DB-Prism: Integrated Data Warehouses and Knowledge Networks for Bank Controlling

Elvira Schäfer [1] Jan-Dirk Becker [1] Matthias Jarke [2]

[1] Global Technologies & Services, Deutsche Bank AG, Prior-Str. 11, 65936 Frankfurt, Germany

[2] GMD-FIT, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Abstract

DB-Prism is an integrated data warehouse system developed for distributed financial and management controlling (data collection, processing, and reporting) at Deutsche Bank. It combines fine-granular availability of historical data with high-actuality reporting and planning facilities. Major components of interest include an OLAP system in the Terabyte range, and a meta database responsible for the whole process from heterogeneous data selection to individual OLAP report positions and back via drill-through to the individual business activity. For these purposes, control workbenches have been developed both at the user –level and at the metadata level.

1 Introduction

The Controlling Warehouse System DB-Prism is a knowledge-driven network of database, processing, and reporting subsystems. Its full functional spectrum is currently used in the controlling of Deutsche Bank AG, the online bank DB 24 AG, and further German subsidiaries including DB Trust, DB Lübeck etc. The system is used in financial accounting, e.g. for creation of balance sheets and profit/loss statements, for information of the bank regulation agencies, and for daily internal status reports. In management accounting, the system is used e.g. for assessments of business units, business areas, and sales evaluation in the commercial banking sector.

The DB-Prism project was motivated by challenges concerning data quality factors [WSF95], such as integrity and transparency/traceability of data. Besides a heterogeneous hardware/software landscape, different principles of modelling and aggregation of multidimensional financial data were followed, targeted to the business goals of individual business units or simply resulting from history. Rule systems were not only heterogeneous but often also hidden in program code. Moreover, data often went through different parts of various controlling systems before ending up in a report. As a consequence of this semantic heterogeneity and knowledge structures hidden both in individual transformation programs and in the overall controlling workflow paths, analysts spent a lot of time reconciling seemingly contradictory information in reports.

The main design goal of DB-Prism has therefore been to clarify this heterogeneity at the conceptual, logical, and physical level, such that integrity of the data is improved, relationships between the different data source and reporting perspectives are made explicit (thus improving transparency and traceability), and reporting as well as refreshment of data are highly efficient. This has been achieved on the one hand by capturing sophisticated domain and systems knowledge about the information network in a metadata repository, on the other by employing and tailoring an efficiently scalable multi-dimensional OLAP system. As a result, heterogeneity is now consolidated into a uniform data warehouse environment based on a uniform historical business data store.

2 DB-Prism Architecture

The system has been tuned for efficient mass data processing and fast reporting. The database is refreshed with current changes on a daily bases, as well as the client-oriented sub-data cubes and reports. Fine-granular historical data are maintained in full detail for four years. The warehouse comprises 1200 GB, of which 250 GB are base data of the controlling warehouse (GDH) stored in

Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000

flat files and 150 GB in DB2 relations; note that these amounts reflect the relatively short operation history of the system and can be expected to grow further. The biggest part and the most important analysis instrument comprises 800 GB of interlinked data cubes stored as collections of sparse matrices in a multidimensional OLAP system called Matplan/b2brain [GMI00].

In the remainder of this paper, we describe the individual components of the DB-Prism architecture shown in figure 1. In the lower part of this figure, we show the basic Controlling Warehouse (GDH) which provides a common cleaned base store integrated from heterogeneous data sources, and the Controlling Workbench CBM which allows the Controller to operate on the GDH; more details can be found in section 3. In the middle of the figure on the right, we find the OLAP environment Matplan/b2brain which contains data cubes extracted and re-organised from the GDH according to various client interests, indicated by the Analyst in the middle; details are described in section 4. GDH/CBM with their underlying data sources (not shown) on the one hand, and OLAP-based analysis on the other, are interconnected in a metadata-driven manner. Metadata, defined and evolved in the CDR workbench for controlling dimensions and reports, are held in a meta database MetaDB. This is described in section 5.

In terms of the traditional data warehouse architecture, as described e.g. in [JLVV99], GDH/CBM roughly correspond to an integrated, cleaned, and historised, but not-yet-aggregated Operational Data Store, whereas Matplan/b2brain corresponds to a MOLAP warehouse and client environment. MetaDB/CDR comprise a significant enrichment of the metadata facilities available in other data warehouses, mostly due to the semantic richness and heterogeneity at the levels of concepts, logical data models, physical efficiency and safety.

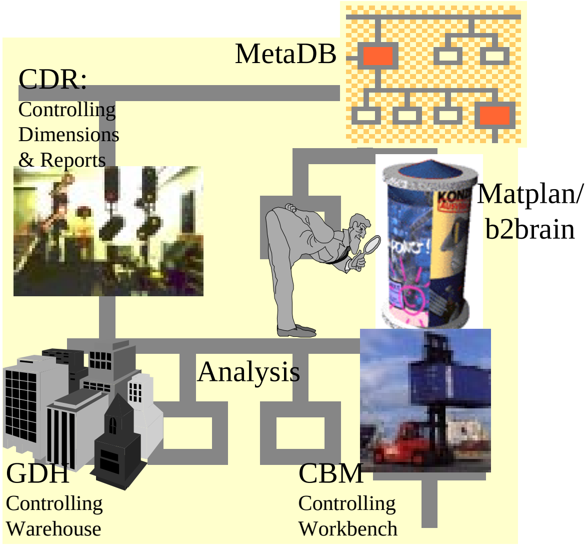


Fig. 1: DB-Prism Components

3 Basic Controlling Warehouse

As stated earlier, data from heterogeneous sources are cleaned, integrated, and historised in a Controlling Warehouse (GDH) supported by the CBM Controlling Workbench which does not only support generic versions of the above operations but also domain-specific operations for Financial and Management Controlling.

3.1 GDH (Controlling Warehouse)

GDH is based on detailed account-level information and is designed for mass data storage needs. GDH consists of five components: import, transformation, storage, access methods and export, following these key ideas:

- *Standard import and export interfaces.* Data are imported from many heterogeneous operative systems, statistics archives and controlling applications mainly through standard interface files. The delivery system does not need any special GDH know-how, as files are created exclusively by GDH modules. For calling up these modules, standardised file interfaces are available. They are generated from the MetaDB and are tailored to the individual delivery system. Also data export takes place only through standard interfaces which are supplied by GDH access.
- *Transformation and ‚cleaning‘* of import data achieve a consistent data base with *uniform data meaning*. Transformation comprises restructuring, redefinition, filtering, derivation, extrapolation and aggregation of data and data registers. Keys and data fields from heterogeneous delivery systems are harmonised, current and historical data undergo the same sort of structural processes.
- *Homogeneous storage of current and past data, methods of storage synchronised with access requests*. There are a variety of access requirements placed on the GDH: insertion of mass data, exchange and deletion of whole data groups, access to individual deals and data groups, processing of all individual deals. Such varied access requirements necessitate varying storage methods if the data availability and response times are to remain acceptable. This requires accepting redundancies in the data storage. As long as the integrity of the redundancies is inherently ensured within the system, such redundancies are useful.
- *Data access with ‚licensed‘ access modules only.* The GDH appears to the external users as a black box. Data access is provided by access functions. The encapsulation of the GDH strictly separates logical and physical aspects. This allows flexible data administration and storage with respect to extensibility of structures or changes in the physical data base.

Further important features include:

- *Possibility of flexible and efficient repetitions.* The GDH components for import, transformation, storage and export can be efficiently and flexibly, even retroactively, repeated. A repetition is executed down to the smallest delivery inventory level and all affected warehouse levels. Repetition needs are recognised and processed automatically.
- *Integrated workflow and control management.* The GDH programs and modules work universally through a common steering and control center. A universal MetaDB records all processing steps and status evaluations by subsequent programs. Thus, event- and result-oriented processing is achieved.
- *Scalability*. The structures in the GDH are constructed flexibly and permit subsequent expansion (variable container concept). Structural changes can therefore be implemented relatively quickly and cost-effectively on all relevant components. It is no problem to incorporate new GDH data inventories into the GDH administration.

3.2 CBM (Controlling Workbench)

CBM is a central workbench for controllers to work on the GDH. It supports additions of relevant information, corrections, interim account shifts and bookings, and plausibility checks. CBM therefore has access to the individual accounts and topic fields of GDH.

4 Matplan/b2brain (Multi-dimensional OLAP System)

Matplan/b2brain is the multidimensional database and OLAP system within the b2brain environment developed and marketed by GMI, a software house based in Aachen, Germany [GMI00]. Matplan/b2brain allows very flexible OLAP functionality (e.g. drill down and roll up between different dimensions and aggregation levels) and provides different application modules for planning, consolidation, data mining, data integration via XML, dynamic pricing & revenue management.

In DB-Prism, Matplan/b2brain is used as the standard instrument for reporting and analysis. Being an end-user task-oriented system, interactive, complex and multidimensional analyses can be carried out.

Matplan/b2brain is based on aggregated reporting data constructed from the Controlling Warehouse (GDH). Main features in the context of DB-Prism include:

- *full integration of Matplan/b2brain in the metadata concept*. All Matplan/b2brain report data taken from GDH are known in the meta database. Their dimensions, source, meaning, structure and composition are recorded. All metadata relevant for the structure of the multidimensional data cubes and all report definitions for data viewing are administered in the MetaDB.
- *Drill through functionality*. The user has the ability to call up as a trace for analysis the basis for the aggregated Matplan/b2brain report data, the GDH individual deals. For a focused analysis, selection criteria for data analysis can be defined (dimensions and their characteristics, cut-offs).
- *Scalability & storage capacity*. In the interest of unlimited analytical possibilities, it might appear useful to declare all dimensions as orthogonal, i.e. to define all data divisible intrinsically by all dimensions at all aggregation levels. However, this would require a storage capacity of 8.56 x 10^{26} cells; all the computers in the world would be insufficient. For this reason, there are several interlinked data cubes of various dimensions and scales. These data cubes constitute a logical hypercube such that all possibilities of multidimensional navigation (e.g. drill-down/ roll-up, slice & dice) remain fully available for the overall system.
- *Scalability & performance*. DB-Prism needs to move a very high volume of current data to make a fully refreshed data warehouse available every day. Matplan/b2brain allows flexible scaling by switching freely between pre-aggregation in batch windows, and dynamic online aggregation. The possible time delay associated with dynamic online aggregation is largely avoided by the usage of efficient hashing techniques.

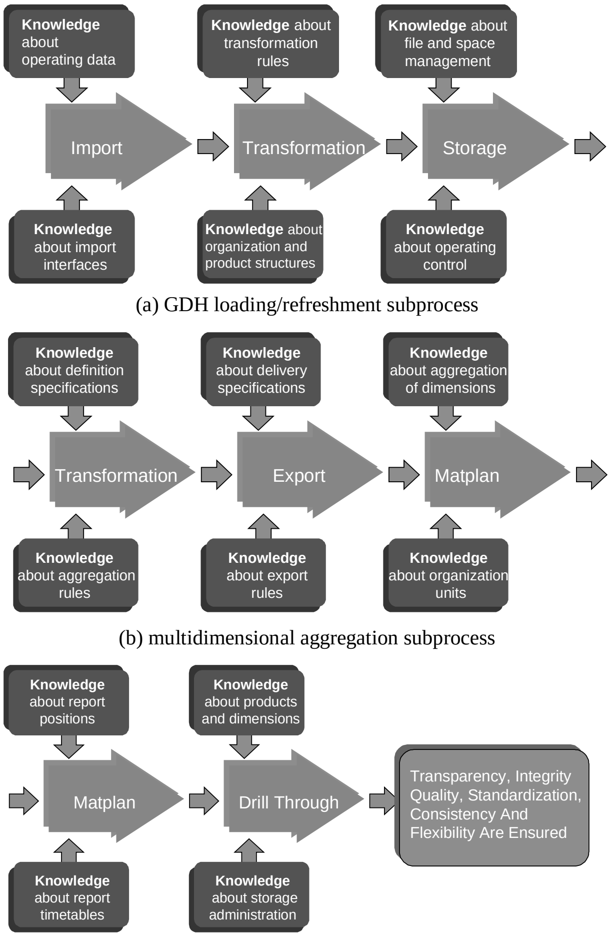
5 Metadata Management

As stated earlier, semantic heterogeneity with domain rules hidden in program code was one of the main motivations for the DB-Prism project. Thus, as postulated in [JJQV99], a conceptual perspective documenting the meaning of data and their relationships has been a critical success factor of the system. However, the mapping of this conceptual approach to the logical level – i.e. the incremental creation of basic GDH data from various sources and the restructuring of GDH data into numerous cube formats and report structures – has been equally important. Given the size and complexity of the system, it is not surprising that physical-level optimisation is also very important. For example, major surmounted challenges at the physical level include the scheduling of join operations in data integration for GDH production and the decision when to pre-materialize cubes and when to do dynamic aggregation.

These aspects cannot be considered independently, but are closely interlinked. The precise documentation and, where possible, automation of these tasks is the goal of the metadata management facilities in DB-Prism; analogous to GDH/CBM, they comprise the meta database MetaDB and an associated workbench called CDR.

5.1 MetaDB

The DB-Prism meta database brings together knowledge about a wide range of heterogeneous application and system aspects. MetaDB describe data from the viewpoints of their origin, structure, meaning, use and relationships; it is the knowledge base for the whole system and for the interaction of its components. As an active knowledge networking system, MetaDB knows, steers and drives the processes within the system as a whole not only at the conceptual but also at the logical and physical level. This way, transparency, integrity, quality, standardisation, consistency and flexibility are ensured. Figure 2, following the steps of the DB-Prism operational processes from source import via GDH and OLAP cube creation up to drill-through back to the operational data, gives an indication of the richness of knowledge covered by the MetaDB.


Knowledge about operating data
Knowledge about transformation rules
Knowledge about file and space management
Import
Transformation
Storage
Knowledge about import interfaces
Knowledge about organization and product structures
Knowledge about operating control
Knowledge about definition specifications
Knowledge about delivery specifications
Knowledge about aggregation of dimensions
Transformation
Export
Matplan
Knowledge about aggregation rules
Knowledge about export rules
Knowledge about organization units
Knowledge about report positions
Knowledge about products and dimensions
Matplan
Drill Through
Transparency, Integrity Quality, Standardization, Consistency And Flexibility Are Ensured
Knowledge about report timetables
Knowledge about storage administration

(a) GDH loading/refreshment subprocess

(b) multidimensional aggregation subprocess

(c) multidimensional analysis subprocess with drill-through

Fig. 2: Knowledge networking in the DB-Prism process

5.2 CDR Tool (Controlling Dimensions & Reports)

CDR is the central user tool for Controllers and system people to manage the MetaDB. Dimensions, products, hierarchies, delivery interfaces and report positions can be defined and adjusted. Processing is carried out online on the operational MetaDB. After an integrated program driven integration and extensive quality checks, changes can be released for real-time effectiveness. Additional important CDR features include:

- time & status concept for defining the scheduling and monitoring policies for operational data flows such as shown in figure 2.
- online status control from data import through GDH creation to data export
- metadata export functionality to automatically or interactively evolve the structure of external systems such as Matplan/b2brain.

6 Conclusions

The DB-Prism system in the form described here has been in operation for about a year. It has had a significant impact on the organisation by achieving the quality goals of integrity, coherence across controlling pathways and reporting systems, currency and availability of data, evolvability and efficient handling of massive amounts of heterogeneous data both at the source and client side.

Primary success factors include the highly efficient and flexible facilities for multidimensional databases provided by Matplan/b2brain, and a metadata model and environment that supports domain-specific quality aspects from the conceptual perspective of bank accounting knowledge in addition to the more traditional logical data transformations and physical data transport and manipulation facilities. Thus, DB-Prism provides an interesting example of a quality-oriented, concept-centred data warehouse architecture as researched in the European DWQ project [JJQV99]. Ongoing work concerns the extension to further parts of the organisation as well as broadened accessibility on the client side.

Acknowledgments. Thanks are due to the people from GMI in Aachen and Frankfurt who contributed significantly to the success of the system described here.

References

[GMI00] GMI mbH. Matplan and b2brain product description (in German). http://www.gmi-mbh.de/nav/produkte.html, June 2000.

[JJQV99] Jarke, M., Jeusfeld, M., Quix, C., Vassiliadis, P. Architecture and quality in data warehouses: an extended repository approach. Special Issue Advanced Information Systems Engineering (Pernici/ Thanos, eds.), *Information Systems 24*(3):229-253, 1999.

[JLVV99] M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis: *Fundamentals of Data Warehouses*. Springer-Verlag 1999.

[WSF95] Wang, R.Y., Storey, V., Firth, C.P. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Eng. 7*(4):623-640, 1995.