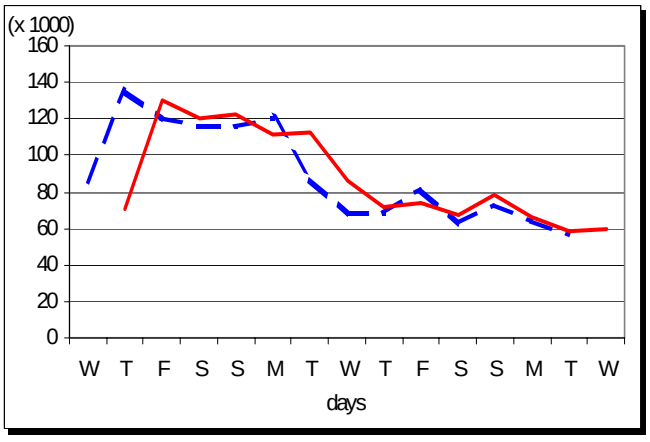

Fig. 4. Bookings daily trend graph (leaving date).

In Figure 5 is reported the bookings monthly trend graph of the four-month period Jan-Apr 2003 *vs* Jan-Apr 2004 (*Monthly Delta* %) relative to the issuing date (gray box) and leaving date (black box).

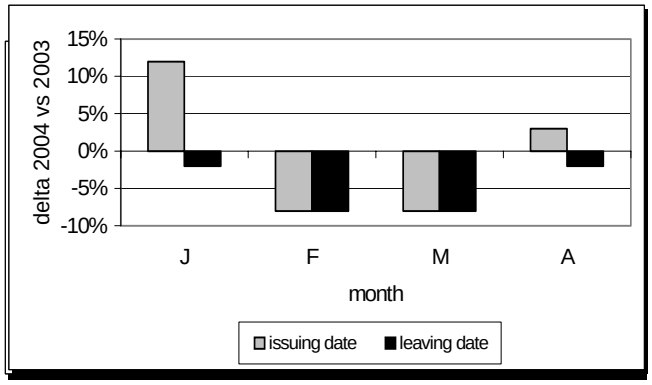

Fig. 5. Bookings monthly trend graph.

The customer also needs a directrix synthesis of the booking service trends. The provided solutions are based on the previous release's tested architecture, as much as possible.

A crucial component of the decisional activity is the summary data section in which a set of statistical indicators regarding the organization data warehouse gives information on the progress of the specific field for which it has been planned.

The indicators for the analysis are: Time, Train, Product, TTF Classification, Service Type, Service, Directrix, Relation, and Source/Target.

The metrics that the analysis and reporting phases will produce, which are self-explaining, are: Railway Km, Train × Km, Delivered seats × Km, Booked travels × Km, Load Factor (*LF*), Medium Charge, Booking monitoring, and Daily Delta, where

LF = (Booked _ travels × Km) / (Delivered_seats × Km).

7 Concluding remarks

In this paper the bases for successful data warehousing have been presented. Further considerations can be made only when designing a warehouse from scratch or when facing with a data warehouse for tuning or re-engineering aims. Materialized views used for query rewrite, for example, can be the key of data warehouse tuning, but only if they are well-written and they are part of well-designed warehouses.

The system we have utilized is owned by the railway telesystems company (*TSF*, *T*ele*S*istemi *F*erroviari) or IT provider, and it comprises:

- the use of the *TSF* IBM host calculus capacity for TPF data acquisition, storing, validation and loading;
- the centralized *TSF* Windows NT server for TTC and NT data acquisition;
- the centralized *TSF* UNIX server, hosting the data warehouse and containing Maintenance Areas, Train Production and Commercial data;
- the centralized *TSF* Windows 2000 professional server, hosting the Web server and the Micro-strategy's Intelligent Server (the reporting engine); and
- the user workstations, linked to these servers by the Geographical Network WAN to use the application.

Data warehousing can be useful for organizations producing or manipulating large or huge amounts of data that need to be suitably and profitably analyzed. In fact, it provides *ad hoc* analysis tools. Nevertheless, organizations attempting to embrace data warehousing must have a high IT maturity degree, because of the costs that data warehouse maintenance implies in terms of constant optimization due to the frequent changes of user's needs. We strongly suggest that organizations still working to meet their operational information needs should not embrace data warehousing: the risk that such organizations run is of increasing efforts for maintaining and re-designing cases together their relative costs.

References

[1] R. Baylis, K. Rich, and J. Fee, *Oracle 9i Database Administrator's Guide*, Release 1 (9.0.1), Oracle Corporation 2001.

[2] M. Boehnlein and A. Ulbrich-vom Ende, Deriving Initial Data Warehouse Structures from Conceptual Data Models of the Underlying Operational Information Systems, *DOLAP '99* Proc. *ACM*, pp 15-21.

[3] C. P. Chua and R. Green, *Data Warehousing Fundamentals*, Oracle Corporation 1999.

[4] L. Copertino, *Building a Successful Data Warehouse: Design, Implementation, and Tuning*, Thesis dissertation 2005.

[5] M. Cyran and C. Dialeris Green, *Oracle 9i Database Performance Guide and Reference*, Release 1 (9.0.1), Oracle Corporation 2001.

[6] S. Chaundhuri, U. Dayal, and V. Ganti, Database technology for decision support systems, *IEEE Computer*, Vol. 34, No 12, 2001, pp 48-55.

[7] C. dell'Aquila, E. Lefons, and F. Tangorra, Decisional portal using approximate query processing, *WSEAS Transactions on Computers*, Vol. 2, No 2, 2003, pp 486-492.

[8] C. dell'Aquila, E. Lefons, and F. Tangorra, Approximate query processing in decision support system environment, *WSEAS on Computers*, Vol. 3, No 3, 2004, pp 581-586.

[9] M. Janesch, *Implementing the Best Data Warehousing Tuning Techniques for Your Environment*, Innovative Consulting 2001.

[10] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, *Fundamentals of Data Warehouses*, Springer-Verlag, 2003.

[11] P. Lane, V. Shupmann, *Oracle 9i Warehousing Guide*, Release 1 (9.0.1), Oracle Corporation 2001

[12] H. Ong, *Data Warehouse Myths and Misconceptions*, Aurora Consulting 1999.

[13] L. McGeen Lusher, *Oracle 9i Database Concepts*, Release 1 (9.0.1), Oracle Corporation 2001.