MASTER IN COMPUTER SCIENCE

Application of Data mart to the field of aviation

Tholomé, Tholomé

Award date:
2005

Awarding institution:
University of Namur

Facultés Universitaires Notre-Dame de la Paix, Namur
Institut d'Informatique.
Année académique 2004-2005

Application of Data mart to the field of aviation

Muriel Tholomé

Mémoire présenté en vue de l'obtention du grade de Licencié en Informatique.

Abstract

This paper presents a Data Mart project to enable the EUROCONTROL Performance Review Unit to analyse the performance of the air traffic management system in Europe.

It gives an overview of the Data Warehouse and Data Mart concept.

Data Warehouse gathers information from different data sources and puts them together into a unified repository for users to access through different sets of tools.

Data Warehouses are often accompanied by Data Marts. Data Marts are smaller data warehouses that are targeted to the specific needs of a department.

Keywords - *Data Warehouse, Data mart*

Acknowledgments

I would like to thank my tutors in the Institut d'Informatique, Facultés Universitaires Notre-Dame de la Paix, Namur.

My family deserve particular thanks for their unfailing support, understanding and good humour.

I would like also to thank my friends, particularly Laure Anne Holemans, for their support.

Last, but certainly not least, I would like to thank my colleagues, especially Stephan Cloquette for working with me on this data mart project, as well as Catherine Hennessy for checking my English.

Table of Contents

ABBREVIATIONS 4

1 INTRODUCTION 5

SECTION 1: LITERATURE REVIEW ON DATAWAREHOUSE AND DATA MART 6

2 EMERGENCE OF DATA WAREHOUSING 7

2.1 HISTORY 7
2.2 MAIN EVOLUTIONARY FACTORS 8

3 DATAMART VS. DATA WAREHOUSE 9

4 OPERATIONAL SYSTEM VS. ANALYTICAL SYSTEM 10

5 DATA WAREHOUSE - MAIN CHARACTERISTICS 12

5.1 SUBJECT-ORIENTED 12
5.2 INTEGRATED 12
5.3 TIME-VARIANT 13
5.4 NON-VOLATILE 13
5.5 MANAGEMENT NEEDS 14

6 DATA STRUCTURE 15

6.1 GRANULARITY OF DATA 15
6.2 LEVEL OF DETAIL 16
6.3 META DATA 16

7 STRUCTURE OF THE DATAWAREHOUSE SYSTEM 17

7.1 ACQUISITION OF DATA 17
7.2 STORAGE OF DATA 19
7.3 ACCESS TO DATA 20

8 DATA WAREHOUSE ARCHITECTURE 22

8.1 ENTERPRISE DATA WAREHOUSE 22
8.2 STAND-ALONE DATA WAREHOUSE 23
8.3 INDEPENDENT DATA MART 23
8.4 INTERDEPENDENT DATA MARTS 24
8.5 INTERDEPENDENT DATA MARTS WITHOUT PHYSICAL DATA WAREHOUSE 25

9 IMPLEMENTATION METHODOLOGIES 25

9.1 TOP-DOWN ARCHITECTURE 26
9.2 BOTTOM-UP ARCHITECTURE 26
9.3 HYBRID ARCHITECTURE 27
9.4 COMPARISON OF THE DIFFERENT ARCHITECTURES 28

10 DATA MODELLING 29

11 DIMENSIONAL MODELLING 30

11.1 DIMENSIONAL MODELLING TERMINOLOGY 30
11.2 STAR SCHEMA/SNOWFLAKE SCHEMA 30
11.3 ADDITIVITY OF FACTS 33
11.4 SURROGATE KEY 33
11.5 SLOWLY CHANGING DIMENSION 33

12 OLAP TOOL 35

12.1 INTRODUCTION 35
12.2 CHARACTERISTICS 35
12.3 DIMENSIONAL DATA STORAGE 38
12.4 OLAP TOOL VENDORS 38

SECTION 2: PRU DATA MART PROJECT 40

13 INTRODUCTION 40

14 PROJECT PLANNING 40

14.1 PROJECT DEFINITION 40
14.2 PROJECT SCOPE 44
14.3 PROJECT PLANNING AND MANAGEMENT 45

15 TERMINOLOGY 46

16 BUSINESS REQUIREMENT DEFINITION 49

17 DIMENSIONAL MODELLING 51

17.1 FLIGHT DATA MART 52
17.2 AIRSPACE DATA MART 56
17.3 REGULATION DATA MART 61

18 PHYSICAL DESIGN 64

18.1 SURROGATE KEYS 64
18.2 MATERIALISED VIEW 64

19 DATA STAGING DESIGN & DEVELOPMENT 64

20 END USER APPLICATION 67

20.1 SELECTION OF THE TOOL 67
20.2 BO TOOL 68
20.3 USER INTERFACE DEVELOPMENT 72

21 DEPLOYMENT 72

22 MAINTENANCE AND GROWTH 73

23 CO-ORDINATION WITH OTHER UNITS 73

24 CONCLUSION 75

ANNEX I - GLOSSARY 76

ANNEX II - BIBLIOGRAPHY 79

Table of Figures

Figure 1: Basic data warehouse system architecture 6
Figure 2: Naturally evolving architecture [INMON,1998] 7
Figure 3: Comparison of data warehouse and data mart 9
Figure 4: Operational system vs. data warehouse 11
Figure 5: Subject-oriented 12
Figure 6: data integration 13
Figure 7: Time variance 13
Figure 8: Non volatility 14
Figure 9: trade-off between detailed analysis and volume of data 15
Figure 10: Structure of data 16
Figure 11: Components of a data warehouse system 17
Figure 12: Meta data 20
Figure 13: Enterprise data warehouse 22
Figure 14: Stand-alone data warehouse 23
Figure 15: Independent data marts 23
Figure 16: Interdependent data marts 24
Figure 17: Interdependent data marts without physical data warehouse 25
Figure 18: Top-down approach 26
Figure 19: Bottom-up approach 26
Figure 20: hybrid approach 27
Figure 21: Comparison of the different architectures 28
Figure 22: Star schema 31
Figure 23: Dimension de-normalisation 31
Figure 24: Snowflake schema 32
Figure 25: Fact constellations 32
Figure 26: Multi-dimensional data cube 36
Figure 27: OLAP operations 37
Figure 28: ROLAP and MOLAP 38
Figure 29: OLAP market [PENDSE,2005] 39
Figure 30: The business dimensional lifecycle diagram 40
Figure 31: CFMU data sources 43
Figure 32: Situation up to November 2002 43
Figure 33: Future situation 44
Figure 34: ATFM delay 46
Figure 35: Hierarchies of entities 47
Figure 36: Relation between traffic and delay 48
Figure 37: Measures of interest 50
Figure 38: Data marts matrix 52
Figure 39: entities in Flight data mart 52
Figure 40: dimension model of flight data mart 53
Figure 41: Date dimension 54
Figure 42: airport dimension 55
Figure 43: aircraft operator dimension 55
Figure 44: aircraft type dimension 56
Figure 45: Entities in Airspace data mart 57
Figure 46: Dimension model of the airspace data mart 58
Figure 47: Entities in Regulation data mart 61
Figure 48: Dimensional model of regulation data mart 62
Figure 49: loading of traffic data per airspace 66
Figure 50: Business Objects tools 68
Figure 51: BO Designer tool 69
Figure 52: BO query interface of the end user 70
Figure 53: BO report 70
Figure 54: Example of BO report with drill down 71
Figure 55: Example of BO report exported for the web 71
Figure 56: EATMP SAMAD Data Warehouse 74

ABBREVIATIONS

AIRAC	Aeronautical Information Regulation and Control
ATM	Air Traffic Management
ANSP	Air Navigation Services Provider
AO	Aircraft Operator
ATC	Air Traffic Control
ATFM	Air Traffic Flow Management
AUA	ATC Unit Airspace
BI	Business Intelligence
BO	Business Object
CTOT	Computed Take Off Time
CFMU	Central Flow Management Unit
CODA	Central Office for Delay Analysis
DB	Data Base
DBMS	Data Base Management System
DM	Data Mart
DSS	Decision Support System
DW	Data Warehouse
EATMP	European Air Traffic Management Program
ECAC	European Civil Aviation Conference
ETL	Extraction, Transformation and Loading
ETOT	Estimated Take Off Time
FIR	Flight Information Region
HTML	Hyper Text Mark-up Language
KPA	Key Performance Area
KPI	Key Performance Indicator
MOLAP	Multi dimensional On Line Analytical Processing
MTOW	Maximum Take Off Weight
NAS	National AirSpace
OLAP	On Line Analytical Processing
OLTP	On Line Transactional Processing
PRC	Performance Review Commission
PRU	Performance Review Unit
RDBMS	Relational Database Management System
ROLAP	Relational On Line Analytical Processing
SAMAD	System for the Acquisition and Management of Airspace Data
STAT-AUA	Statistical ATC Unit Airspace
STAT-TVS	Statistical Traffic Volume Set
TV	Traffic Volume
TVS	Traffic Volume Set

1 INTRODUCTION

This document contains two sections. The first section presents the "state of the art" on the data warehouse and data mart. The second section contains a case study on the realisation of data marts for the Performance Review Unit (PRU) of EUROCONTROL.

In Section 1, the background to data warehousing is described, then the differences between operational and analytical systems as well as the main characteristics of a data warehouse are described.

Next, the data structures and the system structure of a data warehouse are explained. A brief overview of different data warehouse system architectures is then given.

The concept of dimensional modelling, as well as associated On-Line Analytical Processing (OLAP) tools, has been specifically included as they will be used in Section 2.

In Section 2, the background of the data mart project and user requirements are presented. The dimensional models of the data marts, data staging and end user OLAP tool are then presented.

SECTION 1: LITERATURE REVIEW ON DATAWAREHOUSE AND DATA MART

Historically, information systems were geared to support day-to-day business. In recent years, in the context of an increasingly competitive economy, the need to provide decision-makers, analysts and key users with more and better information has become essential.

However, notwithstanding the existence of powerful computers and communication networks, it was difficult for key people to access the information they needed to do their work, even though the information existed in their enterprises.

Thus, in the 1990s, an approach called data warehousing was developed. Its aim was to provide to key users a quick and easy access to information.

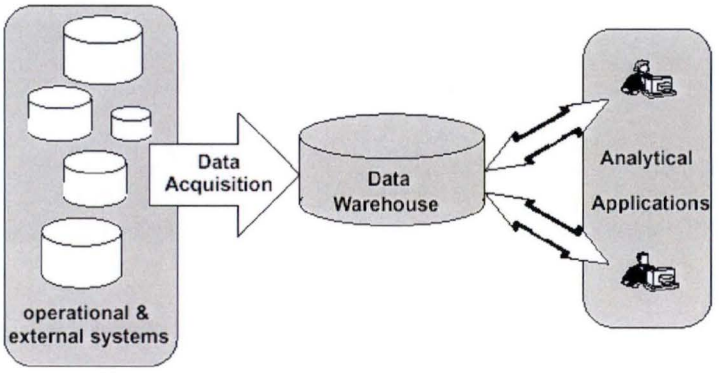


Figure 1: Basic data warehouse system architecture

Mr. Inmon, who is regarded as the one of the "father" of data warehousing, describes data warehousing as "a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, and analyst) to make better and faster decisions". [INMON,1998].

In other words, data Warehousing is the process of extracting, transforming, integrating data from multiple, heterogeneous source systems, and subsequently presenting data as information to business users enabled with analysis and reporting facilities. Data warehousing is supported by an integrated information technology infrastructure comprised of hardware, database management systems, and other business analysis tools.
The central store where the data are stored is called a Data warehouse (DW).

The data warehousing concept is often related to decision support systems, as the data warehouse is mainly utilised as the data repository for these systems.[1]

[1] There are many different terms in today's decision-support systems. Some have been in existence for a long time, e.g. MIS, EIS and DSS. Others are more recent, e.g. Business Intelligence (BI).

2 EMERGENCE OF DATA WAREHOUSING

The following paragraphs give a brief history of the analytical systems and the factors which have favoured the emergence of data warehousing. [Based on GUPTA, BERKELEY, INMON]

2.1 History

The 1960s saw the advent of the direct data storage device and the database management system (DBMS).

Data originating from mainframe computers were extracted in order to generate reports. These reports were generated in batches, usually during off-peak time. A major drawback, however, was that the time needed by IT teams to develop or modify reports was considerable, and thus was not acceptable to end-users.

"In the late 1970s, it became apparent that mainframe-based production systems could not support enterprise-wide decision support. These systems fragmented fundamental business "objects," such as customers and markets, into transaction-level detail data spread across many production databases, and they could not sustain the performance levels required by mission-critical applications while simultaneously servicing knowledge workers' complex queries". [DESMARET,2001]

In the 1980s, with the appearance of personal computers (PC), users were no longer constrained by "dumb" terminals. They were able to create their own applications using spreadsheets and small PC databases. The proliferation of multiple data extracts led to a situation where control and integrity of the data got out-of-control, not to mention the loss of productivity.

"The pattern of this out-of-control extract processing became so commonplace across organisations that it was given the name of "naturally evolving architecture". [INMON,1998]

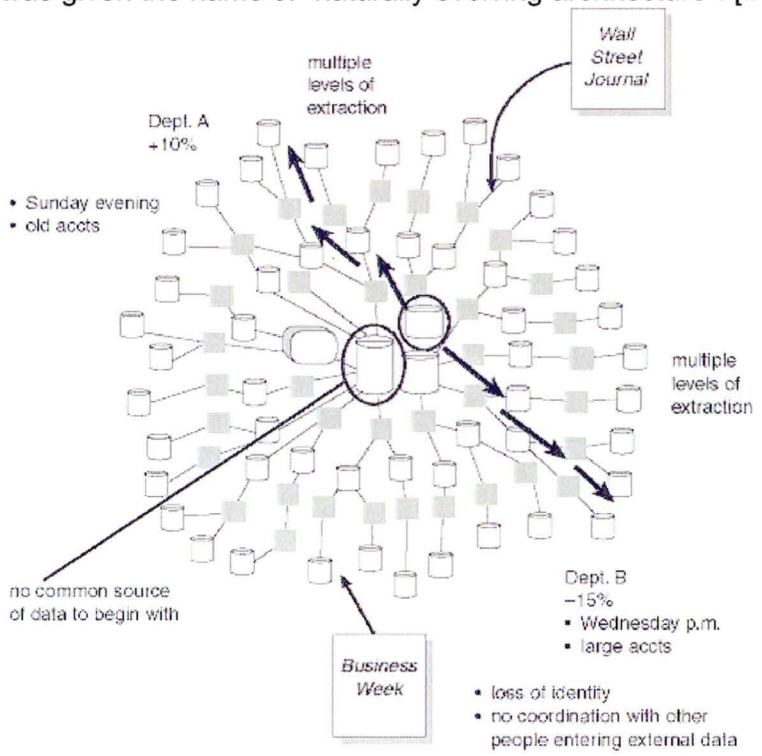


Figure 2: Naturally evolving architecture [INMON,1998]

An example of this "out of control" extract processing can be seen in Figure 2. Department A delivers a report to management claiming activity is 10% up, while Department B says activity is 15% less. The difference can be due to different date extraction (Sunday evening for A and Wednesday p.m. for B) and to algorithm difference, one choosing only large accounts and other one old accounts. Multiple extractions are also a source of discrepancies because of different timings or algorithm differences.

"Two other analysis systems - Decision Support Systems (DSS) in the 1970s and Executive Information Systems (EIS) in the 1980s - may be viewed as the closest precursors to data warehousing systems. Both have data in descriptive standard business terms and generally pre-processed with the application of standard business rules, and both provide consolidated views of the data. However, their designs are also derived from specific requirements, rather than the overall business structure, and the cost and coordination required for their development adversely affected their popularity". [BERKELEY,1997].

In the 1990s, the data warehousing concept with one overall business structure emerged.

2.2 Main evolutionary factors

"Many factors have influenced the rapid evolution of data warehousing. The most significant factors have been the enormous advance in hardware and software technologies. Sharply decreasing prices and the increasing power of computer hardware, coupled with ease of use of today's software, have made possible quick analysis of hundreds of gigabytes of information and business knowledge." [GUPTA,2000]

"What is more, the explosion of intranet and Web-based applications with the open Internet standards has greatly impacted data warehousing as well" [BERKELEY,1997]

"Another very significant influence on the evolution of data warehousing science is the fundamental changes in the business organization and structure during the late 1980s and the early 1990s. The emergence of a vibrant global economy has profoundly changed the information demands made by corporations. The use of technology by mid and upper level managers has increased significantly. This hands-on use of information and technology by upper management has facilitated the sponsorship of larger projects such as data warehousing. "[GUPTA,2000]

3 DATAMART VS. DATA WAREHOUSE

Although no strict definitions of data warehouse and data mart appear to exist, a number of informal definitions can be found in the literature. Some examples are given below:

- Bill Inmon defines data warehouse as: "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process". [INMON,1998]
- Stanford University defines data warehouse as "a repository of integrated information, available for queries and analysis. data and information are extracted from heterogeneous sources.... This makes it much easier and more efficient to run queries over data that originally came from different sources" [STANFORD]
- Webopedia states: "a data mart is a database, or collection of databases, designed to help managers make strategic decisions about their business. Whereas a data warehouse combines databases across an entire enterprise, data marts are usually smaller and focus on a particular subject or department. Some data marts, called dependent data marts, are subsets of larger data warehouses".
- Inmon, who is regarded as the "father" of the data warehouse, states that "A data mart is a subset of a data warehouse that has been customized to fit the needs of a department". [Inmon]
- Ralph Kimball, who is another guru of data warehousing, states that *"...the data warehouse is nothing more than the union of all the data marts."* [KIMBALL,1998].

From these definitions and the literature, it can be seen that there are differences in how the concepts are understood. However, it can also be seen that there is agreement on many common features for data mart. These can be summarised as follows:

- A data mart is a database that has the same characteristics as a data warehouse but is usually smaller and is focused on the data needs of a department.
- Since it is smaller and does not cover all the enterprise subject areas, it is easier and less expensive to build.

It can be difficult to determine the exact boundary between data mart and data warehouse.

In chapter 7, it will be seen that data warehouse and data mart can exist together or independently of each other.

The table below summarises the differences between a data warehouse and a data mart.

Attributes	*Data Warehouse*	*Data Mart*
Scope	Enterprise	Department
Data sources used	Many	Few
Subject areas covered	Many	One or few
Size	GB-TB	MB-GB
Initial effort	More	Less
Initial cost	Higher	Lower
Users	Many	Few

Figure 3: Comparison of data warehouse and data mart

4 OPERATIONAL SYSTEM VS. ANALYTICAL SYSTEM

It is useful to make a comparison between operational systems and analytical systems (or informational system) in order to better understand the data warehouse concept.

"Perhaps the most important concept that has emerged from the data warehouse movement is the recognition that there are two fundamentally different types of information systems in all organizations: operational systems and information systems". [ORR,2000]. From this, the idea has emerged to separate the operational data from the informational data.

The reasons to separate the data are explained below

- The impact on the operational systems is minimal.
- The data warehouse is accessible even if the source of data is not accessible.
- Data can be integrated from different systems to create new, subject-oriented data.
- Analysis based on histories of operational data is possible, independently of whether operational systems provide support for history management or not.

Operational systems, also referred as On-line Transaction Processing (OLTP), handle the day-to-day business of a company.

"Information systems have to do with analysing data and making decisions, often major decisions, about how the enterprise will operate, now and in the future. And not only do information systems have a different focus from operational ones, they often have a different scope. Whereas operational data requirements are normally focused upon a single area, informational data often span a number of different areas and need large amounts of related operational data". [ORR,2000]. Informational systems are also referred to in the literature as On-Line Analytical Processing (OLAP).

The data warehouse is the heart of the information system as it is the data repository that collects, organises, and makes analytical data available.

OLTP applications access detailed, current data, and typically read or update individual data records. Consistency and recoverability of the operational database are critical issues as well as fast response time.

In contrast to operational systems which store detailed and current data, data warehouse aims at providing integrated, consolidated historical data. The workloads are query-intensive with mostly ad hoc, complex queries that can access millions of records.

Data warehouse stores historic information covering many years. This implies considerable capacity requirements for data processing and storage.

Figure 4[2] summarises the differences between operational and analytical systems.

Criterion	Operational systems	Analytical systems
Main aim/focus	-daily business - transaction processing	- decision support - analytical processing
Typical user	Clerk staff	Analysts, decision makers, executive staff, controllers
Data content	-detail -current -isolated	-detail & aggregated - historical -integrated
Queries	-simple -repetitive -read/write	-complex -ad-hoc read mostly
System requirements	-transaction throughput -consistency	-query throughput -quality
Database design	-transaction oriented -normalised	-subject oriented, multidimensional - partly normalized
Data volume	MB-GB	GB-TB

Figure 4: Operational system vs. data warehouse

[2] Source: [VAVOURAS,2002].

5 DATA WAREHOUSE - MAIN CHARACTERISTICS

In the literature, the main characteristics are usually based on Inmon's definition of data warehouse, which has already been given. It is reproduced here again for ease of reference: *"a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process"*. [INMON,1998]

5.1 Subject-oriented

Data are gathered by subject instead of application. In contrast to operational systems which are application-oriented, the warehouse is oriented to the major subject areas of a company.

The differences between process/function application orientation and subject orientation show up as a difference in the content of data at the detailed level as well.

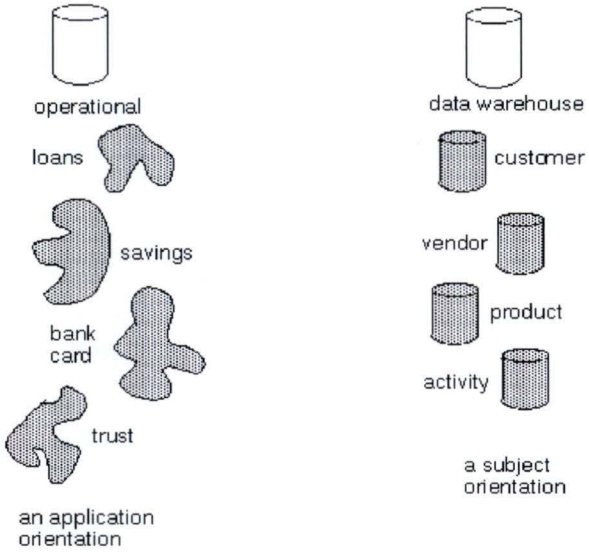


Figure 5: Subject-oriented [3]

5.2 Integrated

Data are integrated from various heterogeneous operational systems and external data sources. All data in the warehouse must be compatible with each other irrespective of whether the underlying source data are stored differently. This includes consistent naming conventions, measurement of variables, encoding structure, physical attributes, etc. Thanks to the integration of data, it is much easier for the user to get the right information. Furthermore these data are clearly defined and quality controlled.

[3] The source of the figures presented in this chapter is [INMON,1995]

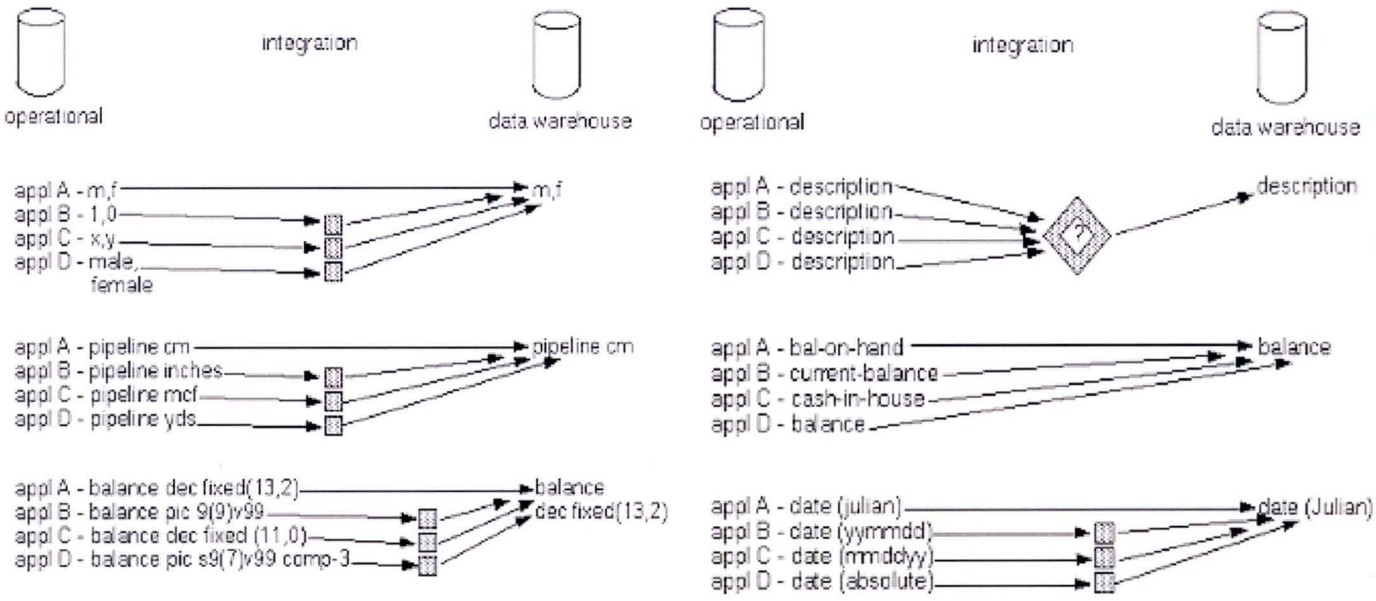


Figure 6: data integration

5.3 Time-variant

A data warehouse maintains historical data (it includes time as a variant). Unlike operational databases where usually only recent data are maintained, data in the data warehouse cover long periods, which enable trend analysis to be performed.

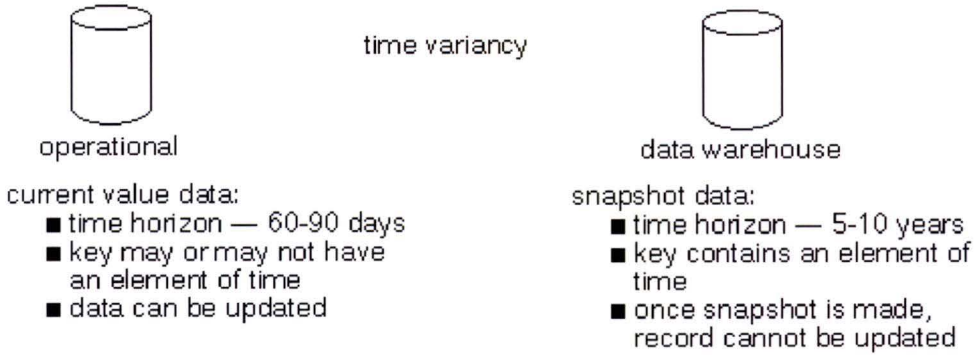
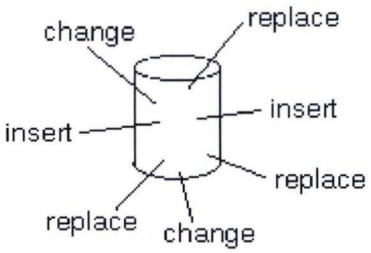


Figure 7: Time variance

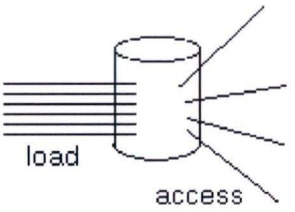
5.4 Non-volatile

The data do not change once they have been collected. Access is typically read-only. Modifications of the warehouse data take place only when modifications of the source data are propagated into the warehouse.



operational

data is updated on a record-by-record basis regularly



data warehouse

data is loaded into the warehouse and is accessed there, but once the snapshot of data is made, the data in the warehouse does not change

Figure 8: Non volatility

5.5 Management needs

The data warehouse is intended for decision-makers, people who need to analyse the data and/or make business decisions. This assumption is somewhat restrictive as data warehouse is not only used in the decision support systems.

6 DATA STRUCTURE

6.1 Granularity of data

Granularity refers to the data levels of detail or summarisation. The more detail there is in a data warehouse, the lower level of granularity. The less detail there is, the higher the level of granularity.

As mentioned by Bill Inmon, "The single most important aspect of design of a data warehouse is the issue of granularity" [INMON,1998].

When choosing the right granularity, not only does existing business needs have to be taken into account but also future business needs.

"The granularity level is significant from a business, technical and project perspective.

From a business perspective, it dictates the potential capability and flexibility of the data warehouse. Without a subsequent change to the granularity level, the warehouse will never be able to answer questions that require details below the adopted level.

From a technical perspective, it is one of the major determinants of the data warehouse size and hence has a significant impact on its operating cost and performance.

From a project data mart, the granularity level affects the amount of work that the project team will need to perform to create the data warehouse since as the granularity level gets into more detail, the project team needs to deal with more data attributes and their relationship". [IMHOFF,2003].

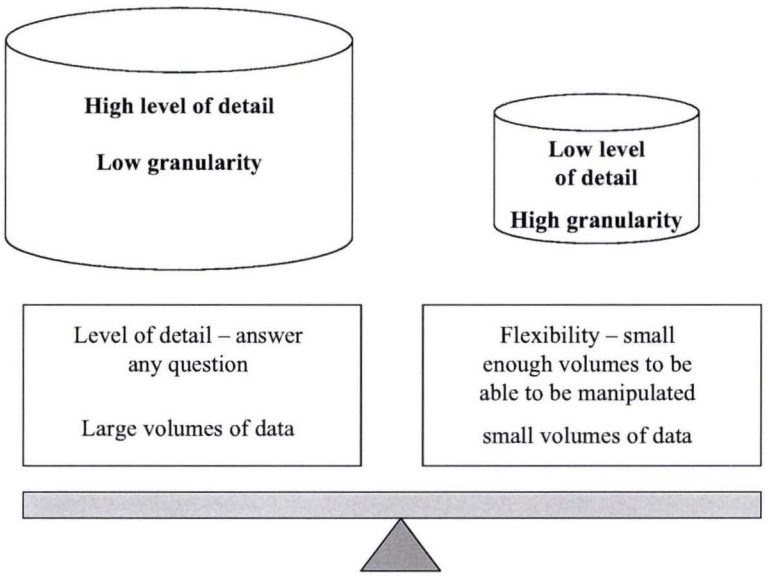


Figure 9: trade-off between detailed analysis and volume of data[4]

[4] The source of the figures presented in this chapter is [INMON,1998]

6.2 Level of detail

According to Inmon, there are different levels of detail in the data warehouse:

- Older detail
- Current detail
- Lightly summarized data
- Highly summarized data

A data warehouse is organised within 2 dimensions, a dimension of time and a granularity dimension.

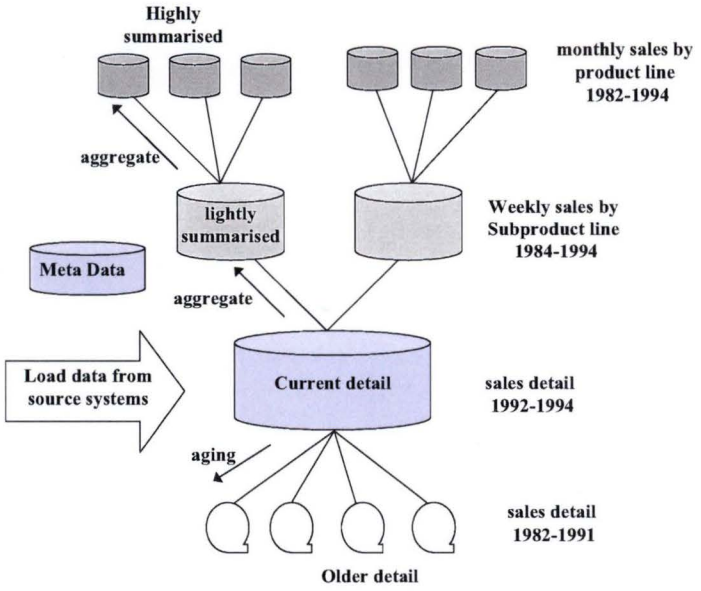


Figure 10: Structure of data

Data loaded from an operational system enters as up-to-date, detailed data. Usually significant transformation occurs at the passage from operational level to data warehouse level. These detailed data are not always identical to the data in the operational system. They can be an aggregation or a simplification of the operational data. Data which have been simplified, summarised or calculated from operational data are called derived data.

All detailed data can be aggregated to yield lightly-summarised data. These summaries can further be aggregated to yield highly-summarised data (see Figure 10).

Thus several levels of granularity are stored in a data warehouse, although this produces some redundancy. Because of the enormous amounts of data stored in a data warehouse, some analytical tasks only are computable within an acceptable time, if some required data is pre-aggregated.

The data ages with time and simultaneously its importance and the chances of accessing it decrease. The older data stays in the data warehouse but moves to external (slower but cheaper) storage media. data stored in these external media are considered part of the DW, because these data can be accessed for analyses, if needed.

6.3 Meta data

Meta data are often defined as the "the data about the data". Meta data are used for building, maintaining, managing and using the data warehouse/mart.

Meta data are a kind of index to the contents of the data warehouse/mart. It keeps track of what is where in the data warehouse. Meta data are dealt with in more detail in Chapter 7 (Section 7.2.3).

7 STRUCTURE OF THE DATAWAREHOUSE SYSTEM

There are three key functions: Acquisition of data, storage of data and access to data, see Figure 11.

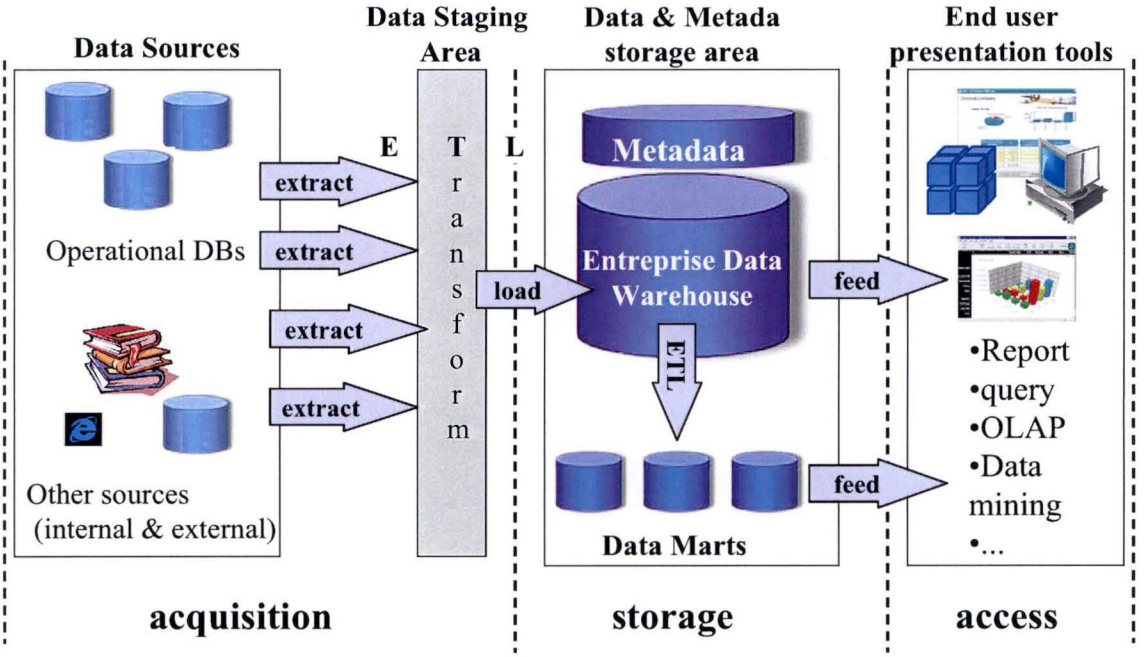


Figure 11: Components of a data warehouse system[5]

7.1 Acquisition of data

"Typically, the source data for the warehouse comes from operational applications. As the data enters the warehouse, it is cleaned up and transformed into an integrated structure and format. The transformation process may involve conversion, summarisation, filtering and condensation of data". [BERSON & SMITH,1997]

7.1.1 Source Systems

Data sources are mainly operational systems whose function is to capture the transactions of the business concerned. data sources can also be external information sources.

The data in these systems can be in many formats from flat files to hierarchical and relational RDBMS. Other sources of data may already be cleansed and integrated and available from operational data stores.

7.1.2 Operational data Store[6]

In some cases, an intermediate layer, called an operational data store (ODS), is introduced between the operational systems and the data warehouse. Its purpose is to address the need of users for an integrated view of current operational data. ODS data is *"subject-oriented, integrated, volatile, and current"* [Inmon]. An ODS is subject to change much more frequently than a DWH, and stores, in contrast to a DWH, no histories over operational data. Thus, an ODS provides support for activities such as collective operational decisions based on current company-wide information. [VAVOURAS,2002].

5 The architecture presented is the Enterprise Data Warehouse (see chapter 8).
6 The operational Data Store is not shown in Figure 11

7.1.3 Data Staging Area and Extraction, Transformation and loading

The data staging area is the portion of the data warehouse restricted to extracting, cleaning, matching and loading data from multiple-source systems. The data staging area is the back room and is explicitly off-limits to end users. The data staging area does not support query or presentation services. A data-cleansing tool may be used to process data in the staging area to resolve name and address misspellings and the like.

A significant portion of the effort is spent extracting data from operational systems and putting it in a format suitable for analytical applications that run off the data warehouse.

Different applications developed at different times for different operational purposes often contain data that are inconsistent or redundant with data in other applications. data elements with the same name may be defined differently. The same elements in two different systems may be stored under a different name.

Data Extraction-Transformation-Load (ETL) tools are used to extract data from data sources, cleanse the data, perform data transformations, and load the target data warehouse and then again to load the data marts.

"The ETL tools produce the programs and control statements needed to move data into the data warehouse for multiple operational systems. These tools also maintain the Meta data". [BERSON & SMITH,1997]

Extraction

Data extraction acquires data from the operational systems and stores data in a temporary processing area.

Transformation

Data are transformed from transaction-level data into information through several techniques: filtering, summarising, merging, transposing, converting and deriving new values through mathematical and logical formulae. Data transformation integrates data into standard formats and applies business rules that map data to the warehouse schema. Issues of data standards, domains and business terms arise when integrating across operational databases

Data Cleansing

Data cleansing is based on the principle of populating the data warehouse with quality data - that is, data that are consistent, are of a known, recognized value and conform to the business definition as expressed by the user.

The cleansing operation is focused on determining those values which violate these rules and, either reject or, through a transformation process, bring the data into conformance.

Data cleansing standardises data according to specifically defined rules, eliminates redundancy to increase data-query accuracy, reduces the cost associated with inaccurate, incomplete and redundant data, and reduces the risk of invalid decisions made against incorrect data.

Loading

Data loading loads the cleaned data into the data warehouse. DW are often taken offline for the loading process, therefore the design of the loading element should focus on efficiency and performance to minimize the data warehouse offline time. This becomes increasingly important for daily updated, where updates are run overnight and little remains for eventual reruns in case of problems.

7.2 Storage of data

The storage component may consist of one or several distinct databases acting as an enterprise data warehouse and of multiple data marts.

7.2.1 Data Warehouse

The data warehouse database is the heart of the data warehousing environment. It collects and stores integrated sets of historical, non-volatile data from multiple operational or external systems and feeds them to one or more data marts. It becomes the one source of the truth for all shared data.

"Because the data contains a historical component, the warehouse must be capable of holding and managing large volumes of data as well as different data structures for the same database over time." [BERSON & SMITH,1997]

7.2.2 Data Marts

A data mart typically contains a subset of the data from a data warehouse. Data are often at a more summarised level than the data warehouse and have a different database design customized to meet the functional and access needs of a specific business area. Very often, an organization will have multiple data marts, each supporting a particular set of user needs.

The data may be customised to the special needs of the target users, and may include, or be combined with data that are not stored centrally in the data warehouse and are relevant only to them. The user community of a data mart is generally much smaller than that of the total data warehouse.

7.2.3 Meta data and Meta data repository

As already stated in Chapter 4, Meta data are "the data about the data". They are critical to the data warehouse's success.

The Meta data repository contains a complete glossary for all components, databases, fields, objects, owners, access, platforms and users of the data warehouse system. The repository offers a way to understand what information is available, where it comes from, where it is stored, the transformation performed on the data, its currency and other important facts about the data.

"Often, a Meta data repository is used to manage all Meta data associated with a data warehouse system. The repository enables the sharing of Meta data among tools and processes for designing, setting up, using, operating, and administering a data warehouse system". [VAVOURAS,2002]

"Meta data are used not only by developers but also by end users. Meta data provide interactive access to users to help understand content and find data". [BERSON & SMITH,1997]

Meta data can be classified as:

- Technical Meta data, which contains information about warehouse data for use by warehouse designers and administrators when carrying out warehouse development and management tasks.
- Business Meta data, which contains information that gives users an easy-to-understand data mart of the information stored in the data warehouse.

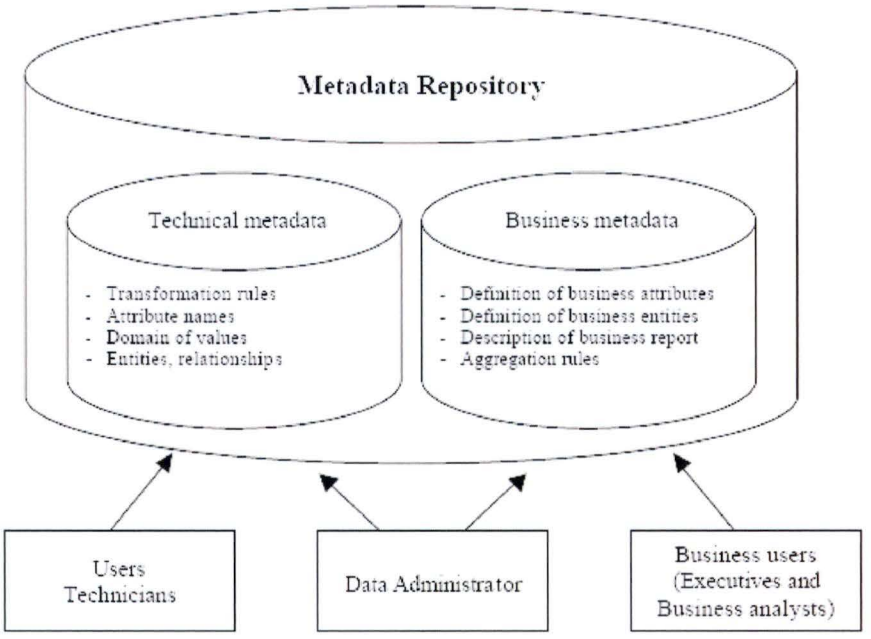


Figure 12: Meta data[7]

"Data administrators are business oriented, focusing on the meaning and use of data. Database Administrators (DBA) are technically oriented, and are concerned with the reliability, integrity and performance of database applications. A data administrator typically deals with business problems due to incorrect data values or invalid use of data due to misinterpretation. Data administration means management of the quality of corporate data." [LAMBERT,1996]

7.3 Access to data

The principal purpose of data warehousing is to provide information to users. In order to access the data, one or more front-end tools are needed, depending on the nature of the user requirements.

Classical tools include, *inter alia*, reporting tools and, in the context of data warehouse, OLAP and data mining tool.

7.3.1 Reporting tools

Static reporting is a repeatable, pre-calculated and non-interactive request for information. Where reporting of this nature is often viewed as hardcopy, it may take on newer forms as the internet and Intranet can become a vehicle for fast dissemination of information.

In interactive reporting, the result set is filtered based on user-entered parameter values. This tool uses a point-and-click display to produce reports with dynamic content.

Ad hoc query provides business analyst the ability to pose specific questions to produce a result.

7 Source: [POLENIS, 2002]

The tools falling into this category offer the ability, often through a point and click interface, to search the database and produce a result that can then be displayed, further refined and analyzed.

7.3.2 OLAP tools

OLAP tools are based on the concepts of dimensional data model and allow users to analyse the data using elaborate, multidimensional views. These tools are considered in more detail in Chapter 12.

7.3.3 Data Mining tools

"Data mining can be described as *the* process of discovering meaningful new correlations, patterns and trends by digging into large amounts of data stored in the warehouse using artificial intelligence, statistical and mathematical techniques." [BERSON & SMITH,1997]

There are two main kinds of models in data mining: *predictive* and *descriptive*. Predictive models can be used to forecast explicit values, based on patterns determined from known results. For example, from a database of customers who have already responded to a particular offer, a model can be built that predicts which prospects are likeliest to respond to the same offer. Descriptive models describe patterns in existing data, and are generally used to create meaningful subgroups such as demographic clusters.

8 DATA WAREHOUSE ARCHITECTURE

As already mentioned, the data warehouse and the data mart can exist together or independently of each other.

The data warehousing architecture is named differently depending on its design. The most common ones are presented below with their advantages and disadvantages.

8.1 Enterprise data warehouse

The data warehouse gathers all the information from the various data sources. Specialised data marts are then created with a subset of the information in the data warehouse. These data marts are easier to use because they only have the particular information that the specific user group needs.

This data mart is called dependent data mart as its data is populated from the data warehouse. It is also possible for locally generated and used information to be store in dependent data mart.

Data marts usually contain lightly and highly summarised data and detailed data are stored in the data warehouse.

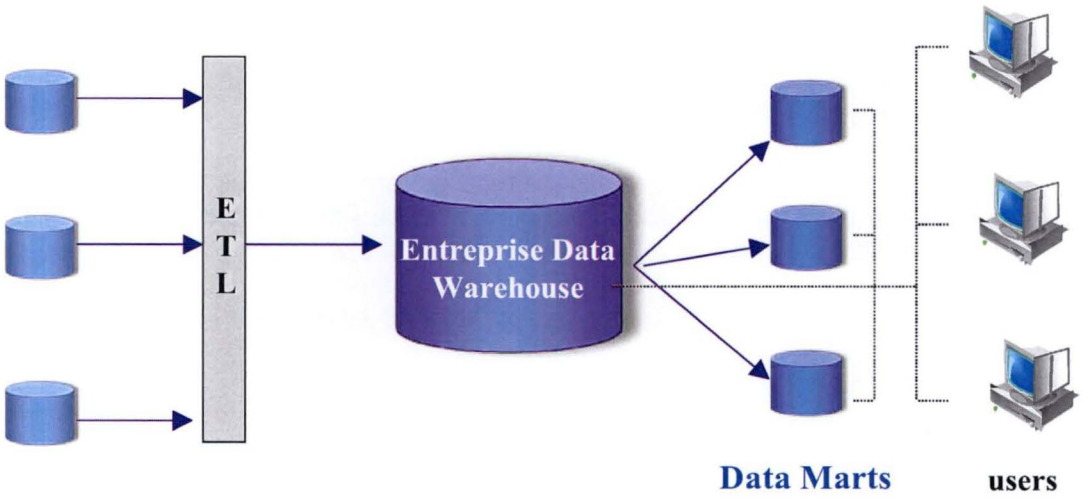


Figure 13: Enterprise data warehouse

According to [Hackney, 2000], the upsides of the classic EDW architecture are:

- Single version of the truth
- One set of extraction processes and business rules
- Common semantics
- Centralized, controlled environment
- Easily created and populated subset data marts
- Single Meta data repository.

The downsides of the classic EDW architecture are:

- Expensive to implement
- Very resource intensive
- Inherent enterprise scale requires enterprise scale systems and resources.
- High exposure risk as it is an enterprise scale

8.2 Stand-alone data warehouse

This architecture is a simplification of the enterprise data warehouse: it is formed only by a data warehouse without the data marts.

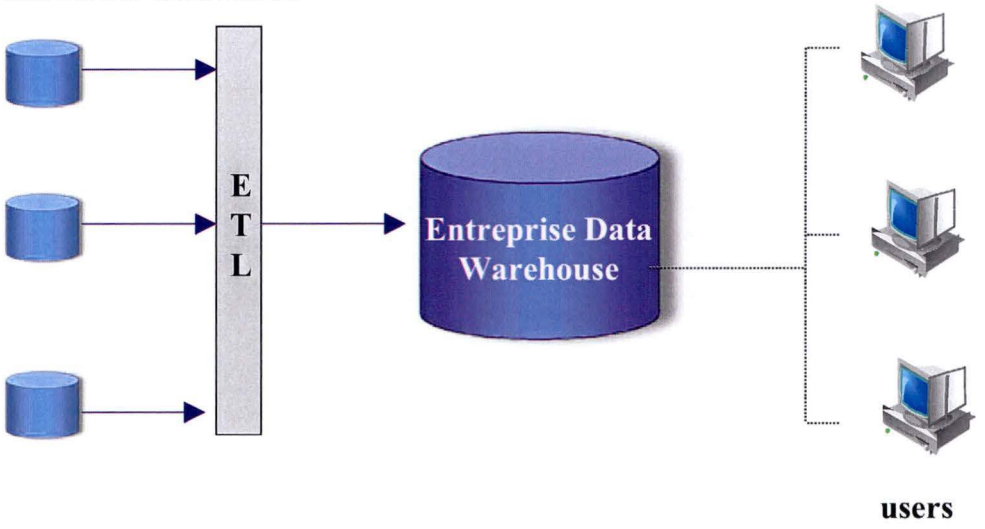


Figure 14: Stand-alone data warehouse

Especially for small data warehouses, it allows the costs of developing and maintaining data marts to be reduced. However performance, availability, and adaptability to various user group requirements could be disadvantaged by this approach.

8.3 Independent data mart

Free-standing data marts are created independently of a data warehouse. It is quicker and cheaper to build a separate data mart instead of building an enterprise-wide data warehouse with data marts derived from it. The drawback of this solution is that the company's data is not integrated. If several separate data marts are built using this strategy, they will usually contain data that are duplicated and inconsistent. These data marts are also called Non architected data marts or LegaMarts

There is generally a separate data staging area for each data mart and there are no common Meta data components across the data marts.

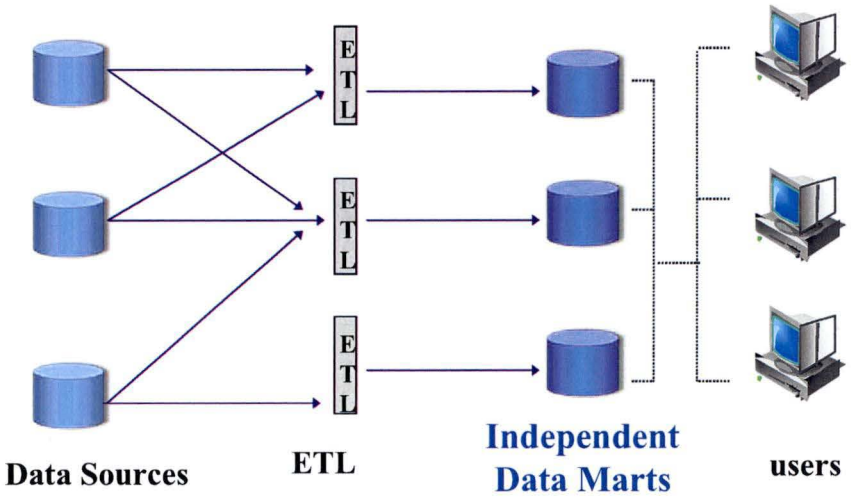


Figure 15: Independent data marts

According to [Hackney,2000], the upsides of independent data marts are: Speed and Low cost. The downsides are:

- Multiple versions of the truth
- Multiple extraction processes
- Multiple business rules
- Multiple semantics
- Extremely challenging to integrate

8.4 Interdependent data marts

This architecture supports an incremental approach to the data warehouse through data mart development by creating a common framework for development.

Individual interdependent data marts are created first and then integrated to create a data warehouse.

Since the data marts are to be the building blocks of the data warehouse, they not only contain summarised data but also detailed data that will appear in the projected data warehouse

The common framework includes enterprise subject areas, common dimensions, metrics, business rules, and data sources, all represented in a logically common Meta data repository. This ensures consistency between data marts.

This common framework is established before incremental process of data mart/ data warehouse development occurs and is updated as new data marts are built.

Central also to the architecture is a common data-staging for extraction, transformation, and loading. This facilitates integration across data marts and with the data warehouse. The unified data staging area along with the global Meta data repository and local data mart Meta data repositories all help to create and maintain semantic consistency in data. [HACKNEY,1998]

The data warehouse is created by moving the detailed data of each data mart to the common data warehouse. Rolling-up an enterprise DW is simply a matter of taking this shared data model to a higher (enterprise) level.

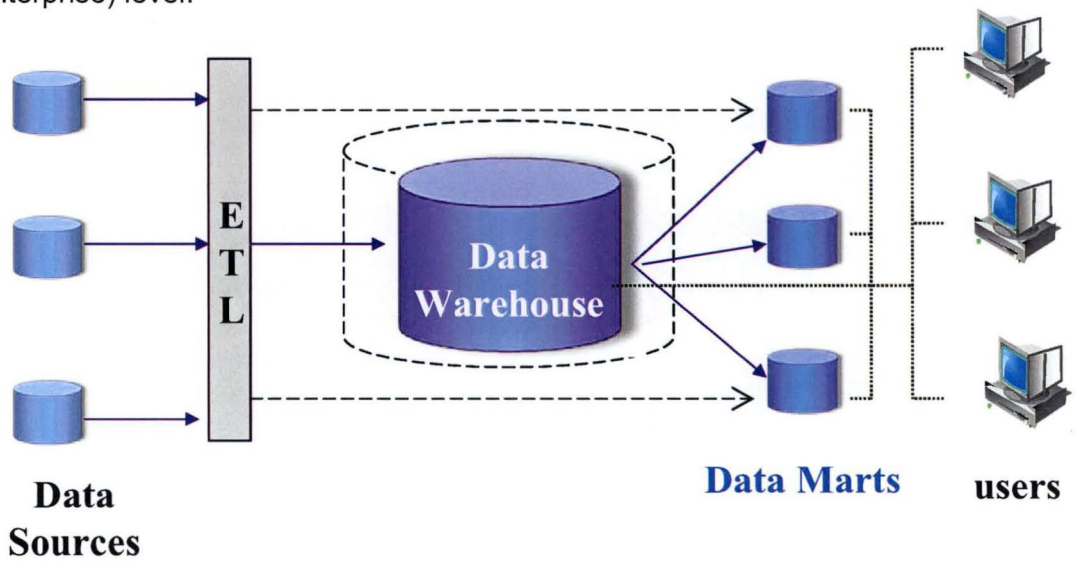


Figure 16: Interdependent data marts

According to [HACKNEY,1998] the upsides of interdependent data marts are:

- Common semantics and business rules.
- Single set of extraction processes.
- Accomplishable scope.
- Inherently incremental.

The downsides of incremental interdependent data Marts are:

- Requires common framework (model).
- Requires compliance with common framework.

8.5 Interdependent data marts without physical data warehouse

This architecture is similar to the previous architecture with the important exception that no physical enterprise-wide data warehouse is implemented. Instead, the data warehouse is viewed as the conjunction of the data marts.

The Meta data repository provides a common view of the resources across data marts. In order to answer an enterprise-wide question, it requires more work to combine the information from individual data marts than from an Enterprise data warehouse.

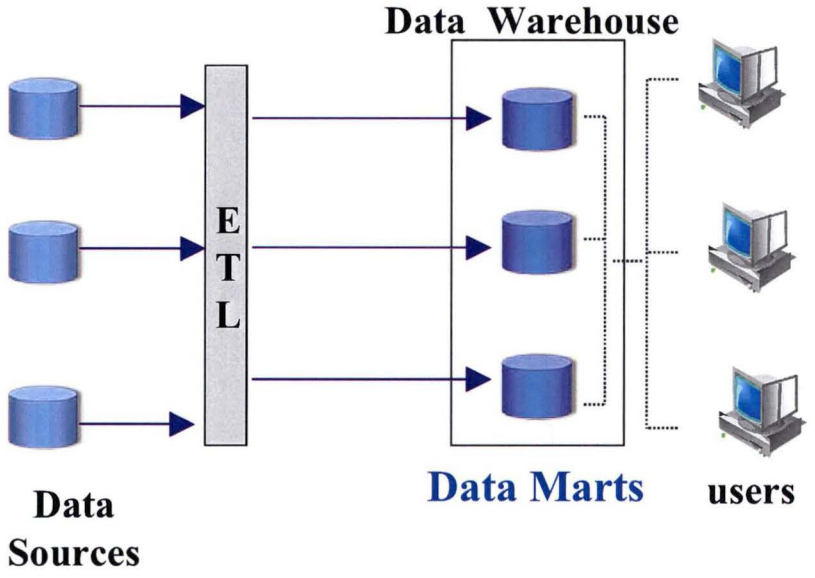


Figure 17: Interdependent data marts without physical data warehouse

9 IMPLEMENTATION METHODOLOGIES

There are two basic methodologies for implementing a three tier data warehouse, the "top-down" and "bottom-up" approaches. The "top-down" approach mandates the construction of an enterprise data warehouse first and then the distribution of subset data marts from that parent data warehouse. The "bottom-up" approach uses a series of incremental, architected data marts to build up toward the goal of the enterprise data warehouse.

Although the "top-down" strategy was favoured in early initial enterprise data warehouse projects and is the most elegant design approach, high failure rates for initial enterprise data warehouse projects have led the majority of current projects to use the "bottom-up"[8] approach. [HACKNEY,1998]

[8] Depending on the authors, the hybrid approach is also called "The bottom-up" approach. The bottom-up approach described by Hackney corresponds to the hybrid approach defined in this document.

"For reasons such as lower costs and risks, the DM solution is gaining in popularity. According to a survey done by the META Group, data mart projects are estimated to make up the majority of decision support systems within the next year. Since data marts are subject specific and smaller in scope, the results are seen much sooner. Furthermore, it is usually much easier to justify costs for a data mart". [TANRIKORUR,1998]

9.1 Top-Down architecture

Introduced by Bill Inmon, this is the first data warehousing architecture.

In this approach, the data warehouse is typically built in an iterative manner, business area by business area, and underlying dependent data marts are created as required from the data warehouse contents.

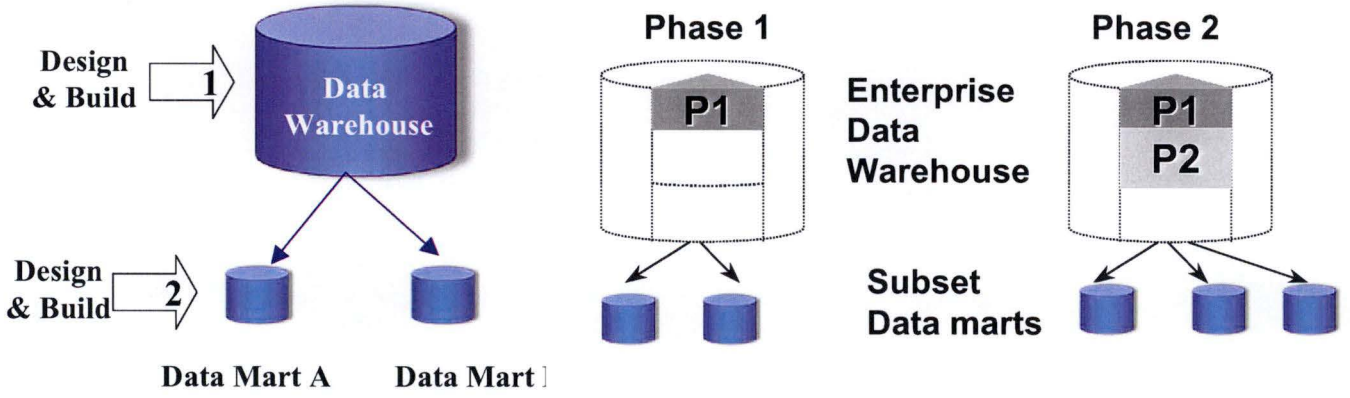


Figure 18: Top-down approach[9]

Even if enterprise data warehouse are built in an iterative manner, it typically takes 15 or more months to bring the first subject area to a production status. This is a very, very long time to maintain political and budget support in the face of ever-shifting priorities, emergencies and changing staff.

9.2 Bottom-up architecture

Bottom-up architecture became popular because Top-down architecture took too long to implement, was often politically unacceptable, and was too expensive [FIRESTONE,1998].

In the bottom-up approach, independent data marts are created with the view to integrating them into an enterprise data warehouse at some time in the future.

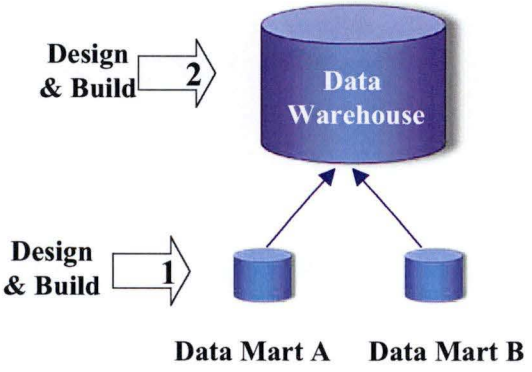


Figure 19: Bottom-up approach

[9] Figures in this chapter based on reference [HACKNEY,2000]

Data marts are built by focusing on one specific business area. This means that, because of its smaller scope, managing user expectations is easier and you are more likely to be on time and budget. A drawback, however, is that multiple data marts become isolated systems. Since each of them requires an extraction of operational data from different sources, they begin to cause redundancy. As their number grows larger, the total time spent managing them becomes longer than envisaged. The biggest drawback is the lack of data integration for the whole company. In other words, "migration" to a common data warehousing model will be harder.

While bottom-up architecture was quite successful in meeting initial expectations in building data marts, it quickly came to be perceived as unacceptable for the long term. The main reason was that it failed to provide a common Meta data component. Without shared Meta data, it is difficult to construct the data warehouse from data marts. So, bottom-up architecture, in its pure form, fails to fulfil its promise of an incremental approach to data warehouse.

9.3 Hybrid architecture

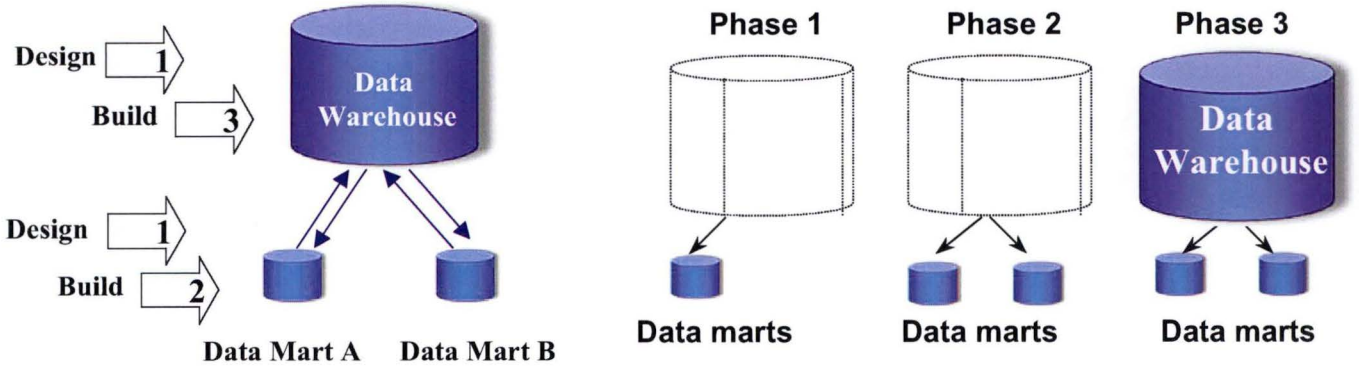


Figure 20: hybrid approach

What organisations require is a solution that offers a low cost and rapid return on investment advantages approach without the problems of data integration in the future.

The hybrid approach recommends developing a subset of an enterprise data model first before developing the first data mart.

The hybrid approach relies on synchronising the enterprise data warehouse models and data marts models, as well as the differences between them. This lets local groups, for example, develop their own definitions or rules for data elements that are derived from the enterprise model without sacrificing long-term integration.

After deploying the first few “dependent” data marts, the data warehouse is created by transferring detailed data from the data marts to the data warehouse. The consolidation of redundant data feeds saves the organization time, money and processing resources. Organizations typically roll-up a data warehouse once business users request views of detailed data across multiple data marts.

The top-down approach encourages the design and building of a data warehouse first and then the data mart. In the bottom-up approach, this order is reversed. It is only the hybrid approach that starts the design at both levels. It then continues to build data marts and finally rolls them up into the data warehouse level.

The major danger of this approach is the creation of LegaMarts (see section 8.3). Since data marts are easy to implement, one may be tempted to save time by building a data mart that does not comply with the common framework.

9.4 Comparison of the different architectures

The table below compares some of the different architectures already described in Section 8:

	Data warehouse	Data mart		
		Dependent	**Independent**	**Interdependent**
Scope	Enterprise-wide	Subject or department		
Basic architectural assumption	Centralised	Centralised	Distributed	distributed
Size	Up to several terabytes	Megabytes to gigabytes		
Approach	Top-down		Bottom-up	hybrid
Data source	Operational systems	Underlying enterprise data warehouse	Operational systems	Operational systems
Level of data	Detailed	Summarised	Detailed and summarised	Detailed and summarised
Next level of migration	data mart	Not applicable	data warehouse	data warehouse
Distinguishing characteristics	Enterprise wide	Derived and aggregated	Independent	Share, conformed dimensions
Benefits	Integrated	Quicker response times than DW		Integrated by design, shorter development time
Drawbacks	Long development time	Requires enterprise data warehouse	Fragmentation of data, integration issues, multiple extracts for update/refresh	

Figure 21: Comparison of the different architectures [10]

[10] Source: [POLENIS,2002]

10 DATA MODELLING

In the data warehouse modelling world, there are two different streams of thought: experts who promote Relational modelling as the best modelling approach for data warehouse modelling, and other experts who promote Dimensional modelling (see section 11). These latter experts consider Relational modelling to be unusable for data warehouse modelling because, in their view, it is too technical and too complex for end users. This is debatable, however.

Other authors recommend choosing between dimensional modelling or relational modelling, depending on the type of analysis to be performed, e.g. ad hoc query, OLAP analysis, data mining. As data warehouse is enterprise focused, it should support all kind of analysis and then should be based on a relational modelling while data marts used for OLAP analysis should be implemented using dimensional modelling.

Relational modelling is an ideal support for studying data structure: simple, intuitive, rigorous mathematical based and accessible.[HAINAUT,2001].

Relational modelling, as already stated, is a proven and reliable data modelling approach. The normalisation rules yield a stable, consistent data model which is flexible in how the data are later analysed by the data marts.
The resulting database is

- Reliable across the business: it contains no contradictions in the way that data elements or entities are named, related to each other or documented.
- Flexible in the types of data analysis it supports
- Correct across the business: the model will provide an accurate and faithful representation of the way the information is used in the business.
- Adaptable to change: the database will be able to accommodate new elements and entities while maintaining the integrity of the implemented ones.

The resulting database is the most efficient in term of storage and data loading as well

IMHOFF defines different steps to transform an enterprise relational model to a data warehouse model [IMHOFF,2003].

1 Select the data of interest: data elements are selected based on the business requirements.

2. Add time to key: time is added to key to accommodate the historical perspective of data warehouse. Adding time is not just about adding a date column, introducing time in a database increase the complexity in the data management as well as in their exploitation. See [FUNDP,2002] for detailed information on how to deal with temporal database.

3. Add derived data: derived data are data that result from performing a mathematical operation on one or more other data elements. The two major reasons to add derived data is to ensure consistency and to improve data delivery performance. Consistency is the most important reason, as one of the common objectives of a data warehouse is to provide data in a way that everyone has the same facts and the same understanding of the facts. For example, the number of days in a month can have multiple meaning (all days, weekdays, working days..).

   Creating a derived field does not usually save space since each of the components used in the calculation may still be stored, but it improves the data delivery performance at the expense of the load performance. Derived data can be calculated at different moments: when loading the data warehouse, when loading the data marts or calculated in the end-user tool. If the derived field is needed to ensure consistency, it should be calculated at the data warehouse level, so that the same definitions and algorithms are used in all the data marts. Another advantage is that the calculation is only performed once.

4. Determine granularity level: it ensures that data exist to address business needs.

11 DIMENSIONAL MODELLING

11.1 Dimensional modelling terminology

To understand dimensional data modelling, it is necessary to define some of the terms commonly used.

A dimensional model includes Fact tables and Dimension tables. Its purpose is to improve performance by matching data structure to the queries.

Fact table contains the information related to factual events, e.g. number of flights. The Fact table contains the measurements, metrics or facts of business processes. The only other things a fact table contains are foreign keys for the dimension tables.

Dimension table *contains c*ontext of the measurements. You can also think of the context of a measurement as the characteristics such as *who, what, where, when, how* of a measurement (subject). Dimension data most often contain descriptive textual information e.g. type of aircraft.

The **Dimension Attributes** are the various columns in a dimension table. In the Time dimension, the attributes can be Year, Month, Week.

Dimensions can have one or more **hierarchies**. A Time dimension, for example, could have a Calendar year hierarchy and a Fiscal year hierarchy.

The hierarchies in dimensions have levels which can be used to view data at various levels of detail. A **level** is a position in a hierarchy. For example, a time dimension might have a hierarchy that represents data at the day, month, quarter, and year levels.

The relationship between hierarchy level is N:1. An attribute hierarchy determines a sequence of functional dependencies.

For example
Day → Month
Month → Quarter
Quarter → Year

Dimension attributes are used as the source of most of the interesting constraints in data warehouse queries. They are virtually always the source of the row headers in the SQL answer set.

Attribute hierarchies are used to analyse facts at different aggregation level.

11.2 Star schema/Snowflake schema

The Dimensional model is represented as a **Star Schema** or **Snowflake Schema**.

11.2.1 Star Schema

In the star schema design, a single object (the fact table) sits in the middle and is radially connected to other surrounding objects (dimension tables) like a star. The dimension tables contain information on particular attributes in the fact table.

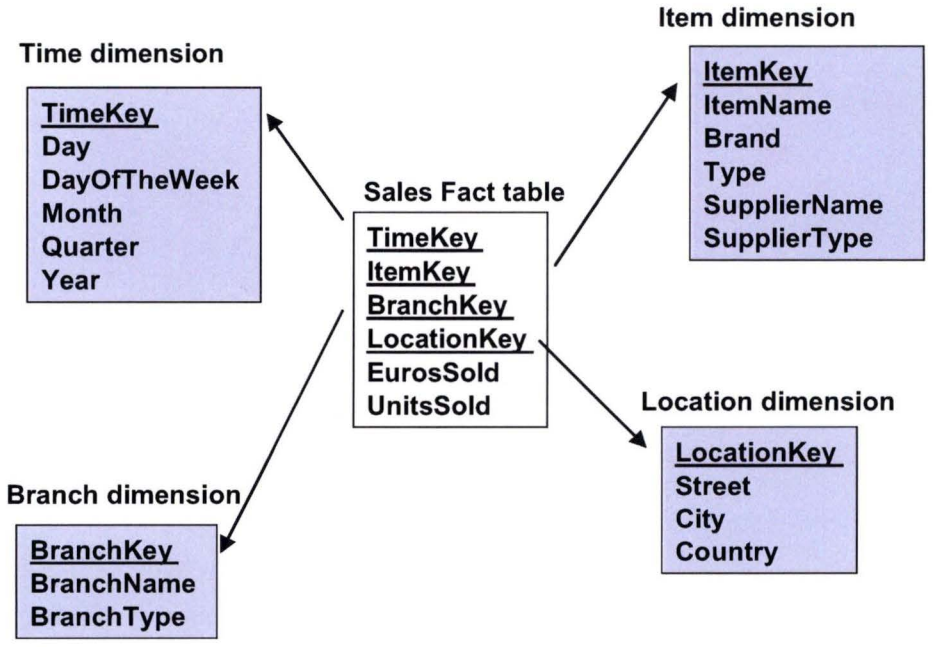


Figure 22: Star schema

Normalisation is a relational database modelling where tables are broken down progressively into smaller tables to a point where all attributes in a table fully depend on the primary key, and on no proper subset of it. This minimises redundancy and the overall size of the database.

A fully normalised data model can perform very inefficiently because of multiple joins needed in queries. For this reason, normalised logical data models are sometimes converted into a physical data model which is significantly de-normalised

In a star schema, dimensions are often de-normalised. For example, Airport and Country dimensions are collapsed into the new Airport dimension.

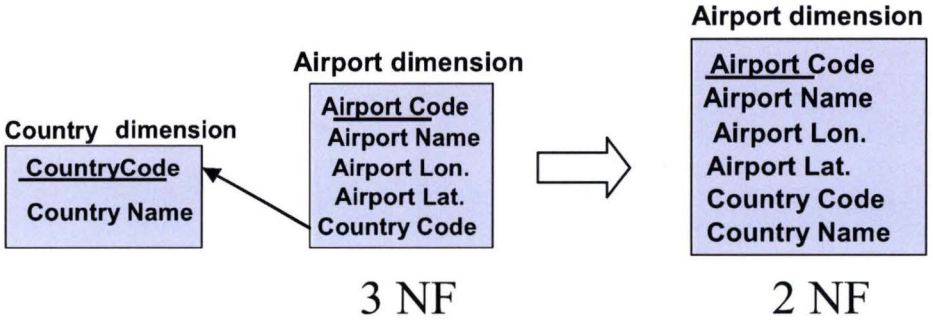


Figure 23: Dimension de-normalisation

As seen in Figure 23, collapsing the Airport and Country dimension, changes the normal form of the table from the third (3 NF) to the second (2 NF) normal form (there is a transitive dependency between Airport Code and Airport Name)

The collapsing causes redundancy but can sometimes improve performance.

11.2.2 Snowflake schema

The star schema is a very simple database design, which clearly presents the multidimensional character of the data and allows for rapid querying of the data in a data warehouse.

In snowflaking, some of the fields of the dimension tables are split off into separate tables. These

dimension tables are "normalised". The star schema presented in Figure 22 then becomes the following snowflake schema.

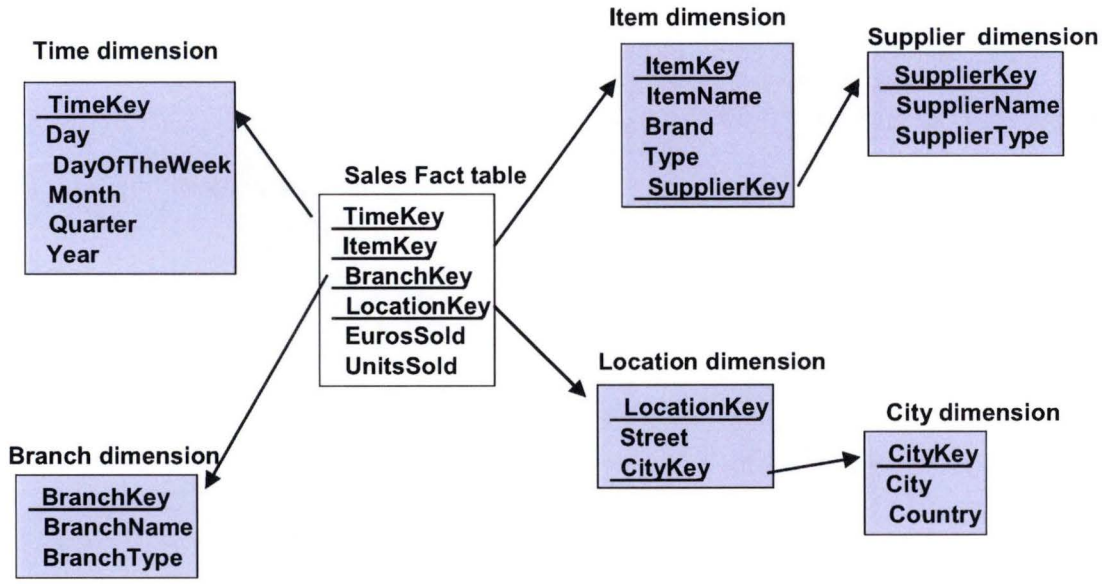


Figure 24: Snowflake schema

11.2.3 Fact constellations

Multiple fact tables share dimension tables, viewed as a collection of stars, and are therefore called galaxy schema or fact constellations.

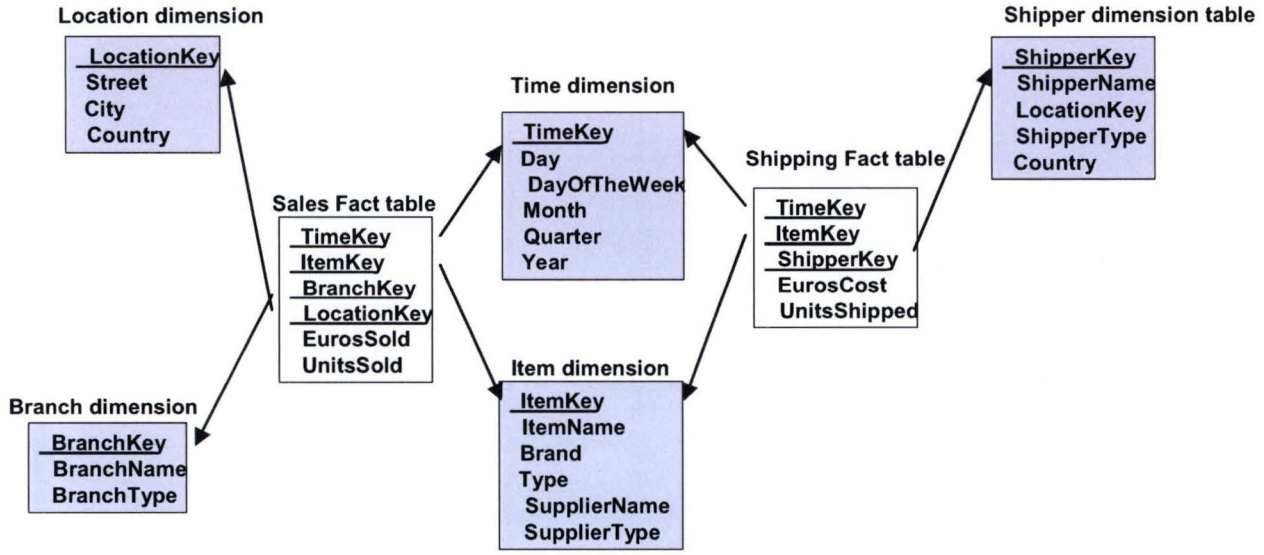


Figure 25: Fact constellations

11.3 Additivity of facts

The word "additivity" means the ability to sum the measures presented in a fact table during drill-up operations according to presented hierarchies.

A measure is non-additive if it cannot be summed in any dimension, such as averages and ratios.

A measure is additive if it can be summed according to every hierarchy presented in any dimension. For example, duration of a flight is an additive measure as it can be summed by time hierarchy and by airspace hierarchy.

A measure is semi-additive if it can be summed according to hierarchies in some dimensions but not in all of them. For example, flights are additive by time hierarchy but not by airspace hierarchy. The number of flights in Europe is not the sum of the number of flights in the countries which comprise Europe, as one flight crosses many national boundaries.

11.4 Surrogate key

In the physical database, sequenced integers called surrogate keys are often substituted for the primary and foreign keys. Since surrogate keys are usually smaller than natural keys, they have the effect of speeding up queries.

In star schema, surrogate keys are important because the size of the fact table is critical to performance.

A primary key is a unique key for each record in a dimensional table. The primary key of each dimension in a star schema is replicated in its fact table where it is referred to as a foreign key.

The dimension table should have a surrogate key as well as a natural key. A natural key serves as the row identifier from the business point of view.

Meaningful key and operational keys should be avoided as operational information can change, for example an operational key could be reused in time.

11.5 Slowly changing dimension

The key purpose of a data warehouse/mart is to collect historical data. Analysis requires dimensional context, which sometimes includes change over time.

Business rules should be defined on how to handle an attribute value that has changed value. The raw data has an operational key (natural key) value which must be matched to the same field in the current dimension table. If the dimensional information has changed, there are 3 possibilities [KIMBALL,1998]

Type 1: rewrite history
Overwrite the current attribute

	Surrogate key	Natural key	Attribute
Current record	101	EBBR	~~Brussels~~ Zaventem

Type 2: keep every change

Create a new dimension record with a new surrogate (meaningless) key

	Surrogate key	Natural key	Attribute
Current record	101	EBBR	Brussels
New record	145	EBBR	Zaventem

The surrogate key for type 2 slowly changing dimension can be found using the following SQL statement:

Select max (surrogate key) from DimensionTable where Natural key = FactSource.NaturalKey

If there is a need to reload historical facts, this technique will not work as it provides only the current record. To resolve this problem, A "FromDate" and "ToDate" fields can be added to the Dimension table. The surrogate key can then be found using the following SQL:

Select surrogate key from DimensionTable where NaturalKey = FactSource.NaturalKey and FactDate between DimensionTable.FromDate and DimensionTable.ToDate.

	Surrogate key	Natural Key	Attribute	FromDate	ToDate
Current record	101	EBBR	Brussels	01/01/1997	12/03/2003
New record	145	EBBR	Zaventem	12/03/2003	99/99/9999

Type 3: keep a limited set

Push down the changed value into an "old" attribute field. In this case it is anticipated that a user needs to refer either to the old value of the attribute or the new values

	Surrogate key	Natural key	Attribute	Old Attribute
Current record	101	EBBR	Zaventem	Brussels

Applying changing dimensions

For type 2 changing dimensions, the fact table always pointed to the instance of the dimension which was true at the time.

Fact table

101		
145		

Airport dimension

Surrogate key	Natural Key	Attribute	FromDate	ToDate
101	EBBR	Brussels	01/01/1997	12/03/2003
145	EBBR	Zaventem	12/03/2003	99/99/9999

Historicity of the dimension is represented here with a valid time defined as a "from date" and a "to Date". The end date is defined as the first date when the state has disappeared. It is a [FromDate, ToDate[interval. See [FUNDP,2002]

12 OLAP TOOL

12.1 Introduction

The term *on-line analytical processing* (OLAP) was introduced in 1993, by Dr Codd, the father of the relational database, as a special category of data processing aiming at *intuitive*, *interactive*, *multidimensional analysis* of integrated data for decision support [CODD,1993].

He outlined 12 rules:

1 Multidimensional conceptual view of data
2 Transparency
3 Accessibility;
4 Consistent reporting performance
5 Client/server architecture
6 Generic dimensionality
7 Dynamic sparse matrix handling
8 Multi-user support
9 Unrestricted cross-dimensional operations
10 Intuitive data manipulation
11 Flexible reporting
12 Unlimited dimensional and aggregation levels

Dr. Codd's 12 rules were not universally accepted. Consequently, the OLAP Report, an independent service, produced the FASMI test (Fast Analysis of Shared Multidimensional Information).

FAST means that the system is targeted to deliver most responses to users within about five seconds, with the simplest analyses taking no more than one second and very few taking more than 20 seconds.

ANALYSIS means that the system can cope with any business logic and statistical analysis that is relevant for the application and the user, and keep it easy enough for the target user.

SHARED means that the system implements all the security requirements for confidentiality (possibly down to cell level) and, if multiple write access is needed, concurrent update locking at an appropriate level. Not all applications need users to write data back, but for the growing number that do, the system should be able to handle multiple updates in a timely, secure manner.

MULTIDIMENSIONAL is the key requirement. The system must provide a multidimensional conceptual view of the data, including full support for hierarchies and multiple hierarchies, as this is certainly the most logical way to analyze businesses and organizations.

12.2 Characteristics

Multidimensional modelling is a powerful conceptualisation technique used in On-line Analytical Processing OLAP applications. Its main advantages are that it is close to analysts' way of thinking and improves query performance.

A user-friendly way to visualise the multi-dimensionality of data is to use data cubes. A data cube is defined over a multi-dimensional space which consists of several dimensions representing the various perspectives of data (e.g., product, region and time). Data cube cells represent the numeric facts.

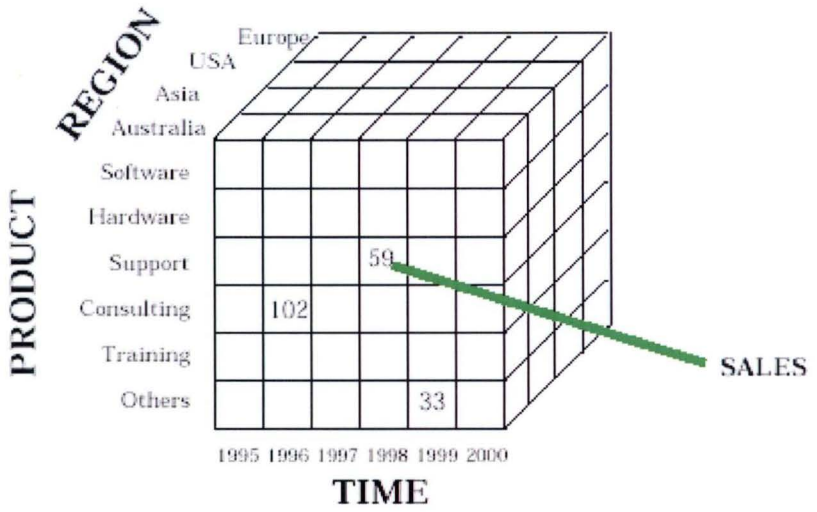


Figure 26: Multi-dimensional data cube

Multi-dimensional analysis using OLAP tools is a dynamic process, whereby users "navigate" across multi-dimensional data structures. Typical operations used during this process are:

- pivoting (rotation): allows for the reorientation of the view by interchanging individual dimensions of the cube.

- drill up (roll up) & drill down: drill up corresponds to aggregating across a hierarchy, i.e. performing a further "group-by" on one of the dimensions. Drill down is the opposite, it corresponds to navigating from aggregated to detail data e.g. Drill down: from year to month, e.g. Drill up: from month to year.

- *Slice & dice:* corresponds to reducing the dimensionality of the data, taking a projection of the data on a subset of dimensions for selected values of the other dimensions and dice allow to pivoting this selection. Slice performs a selection on one or more dimensions of the given cube. For example, see Figure 27, selection "Region = Australia or Region = Asia" AND "Product = Soft or product = Hardware"[11]:

[11] Note that the literature contains different definitions of "slice & dice".

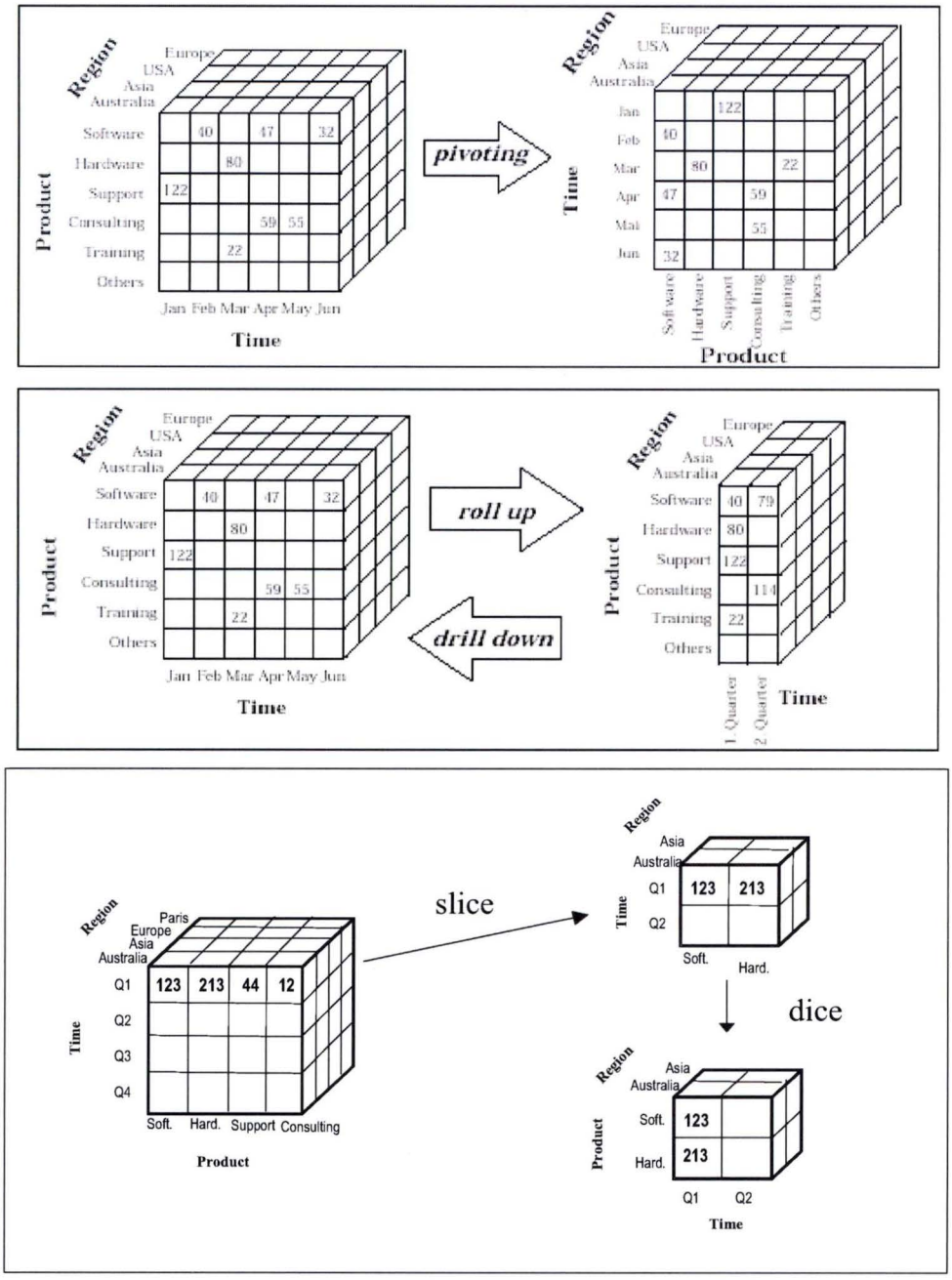


Figure 27: OLAP operations[12]

12 Source: [VAVOURAS,2002]

12.3 Dimensional data Storage

A dimensional data model can be implemented using a familiar RDBMS, or based on proprietary database technology [BERSON & SMITH,1997].

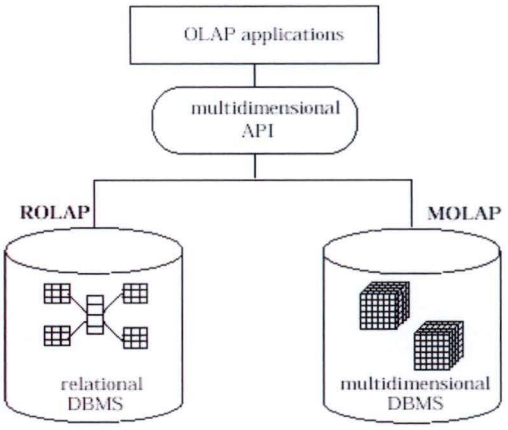


Figure 28: ROLAP and MOLAP

12.3.1 ROLAP

In the *ROLAP* architecture (relational on-line analytical processing), data warehouses are implemented on relational DBMS.

12.3.2 MOLAP

In contrast, in the *MOLAP* architecture (multidimensional on-line analytical processing), data are stored in "pure" multidimensional databases.

MOLAP servers typically store data in multidimensional arrays. Each dimension of the array represents the respective dimension of the cube. The contents of the array correspond to the measure(s) of the cube.

The MOLAP architecture works well with few dimensions. On the other hand, ROLAP is the more efficient way of storing large volumes of data.

12.3.3 Other approaches

Other approaches exist in addition to the ones described above. These include: HOLAP (HYBRID OLAP), DOLAP (Desktop OLAP) etc.

12.4 OLAP tool vendors

A multitude of commercially available tool and tool suites with largely varying characteristics exists. The product landscape is constantly changing due to company mergers, new product releases etc

The OLAP tools are composed mainly of large enterprise application vendors (e.g., Oracle, Microsoft, SAP, PeopleSoft) and medium-sized independent analytic specialists (e.g., Business Objects, MicroStrategy).

Microsoft Analysis Services is typically chosen by smaller organizations, while SAP BW and MicroStrategy are much more likely to be found in the largest organizations. Similarly, Business Objects and SAP are relatively stronger in Europe, while the MicroStrategy and Hyperion customer bases have a North American bias (this is particularly true of the former Brio customer

base). The large non-specialist vendors, such as Microsoft and Oracle, are stronger in the rest of the world than the smaller BI specialists, who tend to be under-represented outside the major markets.

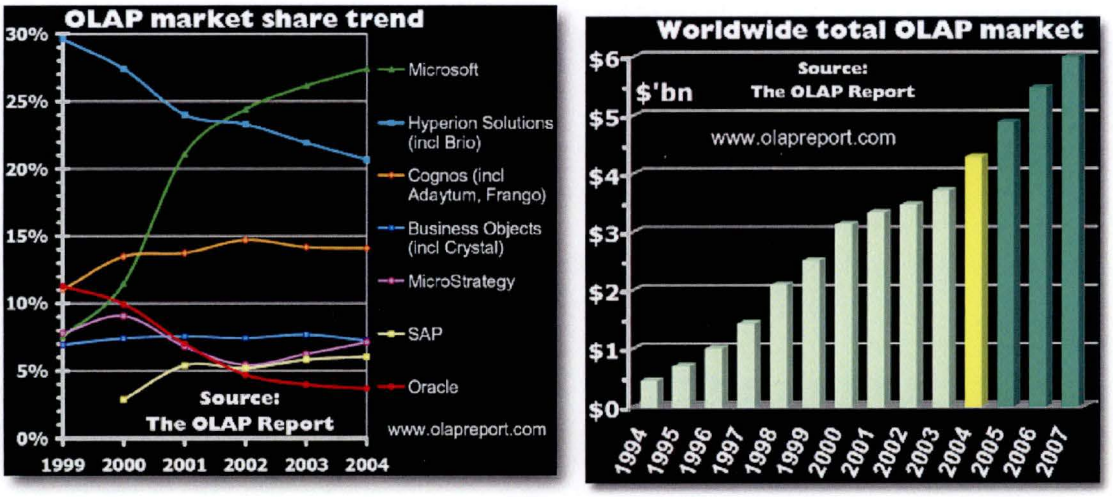


Figure 29: OLAP market [PENDSE,2005]

SECTION 2: PRU DATA MART PROJECT

13 INTRODUCTION

In order to present the Performance Review Unit (PRU) data marts project, I have used the lifecycle presented by R. Kimball in his book "*The data warehouse lifecycle toolkit*" [KIMBALL,1998]. This lifecycle is shown in Figure 30 below.

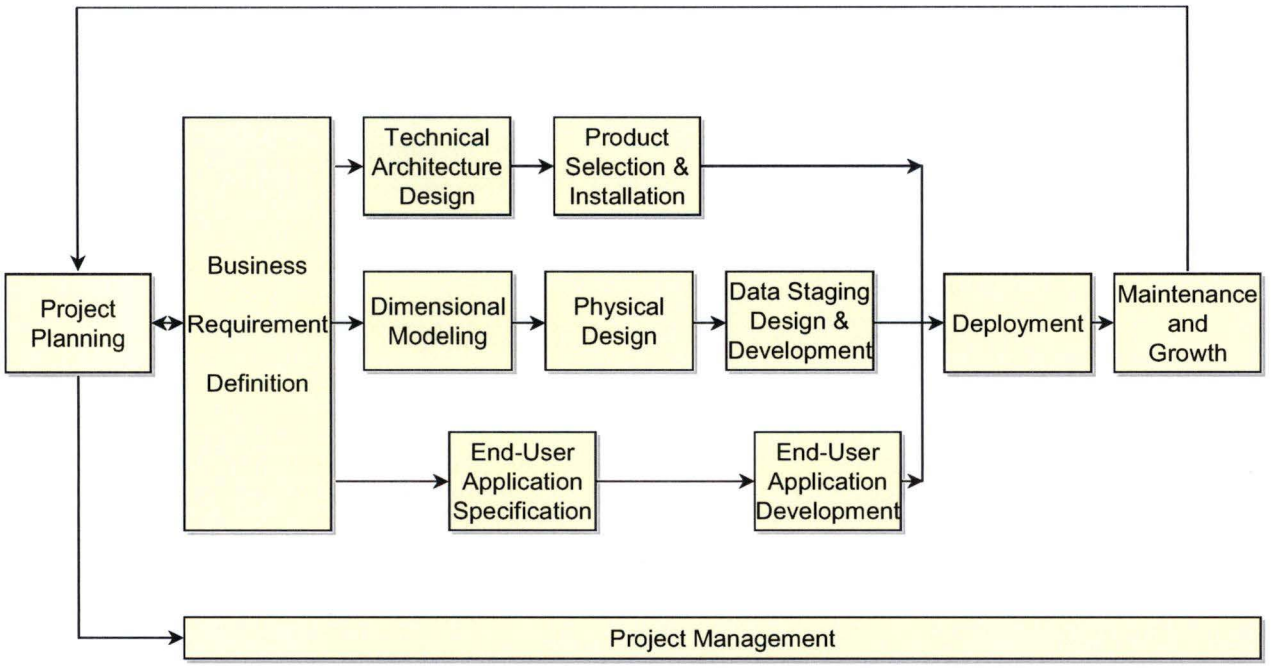
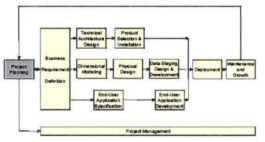


Figure 30: The business dimensional lifecycle diagram

Not all the tasks defined in this lifecycle are covered, since not all are applicable to this relatively small data marts project.

14 PROJECT PLANNING



14.1 Project definition

14.1.1 Background

EUROCONTROL is the European Organisation for the Safety of Air Navigation. This civil and military Organisation, which currently numbers 34 Member States, has as its primary objective the development of a seamless, pan-European Air Traffic Management (ATM) system that fully copes with the growth in air traffic, while maintaining a high level of safety, reducing costs and respecting the environment.

The Performance Review Commission (PRC) is the independent advisory body on ATM performance to EUROCONTROL. It makes recommendations to ensure the effective management of the European Air Traffic Management System through a strong, transparent and independent performance review and target-setting system. This system addresses all aspects of air traffic

management including policy and planning, safety management at and around airports and in the airspace, as well as financial and economic aspects of services rendered.

The PRC was established in 1998 and is composed of twelve independent "wise people" with senior managerial and technical experience of aviation.

The PRC is supported in its work by the Performance Review Unit (PRU) which is composed of 10 persons, who are mainly operational and economic experts.

The PRU produces Performance Review Reports on an annual basis, and on an ad-hoc basis where warranted, to assist all stakeholders in understanding why, where, when, and possibly how, ATM performance should be improved, in knowing which areas deserve special attention, in terms of effort or expertise, and in learning from past successes and mistakes. The spirit of these reports is to help everyone involved to effectively improve performance in the future.

Measuring the performance of the European ATM system was the first task of the PRU. In 1999, it developed an initial performance measurement system, consisting of Key Performance Areas (KPA) and associated Key Performance Indicators (KPI). Thus far, the PRU has focussed on three KPAs: Safety, Delays and Cost-effectiveness.

14.1.2 Key performance areas and indicators

KPA	Definition
Safety	The conformance of air transport to specified safety targets
Delay	The time in excess of the optimum time that it takes a user to complete an operation.
Cost Effectiveness	The value for money that users receive from the supply of air traffic services.
Predictability	The ability of a user to predict variation and to build and maintain optimum flight schedules.
Access	The accessibility of airspace, ATM services and airport facilities under controllable conditions.
Flexibility	The ability of ATM to accommodate changing user needs in real time and without penalty.
Flight Efficiency	The ability of the ATM system to allow a user to adopt the preferred flight profile in terms of flight level and route.
Availability	The availability of critical ATM resources and of the ATM services provided to users.
Environment	The conformance of air transport to environmental regulations.
Equity	Equity of treatment of flights by all aircraft operators within and between specific classes of users.

Some KPAs are constrained by minimum standards or regulatory limits which are imposed by external parties (e.g. safety), while others allow tradeoffs (e.g. delays and cost-effectiveness).

In addition to quantitative measures, there may be a need to evaluate qualitative measures of performance, e.g. users' satisfaction, by means of surveys carried out at regular intervals.

Aggregation of indicators

Several breakdowns of performance items may be needed to understand complex ATM performance issues, e.g. Europe-wide, by State, by Area Control Centre (ACC), by Air Navigation Service Provider (ANSP), by airport, by reference location, by time-series, by phases of flight, by city pair, by airspace user. Some views will be more relevant to some indicators than others. It is intended to report on indicators at an appropriate level for each indicator. For example, the preferred level at which to measure service cost would be the ANSP provider level, which, in most cases, corresponds to the present reporting level, i.e. the State.

Delay Key Performance Indicators

The project covers only data which allow the KPI to be computed for the Delay KPA, and more precisely for ATFM delay.

The aim of the Delay KPA is to ensure that ANSPs have the right level of ATC capacity in relation

to expected demand for ATC services so that delays are maintained at an acceptable level from the user perspective.

A number of indicators of delay are well established and currently reported by the EUROCONTROL central Office for Delay Analysis (CODA) and the Central Flow Management Unit (CFMU). All current delay analyses are based on departure delays. Airborne and arrival punctuality has to be considered as well as a breakdown of different types of delay.

Departure Delay

Departure delay is the difference between actual off-blocks time and the scheduled departure time. Departure delay may have many possible causes and is only partly under the control of ATM. The main cause relevant to ATM is air traffic control flow restrictions (ATFM delay).
The proposed indicator is:
Total minutes of departure delay / Total number of flights
where total minutes of departure delay are accumulated over all flights.

An important indicator for the general severity of the delays for different ATM systems is the balance between the total traffic and the delayed traffic.
The proposed indicator is:
Total number of delayed flights / Total number of flights

Departure delay in relation to the number of flights that are delayed is a measure of the seriousness of delay to an aircraft operator and its passengers. Short delays can be absorbed quite easily but as the delay grows disruption will increase.
The proposed indicator is:
Total minutes of departure delay / Number of delayed flights

14.1.3 Data sources

The PRU's analytical work is based on data coming from different sources that are both internal and external to EUROCONTROL.

As can be seen from Figure 31, the main source of Traffic and Air Traffic Flow Management (ATFM) delays analysis is the EUROCONTROL Central Flow Management Unit (CFMU). The CFMU manages the flow of traffic to assist Air Traffic Control (ATC).

Air Traffic Flow Management protects Air Traffic Control (ATC) services from potentially dangerous overload and also minimises the effects of airspace congestion on aircraft operators.

At times, the capacity of controlled airspace in certain areas or at airports is not sufficient to accommodate traffic demand. A regulation over these areas or airports is requested by the ATC in order to avoid overload.

ATFM measures principally consist of rerouting aircraft over non-congested areas and staggering departure times by imposing appropriate ground delays (slot allocation).

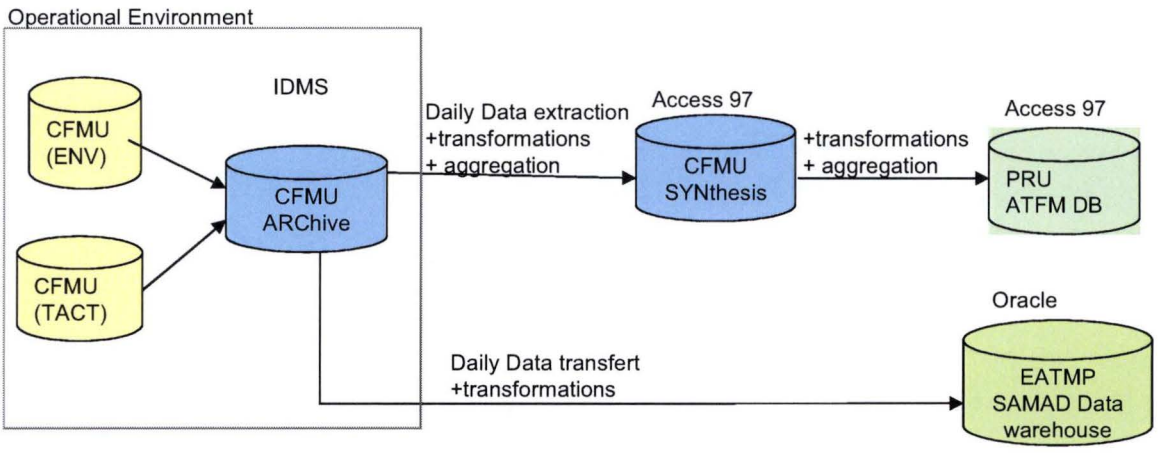


Figure 31: CFMU data sources

The CFMU collects the flight plans of every aircraft planning to fly in Europe (TACT DB).

The ENV database contains comprehensive details of the CFMU area. It includes airports, ATC sectorisation, etc.

The ARChive database records CFMU flight data and processes them to provide historical data.

The CFMU synthesis DB contains tables where daily aggregate counts are stored, since mid 1996. These aggregations include counts of flights, regulations, entered airspaces etc.

The SAMAD data warehouse was created a few years ago to be the common data sources for European Air Traffic Management Program (EATMP) teams. It contains, *inter alia*, Flights plan Information and Environment data. This data warehouse is updated daily with information coming from the CFMU.

The European Air Traffic Management Program (EATMP) is responsible for all development activities relating to the European airspace and its utilisation. These activities include the development of delay indicators (CODA), air traffic statistics and forecasts (STATFOR).

The PRU ATFM DB is the result of a transformation and aggregation of the CFMU Synthesis database which allow a better response to the needs of the PRU analysts.

Up to November 2002, the PRU based its Traffic and ATFM delay analysis on the PRU ATFM DB for requests for aggregated information and on the SAMAD database for more detailed information (Information at flight level). The data were available to the PRU analysts through Excel pivot tables created by the PRU Team.

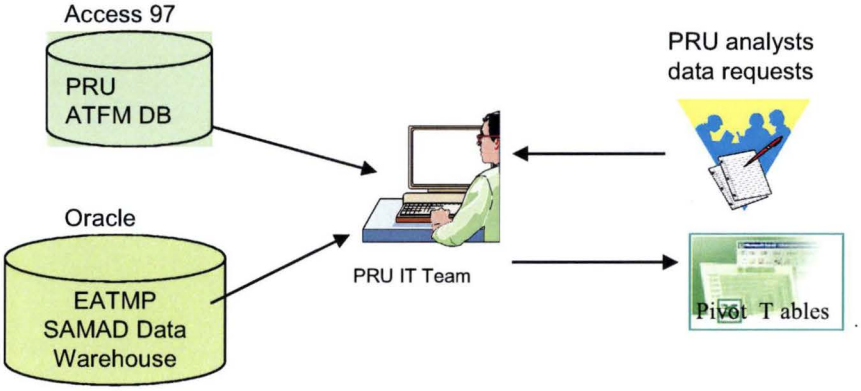


Figure 32: Situation up to November 2002

Since the migration of the CFMU from IDMS to Oracle, the CFMU synthesis database has ceased to exist and consequently the PRU ATFM DB is no longer updated.

The CFMU has created a data warehouse and data marts (see Figure 33) to replace the CFMU Synthesis database.

Since the PRU was not sure of the extent to which it would be granted access to the new CFMU data warehouse and data marts environment, the PRU decided to create its own data marts based on the SAMAD data warehouse in order to:

- add refinements that are only relevant to PRU needs (e.g. information at region level);
- allow more flexibility (change request to CFMU can take more than a year);
- allow the analyst to more easily correlate information from different sources (e.g. financial data with operational data).
- have 2 sources (SAMAD, CFMU) of information (in case access to one of the sources is not possible).

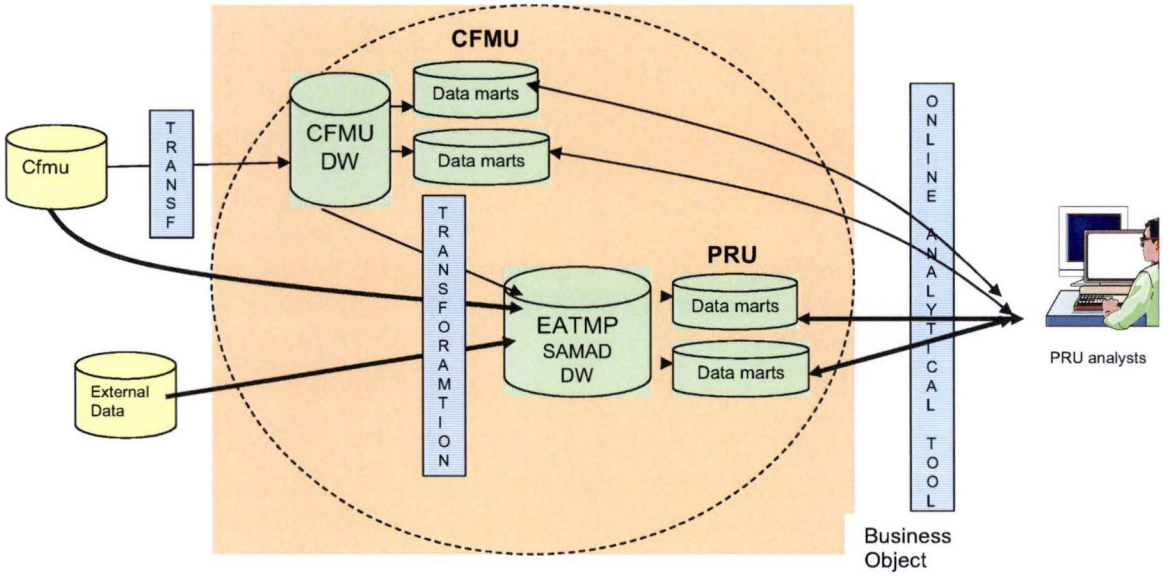


Figure 33: Future situation

14.2 Project scope

The PRU data marts will only be based on operational data derived from the CFMU. The SAMAD data warehouse will be used as the source of these CFMU data.

The data should cover at least the type of data available in the CFMU Synthesis database (see Figure 36 for the list of measures available in the CFMU Synthesis database).

Data in the CFMU Synthesis database are available since April 1996. Data should be available in the PRU data marts from 1997 onwards.

14.2.1 Exclusions from Scope

The following indicators do not form part of this project because data are not readily available to compute them.

- Km flown in airspace
- Data at the airspace level controlled by the Air Navigation Service Providers (ANSP)

A future project will be the integration of ANSP levels which will enable financial data (collected

from ANSP's by the PRU) to be correlated with operational data (collected at operational airspace unit level by the CFMU).

14.2.2 Success criteria

Several key success criteria have been designated:

- Provide easy access to data for the analysts
- Have access to up to date information
- Provide a single interface for the analysts to access the data

14.2.3 Risks

- Learning curve of new tools for the analysts
- Inconsistency between indicators produced by the CFMU

14.3 Project planning and management

Front Office	
Sponsor	PRU head of unit
Coaches	
Project Manager	PRU Team
Business Project Lead	PRU Operational expert
Regular Line up: Core Project Team	
Business Systems Analyst	PRU Team
Data Modeller	PRU Team
Data Warehouse database Administrator (DBA)	EATMP Data Warehouse team
Data Staging Designer & Programmer	PRU Team
End User Application Developers	PRU Team
Data Warehouse Educator	PRU Team
Special Teams	
Technical/Security Architect	EATMP Data Warehouse team
Technical Support Specialists	EATMP Data Warehouse team
Data Steward	EATMP DW team & PRU Team
Data Warehouse Quality Assurance Analyst	EATMP DW team & PRU Team
Fans	
Business Users by Group/Function	PRU analysts

The PRU team comprises two persons: a consultant and me (the PRU technical assistant). The consultant is specialised in Oracle 9 database, has previous experience in a data warehouse project but has no background knowledge of aviation. I have very good knowledge of aviation and of the PRU experts' data requirements.

All the project phases will be followed by the PRU team. Contact with other units in EUROCONTROL is coordinated by me. The data staging design and program is driven by the consultant and the end-user application by me.

I am mainly involved in collecting the user requirements, finding and validating the data source and in the development of the end-user application interface and pre-defined reports.

Depending on the current PRU team workload and on the data quality and availability in the SAMAD data warehouse, the project was planned to be completed within 6 months, in order to be ready for work starting on the next performance review report.

Every month, a meeting is held with the PRU operational expert to inform him on the progress of the project and on the direction chosen.

15 TERMINOLOGY

The terminology used in subsequent chapters is explained here.

As mentioned before, at times, the capacity of controlled airspace in certain areas or at airports is not sufficient to accommodate traffic demand. A regulation over these areas or airports is requested by the ATC in order to avoid overload.

The Central Flow Management Unit (CFMU) can delay departure times by imposing a take off time (slot allocation) to the flights in order to limit the number of flights allowed in the regulated areas or airports. This can result in ATFM delay.

"ATFM delay" is defined as the duration between the last Take-Off time (ETOT) requested by the aircraft operator and the Take-Off slot (CTOT) given by the Central Flow Management Unit (CFMU).

A flight is called a delayed flight if it has ATFM delay.

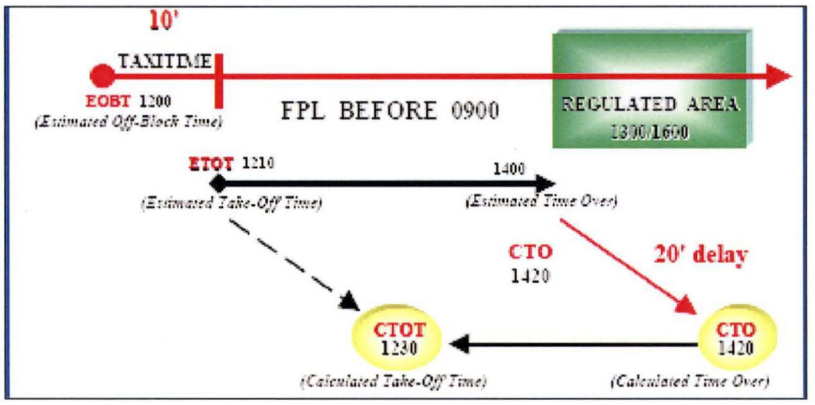


Figure 34: ATFM delay

If a flight is subject to several regulations, it is given the delay of the most penalising regulation and is forced with that delay in all the other regulations.

A regulated flight is a flight affected by a most penalizing regulation

A regulation is put to protect a reference location which is an airport or a sector. A regulation is considered to generate airport delay if its reference location is an aerodrome and en-route delay if is reference location is a sector.

Each regulation has a reason associated to it (e.g. bad weather).

A regulation affects the time of departure of flights entering a defined traffic volume. A flight is submitted to the regulation when it follows the traffic volume and the estimated time of over flight/entry is between the regulation start and end time.

Each flight is associated with a traffic volume profile which describes the path (represented in four dimensions) that an aircraft is expected to follow between the departure and the arrival airport in terms of traffic volumes that have been encountered.

Traffic volumes (TV) are the operational entities to which regulation measures are applied.

A Traffic Volume is part of a Traffic Volume Set (TVS). ATFM delays generated by a regulation applied on a traffic volume are associated to a geographic area through a traffic volume set.

Statistical traffic volume set (STAT-TVS) has been created by the CFMU to be used for statistics. In opposition to TVS which are operational entities used at operational level, statistical entities are entities that do not change for a long period of time. Only the mapping with the underlying TVS changed.

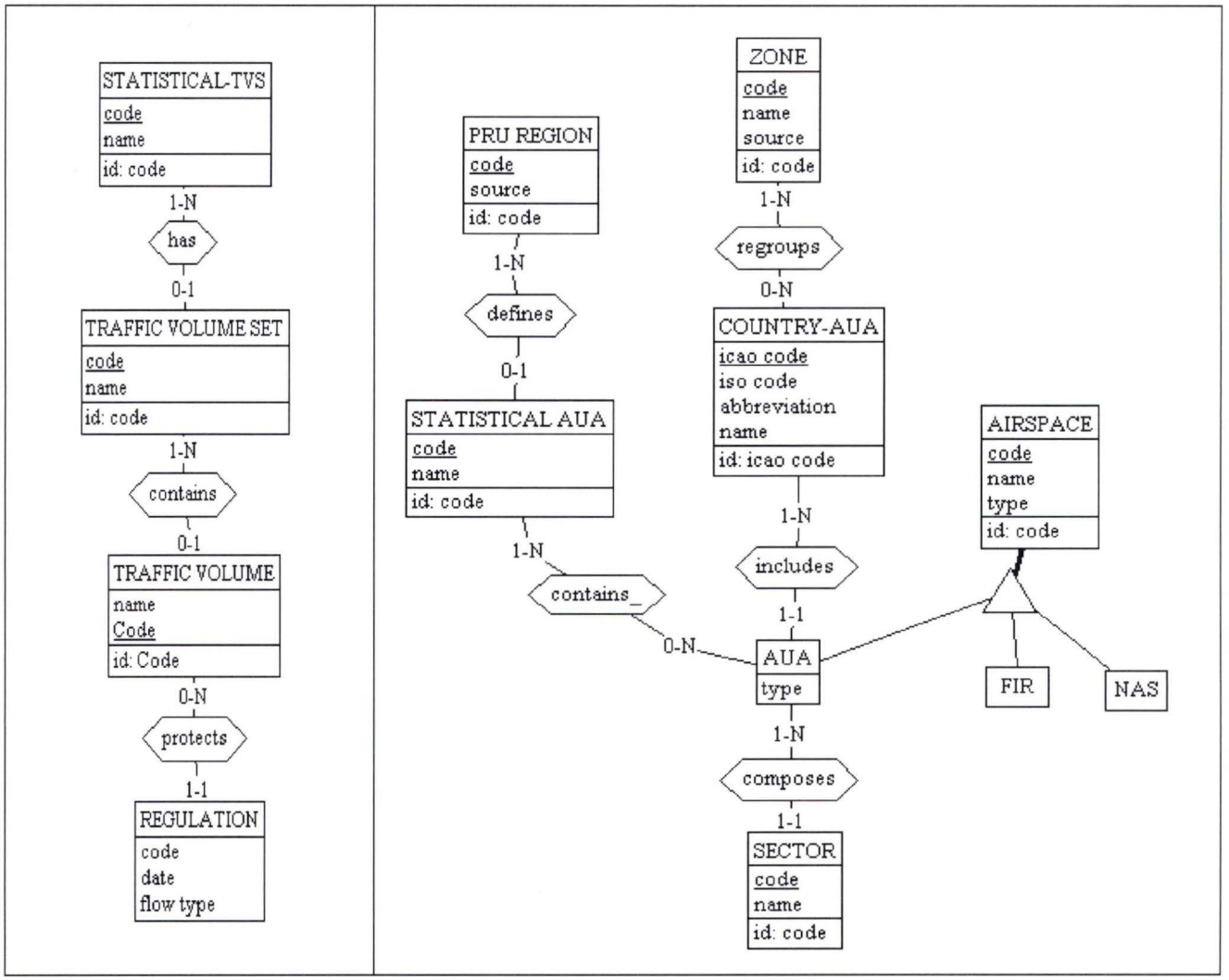


Figure 35: Hierarchies of entities

At each flight is associated an airspace profile which describes the path (represented in four dimensions) that an aircraft is expected to follow between the departure and the arrival airport in terms of airspace volumes.

Airspace volumes reflect the operational ATC airspace structure (AUA logic) and the administrative airspaces (FIR and NAS logic).

Sectors make up the airspace of an ATC Unit Airspace (AUA).

Several AUAs assembled together form one Statistical AUA (STAT-AUA). Statistical ATC Unit Airspace (STAT-AUA) has been created by the CFMU to be used for statistics. In opposition to AUA, statistical entities are entities that do not change for a long period of time.

PRU regions are composed of STAT-AUA and have been defined by the PRU for specific analysis.

All the AUAs of a same country form one COUNTRY-AUA airspace. The country is considered as the ICAO country. For example Canarias Island will be considered as a different country-AUA than Spain since its ICAO code is different from Spain's ICAO code.

A zone is composed of several COUNTRY-AUAs.

A Flight Information Region (FIR) is an airspace where Flight Information Services are provided and an NAS corresponds to National airspace.

Traffic Measures derived from Airspace profile by Stat AUA airspace can be associated with delay measures from regulations, via the relation between Stat AUA and Stat-TVS (see Figure 36).

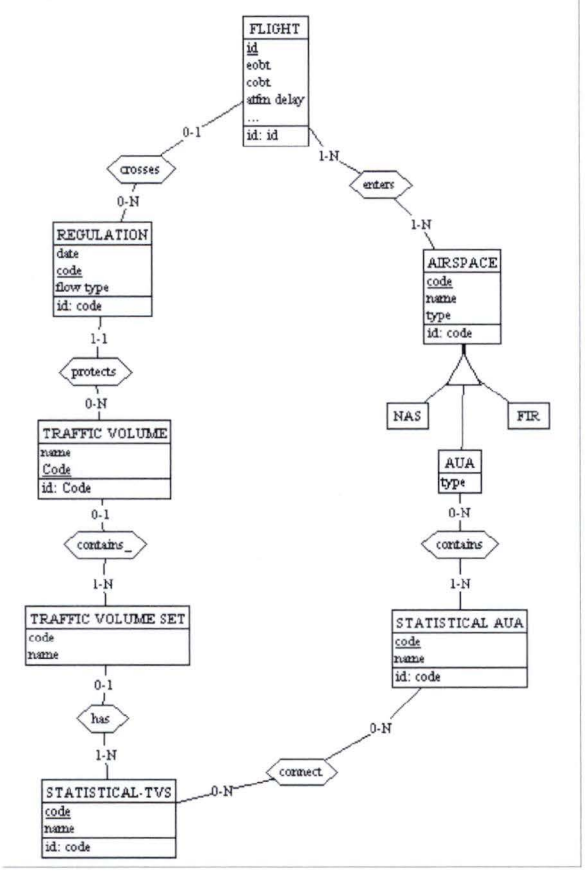


Figure 36: Relation between traffic and delay

Number of flights in an airspace.

The calculation is made taking into account the first entry time of the flight in the airspace, using the last available updated flight plan.

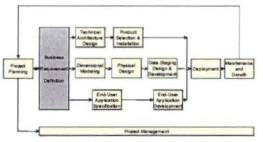
A flight entering an airspace will be counted only once in that airspace. It will be counted on the day of the first entry in the airspace. So, if a flight leaves the airspace and later re-enters it again, it will not be counted twice, even if it re-enters the airspace the next day (for flights crossing the 24:00 UT mark).

This definition has been used in order to be consistent with the definition used in the CFMU Synthesis database.

The number of flights within a given airspace for any given day is the sum of all flights entering for the first time in their journey the airspace on that day.

Duration of flights per airspace is calculated using (Exit Time – Entry Time) in the Airspace using the last available flight plan from the CFMU.

Only flights with the status "activated" are taken into account in the measures.



16 BUSINESS REQUIREMENT DEFINITION

User requirements have been compiled through analysing data in the CFMU Synthesis database, interviewing analysts and reviewing PRU reports.

For the project, my first task was to list all the data available in the CFMU Synthesis database. Then, using the annual reports published by the PRU which assess the air traffic management in Europe as well as specific reports (specific delay report in 1999), I updated the list accordingly.(These reports are available online at "http://www.eurocontrol.int/prc")

I then interviewed the 3 analysts who are in charge of the operational analysis using this list as the basis of discussion. New requirements were then formulated by the analysts, which correspond to current and future analysis they want to perform.

Some of these new requirements have been excluded (out of scope) from the project as the data were not readily available.

The final list of measures was then sent to all the analysts for approval. The list contained in Figure 37 defines the measures which are of interest for the analysts.

Measures	Dimensions	Available in CFMU Synthesis	Out of scope
Number Of Flights	Per aircraft Type		
	Per Aircraft Operator	x	
	Per Aircraft Operator Category		
	Per Zone (CFMU, ECAC, EURO 88, ...)	x	
	Per PRU Region		
	Per Country-AUA	x	
	Per ANSP		x
	Per STAT- AUA	x	
	Per AUA	x	
	Per departure airport	x	
	Per arrival airport		
	Per departure country	x	
	Per arrival country		
	Per FIR		
	Per NAS		
	Per Airport pairs		
	Per aircraft type and departure/arrival airport		
	Per aircraft type and departure/arrival country		
Nb of international flights	Per Country-AUA		
Nb of Domestic Flights	Per Country-AUA	x	
Number of Over Flights	Per Country-AUA	x	

Flight Duration	Per Zone (CFMU, ECAC, EURO 88, ...)		x	
	Per ANSP			x
	Per PRU Region		x	
	Per Country-AUA		x	
	Per STAT-AUA		x	
	Per AUA		x	
	Per FIR			
	Per NAS			
Km Flown	Per Zone (CFMU, ECAC, EURO 88, ...)			X
	Per ANSP			X
	Per Country-AUA			X
	Per STAT-AUA			X
	Per AUA			X
	Per FIR			
	Per NAS			x
ATFM Delay	Per Reference location		x	
Delayed flight	Per Sector		x	
Regulated flight	Per Airport		x	
ATFM Delay > 15 min	Per Regulation		x	
Delayed flight > 15 min	Per TV	>> Per Reason of delay	x	
	Per TVS	>>Per Type of Delay	x	
	Per STAT-TVS		x	
	Per STAT-AUA		x	
	Per PRU-Region		x	
	Per Country-AUA		x	
	Per Zone		x	
ATFM Delay	Per departure / arrival airport			
Delayed flight	Per airport pair			
Regulated flight				
ATFM Delay 1-15 min				
ATFM Delay 16-30 min				
ATFM Delay 31-60 min				
ATFM Delay + 60 min				
Delayed flight 1-15 min				
Delayed flight 16-30 min				
Delayed flight 31-60 min				
Delayed flight + 60 min				
Expected delay	Per Stat Aua			

Figure 37: Measures of interest

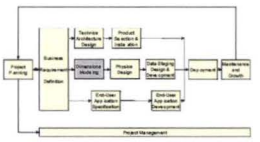
From the discussion with the analysts, it has emerged that there is a general need to

- Access very large amounts of data, e.g. several years of traffic.
- Analyse relationships between many types of business elements e.g. Aircraft, region, ...
- Compare aggregated data over hierarchical time periods

The following requirements have also been formulated:

- Use of Excel to perform analysis
- Easy to use tool to perform analysis and generate graphs and reports
- Availability of data: a delay of around 15 days between current day and data provided is acceptable to the PRU analysts.

17 DIMENSIONAL MODELLING



Three different perspectives of analysis (group of dimensions and measures) have been defined:

- the flight data mart
- the regulation data mart
- the airspace data mart

These data marts correspond to the requirements of the PRU analysts.

Data Mart Name	Data Mart Description
Flight	Describes group of flights along different characteristics.
Airspace	Describes group of flights by entered airspace.
Regulation	Describes the regulations and related ATFM delays.

The data mart matrix shows the relationship between the possible data marts and dimensions. Any dimension (column) with more than one X implies that this dimension must be conformed across multiple data marts. A brief description of each data mart and dimension follows the matrix.

Data mart / Dimension	Date	Aircraft Operator	Aircraft type	Airport	AUA	STAT-AUA	Country-AUA	PRU Region	ZONE	FIR	NAS	Regulation	Reference Location	Sector	TV	TVS	STAT-TVS	Type of delay	Reason of delay
Flight	x	x	x	x															
Airspace	x				x	x	x	x		x	x							x	
regulation	x			x								x	x	x	x	x	x	x	x

Dimension Name	Dimension Description
Date	Contains all of the attributes associated with the date that activity occurred.
Aircraft operator	Contains all of the attributes associated with the aircraft operator of the flight
Aircraft type	Contains all of the attributes associated with the type of aircraft used for the flight
Airport	Contains all of the attributes associated with the airport on which the flight has taken off or land
AUA	Describes the ATC Unit Airspace
STAT-AUA	Describes the airspace defined for statistical reason by the CFMU based on set of AUA.
Country-AUA	Describes the Country defined as a set of AUA by the CFMU
PRU Region	Describes the region defined as a set of STAT-AUA by the PRU
Zone	Describes the zone defined as a set of Country-AUA by the PRU.
FIR	Describes the Flight Information Region
NAS	Describes the National airspace
Regulation	Contains all the attributes associated to the regulation put in place
Reference location	Describes the reference location on which a regulation is put
Sector	Describes the sector associated with the reference location
TV	describes the Traffic Volume concerned by the regulation

TVS	describes the Traffic Volume Set defines by the CFMU as a set of Traffic Volume
STAT-TVS	Describes the Statistical Traffic Volume Set defined for statistical reason by the CFMU based on set of TVS.
Reason of delay	Contains all the attributes associated with the reason of delay
Type of delay	Describes if the regulation is put to protect airport or en-route reference location

Figure 38: Data marts matrix

17.1 Flight data mart

17.1.1 Purpose

- Flight data mart describes a group of flights and associated delay (measures), along the different characteristics of these flights (dimensions).
- The characteristics of the flights taken into account are the type of aircraft, the aircraft operator and the departure and arrival airports.

17.1.2 Granularity

The granularity is by aircraft type by aircraft operator by airport pair by day.

17.1.3 Overview of the concerned entities from which the Flight data mart is derived

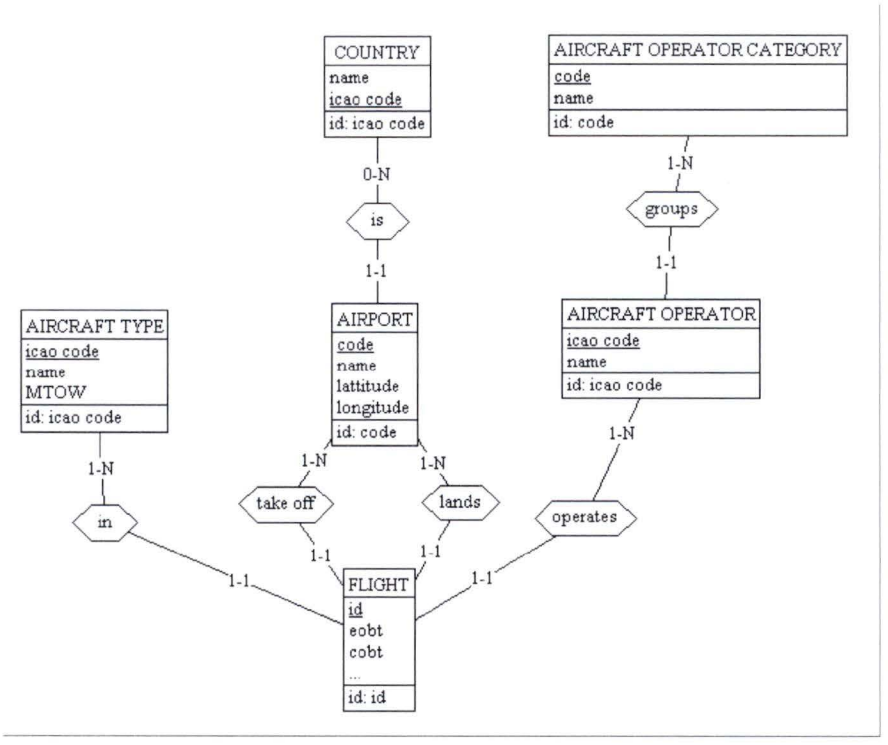


Figure 39: entities in Flight data mart

Figure 39 represents a point in time. For example, over time an aircraft operator can change category.

17.1.4 Dimensional model

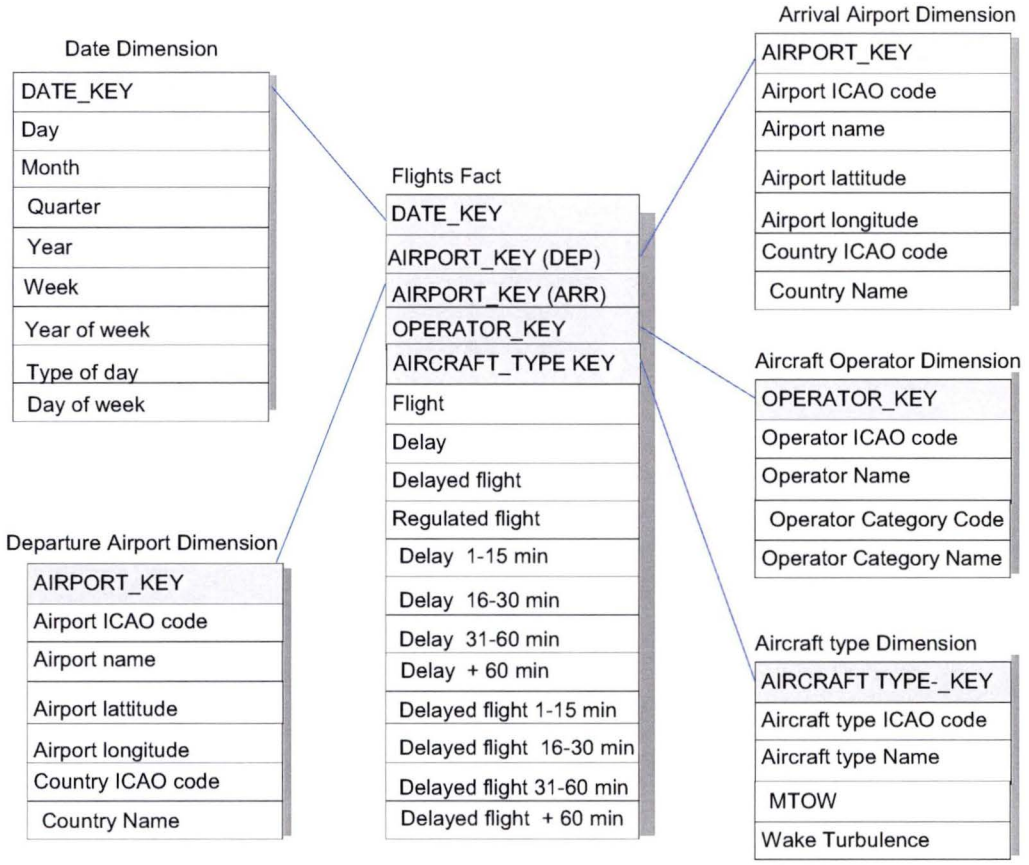


Figure 40: dimension model of flight data mart

17.1.5 Measure-Related

Measures	Description
Flight	Daily counts of flights where the time on which the flight leaves the gate for departure is on the selected date.
ATFM Delay	Daily counts of minutes of ATFM delay encounter by the flights. This ATFM delay is not specifically due to a regulation put to protect the departure or arrival airport but it is the delay related to the flights selected. These flights can be delayed by a regulation put to protect en route airspace.
Delayed Flight	Daily counts of flights which are delayed.
Regulated Flight	Daily counts of flights which are regulated.
Delay 1-15 min	Daily counts of minutes of ATFM delay for the flights where delay is between 1 and 15 minutes included.
Delay 16-30 min	Daily counts minutes of ATFM delay for the flights where delay is greater than 15 minutes and less or equal to 30 minutes.
Delay 31-60 min	Daily counts of ATFM delay for the flights where delay is greater than 30 minutes and less than or equal to 60 minutes.
Delay +60 min	Daily counts of ATFM delay for the flights where delay is greater than 60 minutes.
Delayed flight 1-15 min	Daily counts of flights where delay is between 1 and 15 minutes included.
Delayed flight 16-30 min	Daily counts of flights where delay is greater than 15 minutes and less or equal to 30 minutes.
Delayed flight 31-60 min	Daily counts of flights where delay is greater than 30 minutes and less than or equal to 60 minutes...
Delayed flight +60 min	Daily counts of flights where delay is greater than 60 minutes.

17.1.6 Dimension-Related

- Date
- Airport
- Aircraft Operator
- Aircraft type

17.1.7 Date dimension

The date corresponds to the date on which the aircraft leaves the departure gate.

As we can see from Figure 41 below, multiple hierarchies exist in the date dimension. These hierarchies allow to drill up and down.

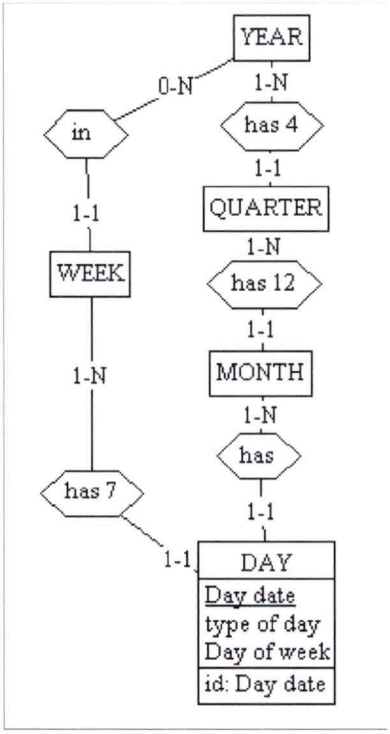


Figure 41: Date dimension

Attribute Name	Attribute Description	Sample Values
Day date	The specific day that an activity took place.	06/04/2005
Day of week	The specific name of the day abbreviated to 3 letters.	Mon; Sun
Type of Day	Indicates whether or not this day is a weekday or a weekend day.	W(Weekend), N(Weekday)
Week	The calendar week, week start on Monday.	1, 52
Year of week	The calendar year except for the first week where the year of the last day of the first week is taken and for the last week where the year of the first day of the last week is taken. This allows performing analysis on full week for a specific year.	2004
Month	The calendar month.	1, 12
Quarter	The calendar quarter	1, 4
Year	The calendar year.	2002

17.1.8 Airport dimension

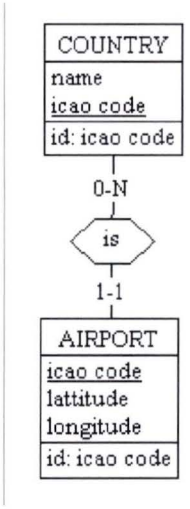


Figure 42: airport dimension

Attribute Name	Attribute Description	Sample Values
Airport ICAO code	4-letter ICAO code as listed in the ICAO Document 7910.	EBBR, LFPG
Airport ICAO name	Name of the Airport as listed in the ICAO Document 7910	Brussels,
Country ICAO code	ICAO code of the country where the airport is located	EB, LF,
Country ICAO name	Name of the Country where the airport is located	Belgium, France
Airport longitude	Longitude at which the airport is located	4.486111
Airport latitude	Latitude at which the airport is located	49.00972

17.1.9 Aircraft operator dimension

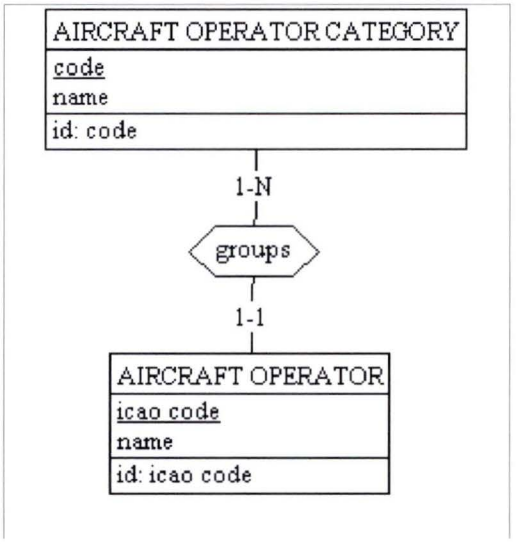


Figure 43: aircraft operator dimension

ttribute Name	Attribute Description	Sample Values
Aircraft Operator ICAO code	ICAO code of the aircraft operator	DAT, VEX
Aircraft operator name	Name of the aircraft Operator	Air France
Aircraft Operator category name	Category of aircraft operator	Military

17.1.10 Aircraft type dimension

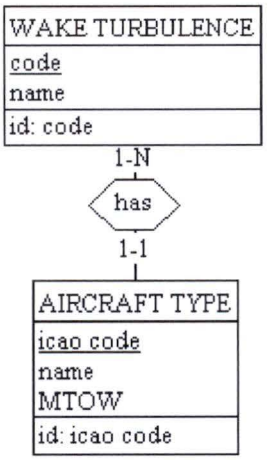


Figure 44: aircraft type dimension

Attribute Name	Attribute Description	Sample Values
Aircraft type ICAO code	ICAO code of the aircraft type	A332
Aircraft type name	Name of the aircraft Operator	AIRBUS A-330-200
MTOW	Maximum take off weight	5300 kg
Wake turbulence	Wake turbulence category	H (heavy)

17.2 Airspace data mart

17.2.1 Purpose

Airspace data mart describes the flow of flights by airspace entered and ATFM delays which are associated to these airspaces.

17.2.2 Granularity

The granular level in this data mart is the airspace entered per day.

17.2.3 Non-additive flight counts

A measure is semi-additive if it can be summed according to hierarchies in some dimensions but not in all of them. Flights are additive by time hierarchy but not by airspace hierarchy. For example, the number of flights in a STAT-AUA is not the sum of the number of flights in the AUA which composed the STAT-AUA.

Drill-up of number of flights though the airspace level lose their meaning, as a flight crossing the different airspace's would wrongly be attributed several times to the airspace entity that would hierarchically contain this group of airspace's. Therefore, to compensate this limitation, aggregate counts will be stored for different levels of airspace.

In contrary to duration which is an additive measure as it can be summed by time hierarchy and by airspace hierarchy. The duration in a STAT-AUA is the sum of the duration in the AUA which composed the STAT-AUA.

17.2.4 Overview of the concerned entities from which the Airspace data mart is derived

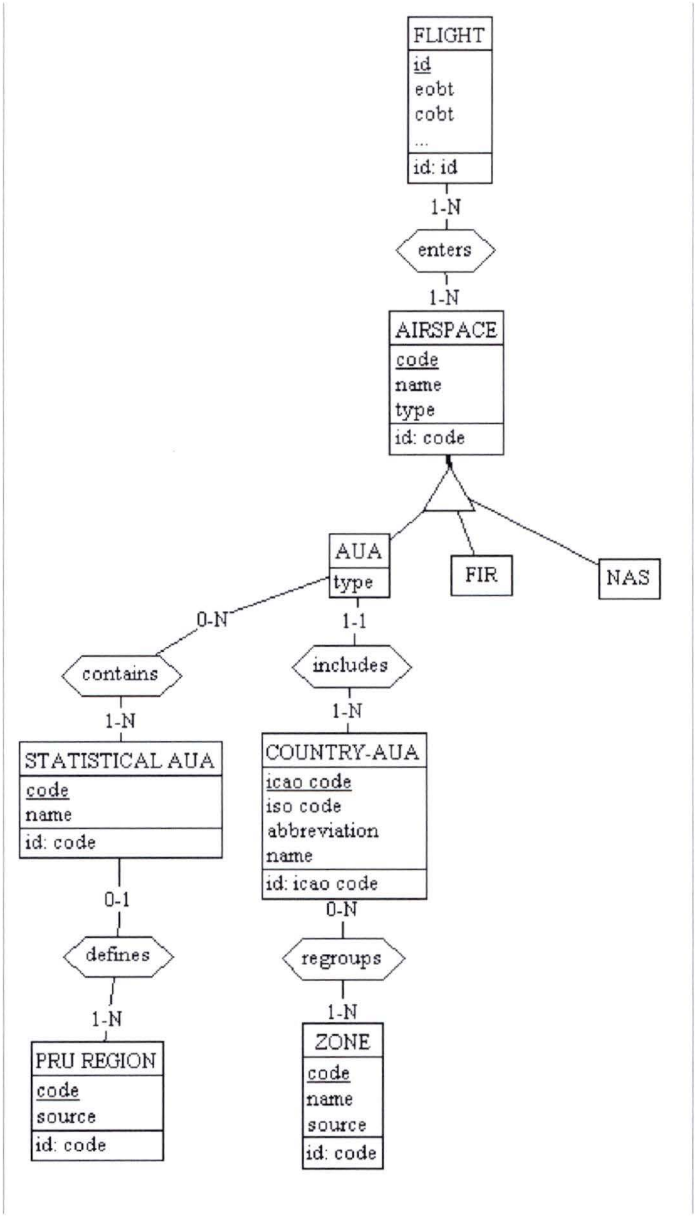


Figure 45: Entities in Airspace data mart

17.2.5 Dimensional model

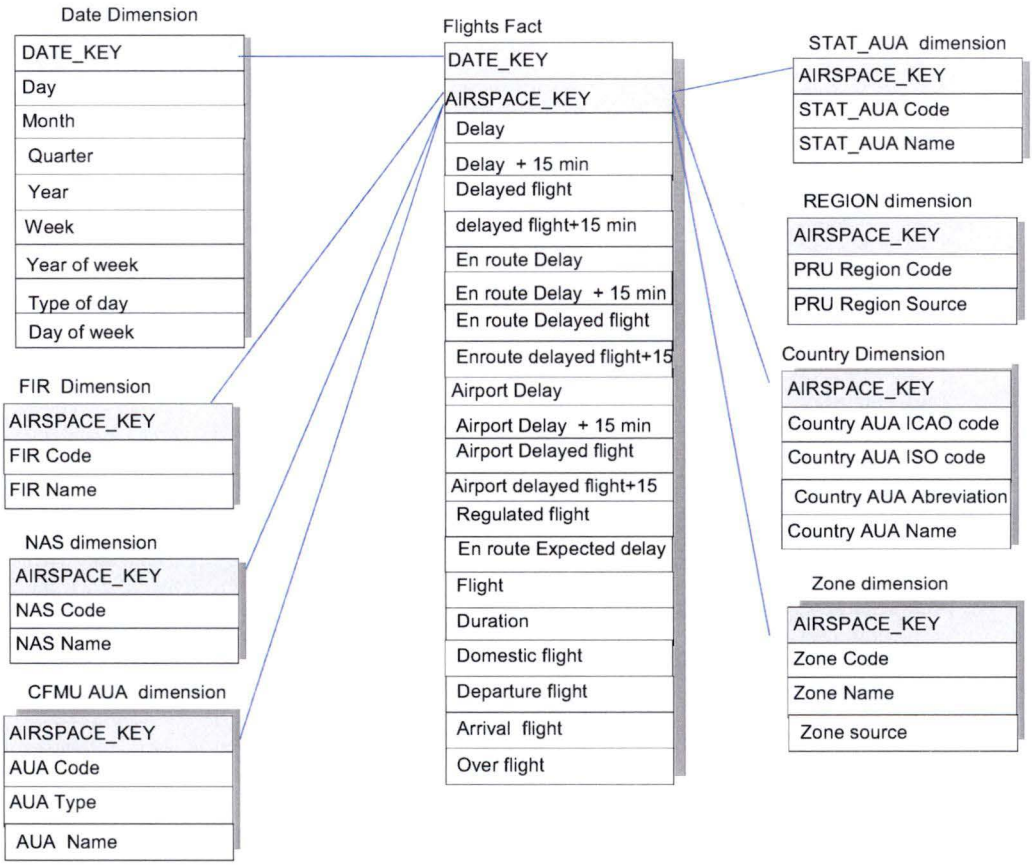


Figure 46: Dimension model of the airspace data mart

17.2.6 Measure-Related

Measures	Description
ATFM Delay	Daily counts of minutes of ATFM delay attributed to the airspace
En route ATFM delay	Daily counts of minutes of en route ATFM delay attributed to the airspace
Airport ATFM delay	Daily counts of minutes of airport ATFM delay attributed to the airspace
Regulated Flight	Daily counts of flights regulated due to the airspace.
ATFM Delay + 15 min	Daily counts of minutes of ATFM delay attributed to the airspace for flights delayed more than 15 minutes.
En route ATFM Delay + 15 min	Daily counts of minutes of en route ATFM delay attributed to the airspace for flights delayed more than 15 minutes.
Airport ATFM Delay + 15 min	Daily counts of minutes of airport ATFM delay attributed to the airspace for flights delayed more than 15 minutes.
Delayed flight	Daily counts of flights which have ATFM delay attributed to the airspace.
En route Delayed flight	Daily counts of flights which have en route ATFM delay attributed to the airspace.
Airport Delayed flight	Daily counts of flights which have airport ATFM delay attributed to the airspace.
Delayed flight + 15 min	Daily counts of flights which have ATFM delay greater than 15 minutes attributed to the airspace.
En route Delayed flight + 15 min	Daily counts of flights which have en route ATFM delay greater than 15 minutes attributed to the airspace.
Airport Delayed flight + 15 min	Daily counts of flights which have airport ATFM delay greater than 15 minutes attributed to the airspace.
Flight	Daily counts of flights entering in airspace (non additive measure).

Duration	Daily counts of minutes flown in the airspace.
En route expected delay	Daily counts of minutes of En route expected delay based on the current number of flights in the STAT-AUA.
Domestic flight	Daily counts of domestic flights in the Country AUA.
International departure flight	Daily counts of international departure flights in the Country AUA.
International arrival flight	Daily counts of international arrival flights in the Country AUA.
Over flight	Daily counts of over flights.

- Nb of Over flight, Domestic, International Departure and Arrival traffic: only at Country-AUA level
- En route expected delay: only at STAT-AUA level
- Delay
- En Route delay
- Airport Delay
- Delayed Flight
- Regulated flight

} In Stat AUA, Region, Country-AUA, Zone levels

17.2.7 Key Indicators

- Key indicators used by the PRU analysts which are derived from the measures.
- The Key indicators are not pre-computed in the data marts but are calculated on the fly.

Key Indicators	Formula
Delay per delayed flight	Delay / Delayed Flight
Delay per flight	Delay/ flight
% of delayed flight +15 min	Delayed flight +15 min / Flight
% of delayed flight	Delayed flight / flight
En route delay per flight	En route delay / flight
% airport delay	Airport delay / total delay
% en route delay	En route delay / total delay
% of regulated flight	Regulated flight / flight

17.2.8 Dimension-Related

- Date.
- Airspace
- AUA, STAT-AUA, Country-AUA, PRU Region; Zone, FIR, NAS

17.2.9 Date dimension

Date dimension corresponds to the first entry date in the airspace considered. See paragraph 17.1.7 for a detailed description.

17.2.10 AUA dimension

Attribute Name	Attribute Description	Sample Values
AUA code	Code of the ATC Unit Airspace	LEBLAPP, EDYYDUAC
AUA type	Type of the AUA	APP,.ACC.
AUA Name	Name of the AUA	Barcelona Approach,

17.2.11 STAT- AUA dimension

Attribute Name	Attribute Description	Sample Values
AUA code	Code of the STAT- AUA	EBBUACC, EDYYUAC
AUA Name	Name of the STAT-AUA	Brussels; Maastricht

17.2.12 Region dimension

Attribute Name	Attribute Description	Sample Values
Region code	Code of the Region	AREA SOUTH, BENELUX
Region source	Source of the definition of the region	PRU

17.2.13 FIR dimension

Attribute Name	Attribute Description	Sample Values
FIR code	Code of the FIR	EDFFFIR
FIR name	Name of the FIR	FRANKFURT FIR

17.2.14 NAS dimension

Attribute Name	Attribute Description	Sample Values
NAS code	Code of the National Airspace	LC
NAS name	Name of the National Airspace	Cyprus

17.2.15 Country- AUA dimension

Attribute Name	Attribute Description	Sample Values
Country- AUA ICAO code	ICAO Code of the Country	EB
Country- AUA ISO code	ISO Code of the Country	BE
Country- AUA Abbreviation Name	Abbreviation of the Country used in PRU report	Bosnia & Herz.
Country- AUA Name	Name of the Country	Belgium

17.2.16 Zone dimension

Attribute Name	Attribute Description	Sample Values
Zone code	Code of the Zone	ESRA2004
Zone name	Name of the Zone	ESRA 2004
Zone source	Source of the Zone	STATFOR

17.3 Regulation data mart

17.3.1 Purpose

The regulation data mart describes the regulations and associated ATFM delays.

17.3.2 Granularity

Granular level in this data mart is by regulation by day.

17.3.3 Overview of the concerned entities from which the Regulation data mart is derived

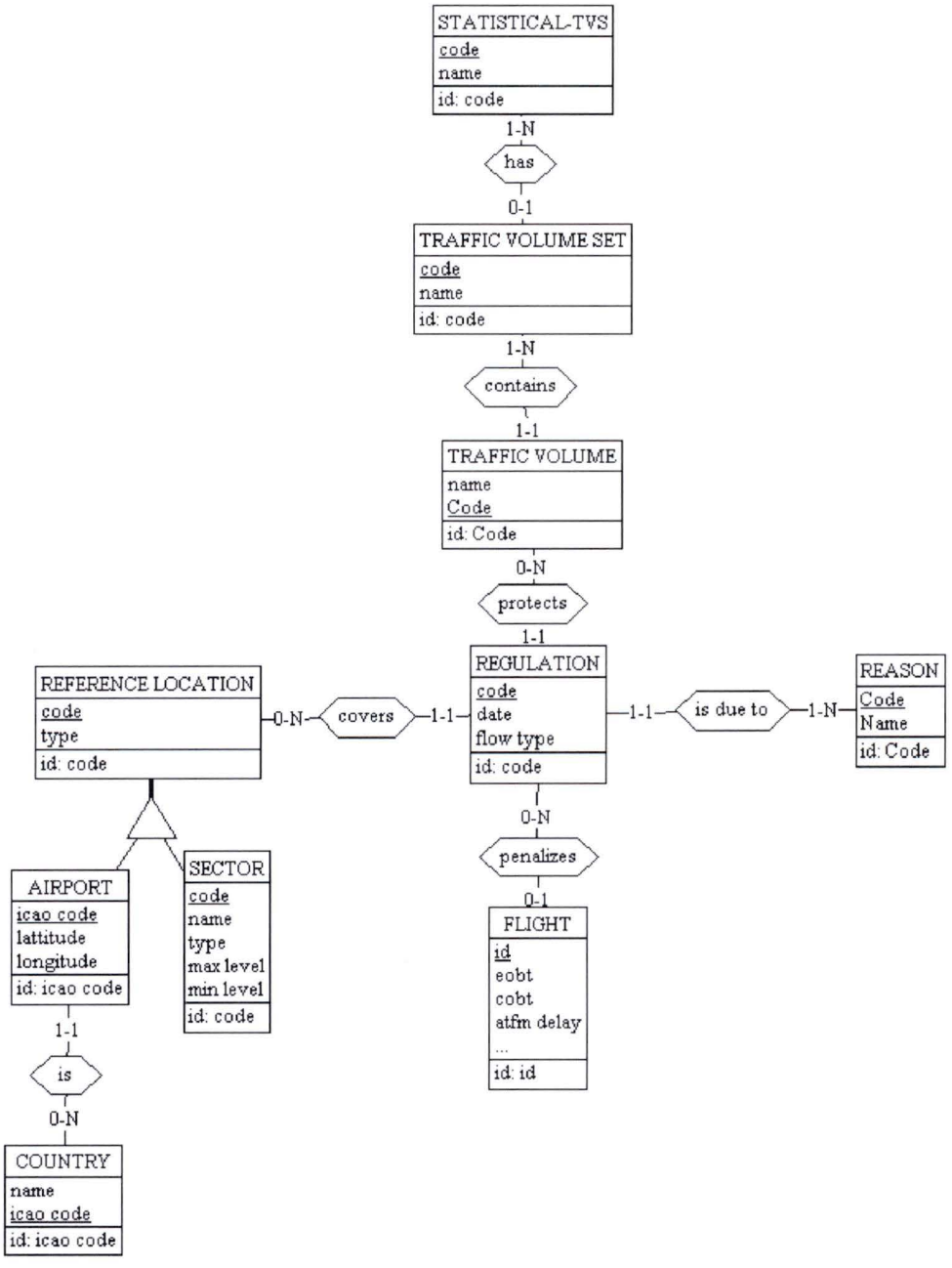


Figure 47: Entities in Regulation data mart

17.3.4 Dimensional model

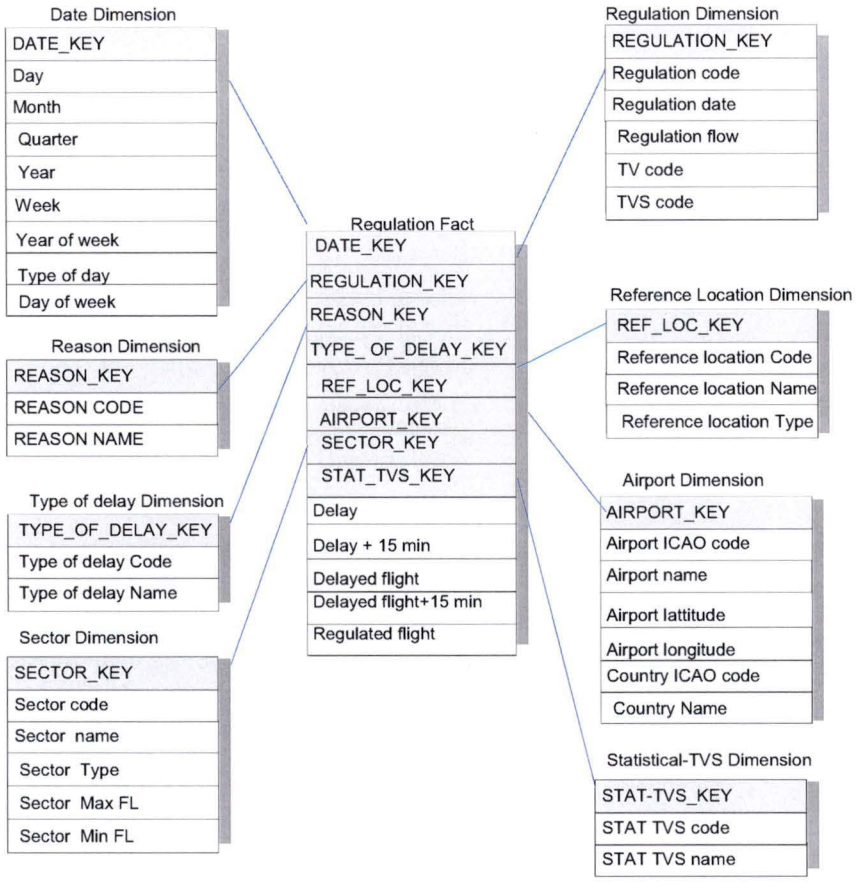


Figure 48: Dimensional model of regulation data mart

17.3.5 Measure-Related

Measures	Description
Delay	Daily counts of minutes of ATFM delay.
Delay + 15 min	Daily counts of minutes of ATFM delay where delay is greater than 15 minutes
Delayed Flight	Daily counts of delayed flights
Delayed Flight + 15 min	Daily counts of flights which have an ATFM delay greater than 15 minutes.
Regulated Flight	Daily counts of regulated flights

17.3.6 Key Indicators

Key indicators used by the experts which are derived from the measures.
The Key indicators are not pre-computed in the data marts but are calculated on the fly.

Key Indicators	Formula
Delay per delayed flight	Delay / Delayed Flight
% of delayed flight +15 min	Delayed flight +15 min / Delayed Flight

17.3.7 Dimension-related

- Date
- Regulation
- Reference Location:
- Airport
- Sector
- Statistical TVS

- Type of delay (En Route/Airport)
- Reason of Delay

Date dimension

The date corresponds to the date at which the aircraft leaves the departure gate. See paragraph 17.1.7 for a detailed description.

17.3.8 Regulation dimension

Attribute Name	Attribute Description	Sample Values
Regulation code	Regulation operational code	LKNEM01
Regulation date	Date of the first day of the regulation	10/03/2003
Regulation flow	Type of flow regulated (Arrival, Departure or global flow)	A, D, G
TV code	Traffic volume operational code	LKAANEM2
TVS code	Traffic volume set operational code	EKDKFMPW

17.3.9 Reference Location dimension

Attribute Name	Attribute Description	Sample Values
Reference Location Code	Reference Location operational Code	EBBUHES
Reference Location Type	Type of reference location, Airport(A) or Sector (S)	10/03/2003

17.3.10 Airport dimension

See paragraph 17.1.8 for detailed description

17.3.11 Sector dimension

Attribute Name	Attribute Description	Sample Values
Sector Code	Sector operational Code	EBMAWST
Sector Name	Name of the sector	MAASTR. UPPER WEST
Sector type	Type of sector (Elementary (ES) or Collapsed (CS)	ES, CS
Sector Max FL	Upper boundary of sector in Flight level	340
Sector Min FL	Lower boundary of sector in Flight level	245

17.3.12 Statistical TVS dimension

Attribute Name	Attribute Description	Sample Values
STAT TVS Code	Statistical TVS code	LIMMFMP
STAT TVS Name	Statistical TVS name	MILANO FMP

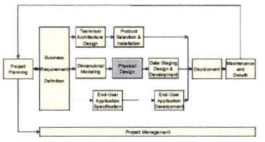
17.3.13 Type of delay dimension

Attribute Name	Attribute Description	Sample Values
Type of delay Code	Code of the type of reference location to protect (Airport (A), En route (E)	A, E
Type of delay Name	Type of delay name	Airport

17.3.14 Reason dimension

Attribute Name	Attribute Description	Sample Values
Reason Code	Reason of delay code	W
Reason of delay Name	Reason of delay name	Weather

18 PHYSICAL DESIGN



18.1 Surrogate keys

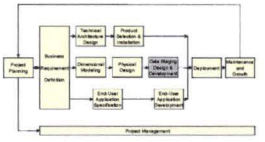
In the PRU data marts, surrogate (meaningless) key have been used as primary key for dimension tables except for the date key.

18.2 Materialised view

For the Flight data mart, the size of the fact tables is quite big as there are some 8-9 million flights per year in Europe. Since the grain used is quite low, the fact table contains approximately 10 millions rows since 1997.

To avoid duplicating this amount of data, the contents of a materialised view have been agreed between the different units concerned. This materialised view is managed by the EATMP SAMAD Data Warehouse team. The flight data mart is based on this materialized view.

19 DATA STAGING DESIGN & DEVELOPMENT



Data availability and quality

It was decided early in the process that, for consistency reasons with the CFMU, the same definitions and business rules used by the CFMU should be applied.

One of the first steps was to compare a sample of data computed by the PRU with the CFMU data. After investigation, we noted that not all information needed for the computation was transferred in the EATMP SAMAD Data Warehouse. A change request had to be made to get the required information.

We also noticed that, during the loading of the CFMU data in the EATMP SAMAD Data Warehouse, some assumptions were made by the EATMP SAMAD Data Warehouse team (e.g. delays less than 5 minutes are not taken into account when calculating the derived data: ATFM delay). Since no Meta data or clear documentation were available, it was sometimes quite time-consuming to find out the reason or the sources of differences.

The quality of the data was revealed to be sometimes problematic. For example, duration of flights were double counted in some airspaces for domestic flights. So specific algorithms had to be created in the Data warehouse to resolve this problem.

For all these reasons, the extraction process has been much longer than anticipated.

Data loading

The extraction, transformation and loading of the data marts have been coded manually using PL/SQL procedure.

Data before 2003 have been loaded from the CFMU Synthesis access database since not all data were available in the SAMAD data warehouse before that time. Data after 2003 are extracted from the SAMAD data warehouse (Oracle 9) and for some specific dimensions from external data sources.

Data sent by the CFMU are transformed, extracted and loaded in 3 different schemas by the EATMP Data warehouse Team.

- FLX schema contains data on individual flights and regulations.
- FSD schema contains the airspace profile of each flight per AUA, FIR and NAS.
- ENV schema contains environmental data such as airports, airspace, ...

FLX and FSD schema are appended daily while ENV schema is appended only every 28 days (AIRAC cycle). A snapshot of the ENV data is loaded every AIRAC cycle. The time key used is the AIRAC Cycle.

In the Staging area, the natural keys need to be transformed to surrogate keys in order to be used in the dimensional and fact tables.

Surrogate keys

Most of the PRU dimension tables are created based on ENV data. Every AIRAC Cycle, the dimension tables are updated following the method explained below.

Any new records from the operational source are inserted into the dimension table and are assigned the next surrogate key in sequence.

Existing dimension records that have changed are detected and the nature of their change examined. Depending on the policy chosen, the current dimension record is overwritten (type 1) or a new dimension record possessing the same natural key is created using the next surrogate key in sequence (type 2).

In loading the fact tables, the natural key must be replaced by the Surrogate key in the fact tables.

Every Airac cycle, "ToDate" field of records in dimension tables that cease to exist, or that have been replaced with a new record (type2), are updated (see section 11.4).

Relation Tables

Relation between entities can change in time. A 1 to N relation can become a M-N relation over time. For example an Aircraft operator can change Category over time (e.g. Regional airline to Cargo airline).

Relation tables are maintained in the staging area. They contain the surrogate key of the concerned entities and validity date of the relation (FromDate and ToDate). These relation tables are used for example to calculate the traffic in a STAT-AUA based on the relationship between AUA and STAT-AUA. Since only the AUA airspace entered by a flight is provided by the CFMU.

Figure 49 shows the loading of traffic data per airspace derived from AUA airspace.

Information at higher hierarchy level of airspace is computed by using Relation tables.

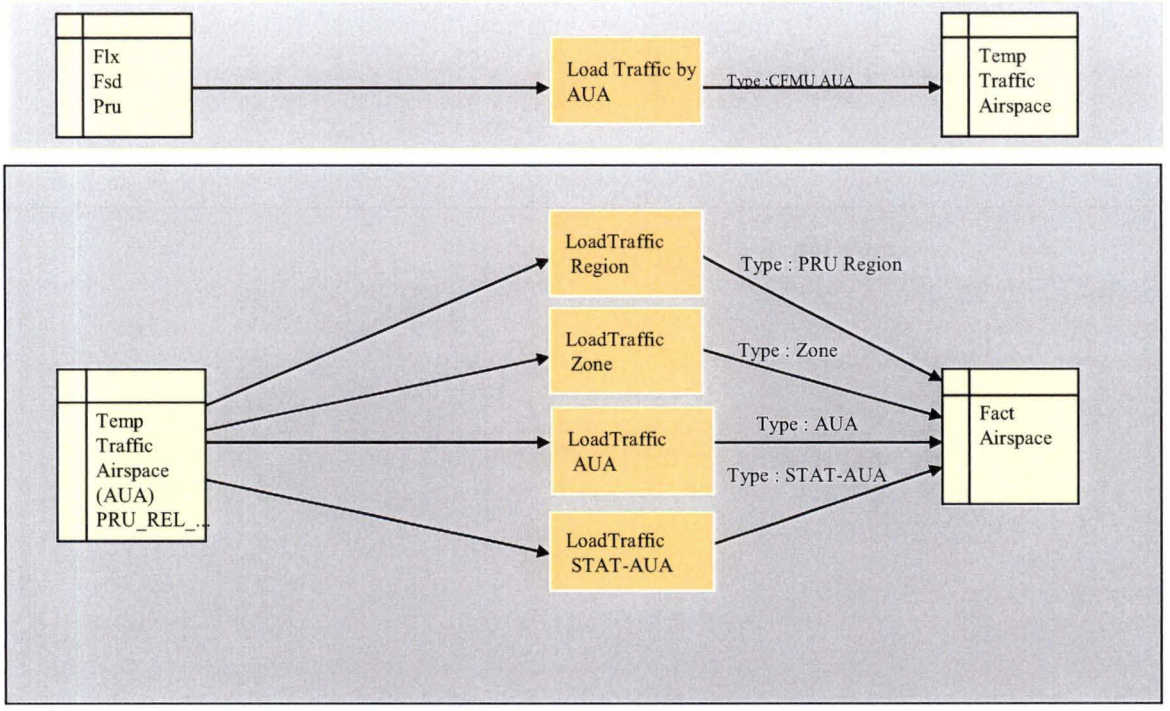


Figure 49: loading of traffic data per airspace

Manual update

Some dimensions are manually updated. These updates are not frequent and are usually performed once a year, for example when new zones are defined.

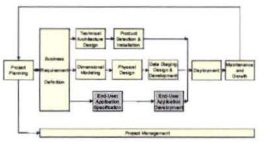
Date	Loading on demand by procedure	PRU
Zone	Updated Manually	PRU
PRU Region	Updated Manually	PRU
Aircraft Category	Updated Manually	PRU
STAT-AUA	Updated Manually	CFMU
STAT-TVS	Updated Manually	CFMU

Naming convention

There is no formal naming convention defined at SAMAD data warehouse level. Where possible, the same naming convention as in the source schemas has been followed:
E.g. TDM: total delay minutes, TTF: total traffic,....
Views are prefixed by V_.
Here are some of the naming conventions used in the PRU schema:

- The table and package prefixed by PRU_STG have been used to load data before 2003.
- The temporary tables are prefixed by PRU_TEMP.
- Fact tables are prefixed by PRU_FACT
- View is prefixed with V_PRU
- Surrogate key are named ID.

20 END USER APPLICATION



20.1 Selection of the tool

The end-user application chosen is an OLAP tool called Business Objects 5.0 (BO) which allows standard reporting and OLAP analysis.

Business Objects allows easy access for the user to the data, user can create their own queries through a graphical user interface by using business terms that are familiar to them without having to know SQL or the data structure behind the data.

The main reasons to use BO are:

- it is a standard tool in EUROCONTROL, so the support and the licence are free for the PRU.
- It is also the tool chosen by the CFMU. Thus, the PRU analysts can access the data marts of the PRU and the CFMU using only one tool.
- This tool can be interfaced with Excel via Business Query.
- It allows user to create their own queries

There is the possibility to create reports in a central repository, so that analysts just need to refresh these predefined reports in order to have up- to-date information. These reports can also be used by the analysts as a basis to create their own reports.

Predefined reports can also be saved in HTML. This enables quick access to reports through a web browser. (Web-I is the web version of BO but is not implemented as of now)

The initial assessment of BO shows that it is a suitable tool to use for the displaying of basic information but it is not an easy tool to use to perform complex analysis (calculating ratio…). The user interface is not intuitive and specific training is required.

From experience gained with BO tools, it can be seen that the PRU analysts generally use Business Object to find quickly some information and for drilling down. For example, to better understand if a monthly delay is due to a specific day or if it is spread over all the days of the month.

When the analysts need to perform specific calculations (running total...) they prefer to use Business Query (BO via Excel) via Pivot Table. First, because it is a tool with which they are familiar, and also because they find that performing calculations is not really straightforward in BO. This confirms our assessment of BO.

They also use Business Query when they have to produce graphs for publishing into reports, as they find that the BO's graphic features are too constraining.

The PRU analysts found that using BO allows them to perform their own queries without specific knowledge of SQL. This gives them more flexibility than having to request specific data extraction each time.

20.2 BO tool

BO suite includes the Designer tool which provides a graphical environment to map database structure to business terms, Business Objects, the end user tool and Business Query which allow to use Business Objects through Excel.

BO Designer

Business Objects Designer allows you to create Business Objects universes.

A universe is a file that contains the following:

- Connection parameters for one or more database middleware.
- SQL structures called objects that map to actual SQL structures in the database such as columns, tables, and database functions.
- A schema of the tables and joins used in the database. Objects are built from the database structure that is included in the schema. The schema is only available to Designer users. It is not visible to Business Objects users. Business Objects users connect to a universe, and run queries against a database. They can do data analysis and create reports using the objects in a universe, without seeing, or having to know anything about, the underlying data structures in the database.

The role of a universe is to provide an easy to use and understand interface for non technical BUSINESSOBJECTS users to run queries against a database to create reports and perform data analysis.

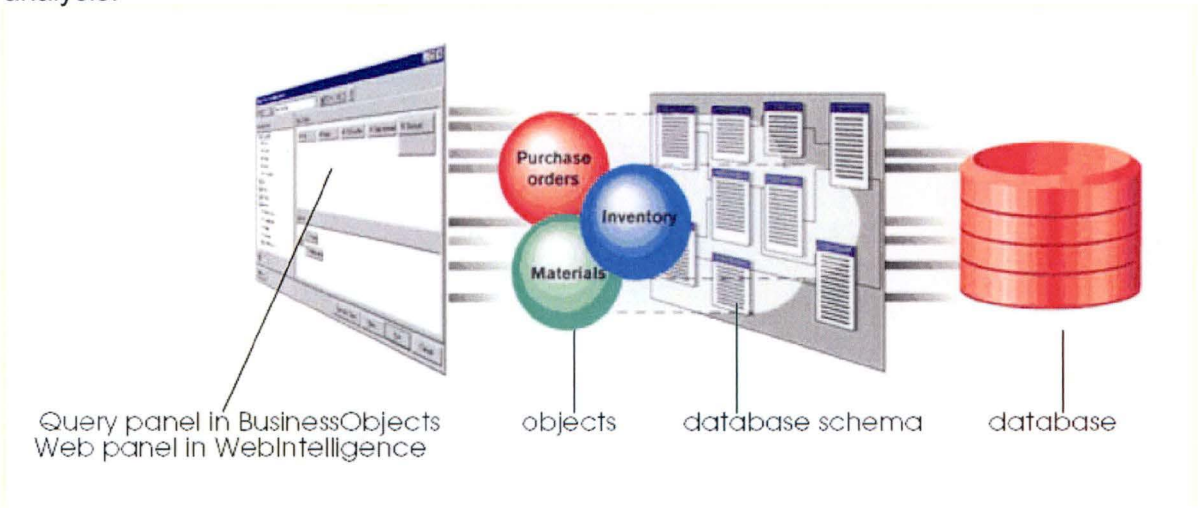


Figure 50: Business Objects tools

Figure 51 shows the design of the airspace data mart as seen in the Designer tool.

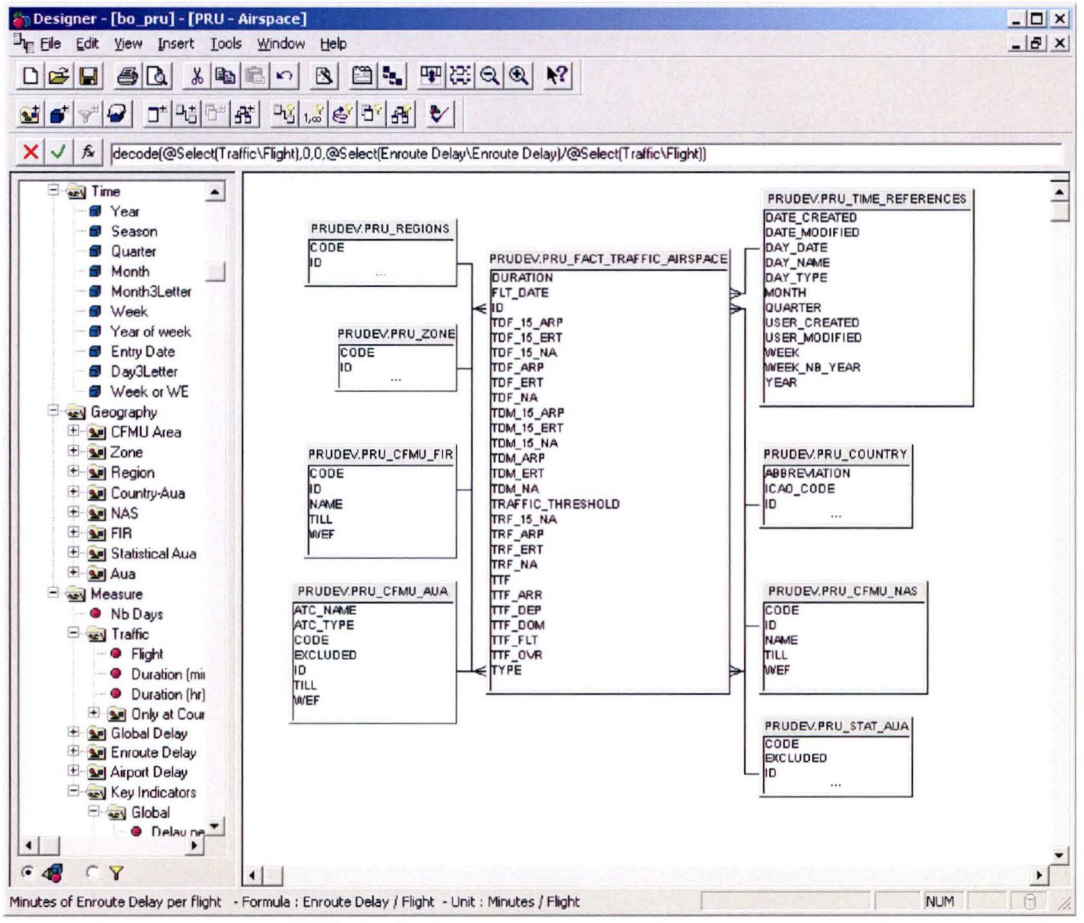


Figure 51: BO Designer tool

Business Objects end user tool

Business Objects is an integrated query, reporting and analysis tool. Using an editor called the Query Panel the user can create its query by selecting objects and applying conditions on these objects. Objects are elements that map to a set of data from a relational database in terms that pertain to the business situation. When the query is run, BUSINESSOBJECTS connects to the database and retrieves the data mapped to the objects selected.

Dimension object

An object can be qualified as a dimension, a detail, or a measure. Each type of object serves a different purpose:

• Dimension objects retrieve the data that will provide the basis for analysis in a report. Dimension objects typically retrieve character-type data (airport code,..), or dates (year, month,..)

Detail object

• A detail object is always associated to one dimension object, on which it provides additional information. For example, airport name is a detail object that is associated to Airport Code

Measure object

• Measure objects retrieve numeric data that is the result of calculations on data in the database.

Figure 52 shows the airspace data mart as seen by the end-user.

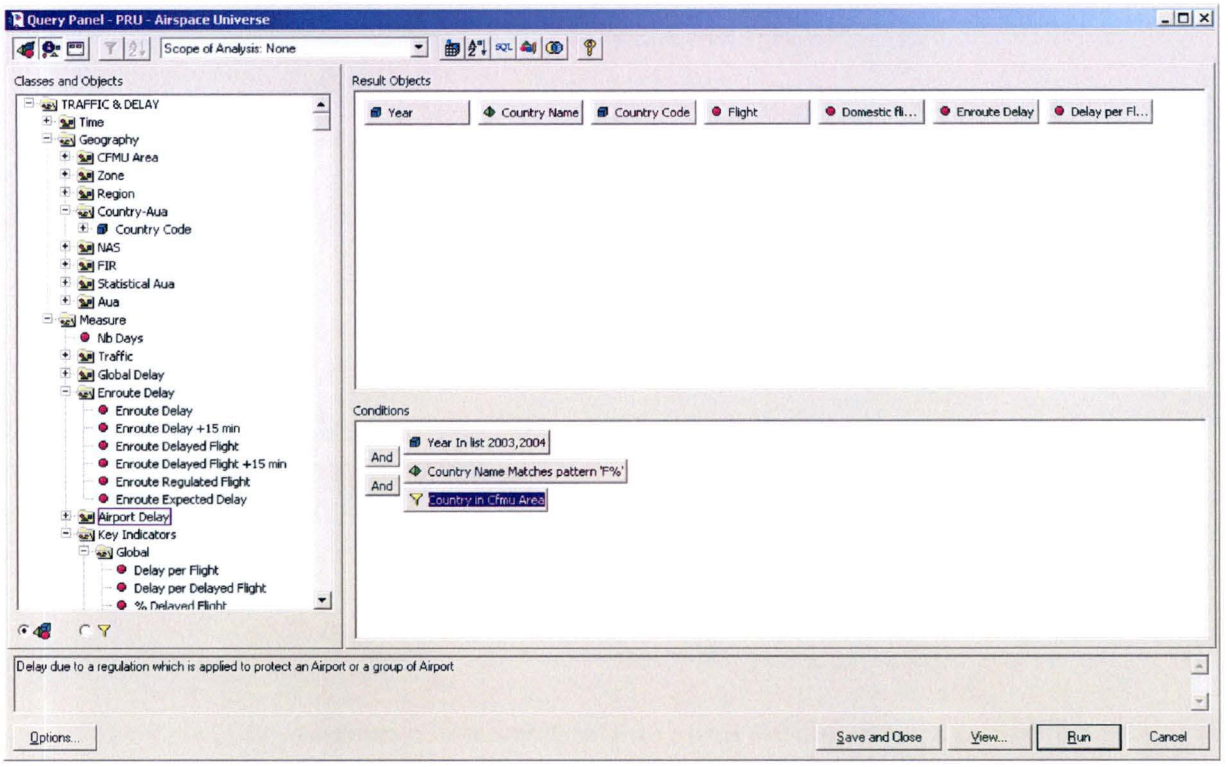


Figure 52: BO query interface of the end user

The result of the query performed in Figure 52 can be seen in Figure 53.

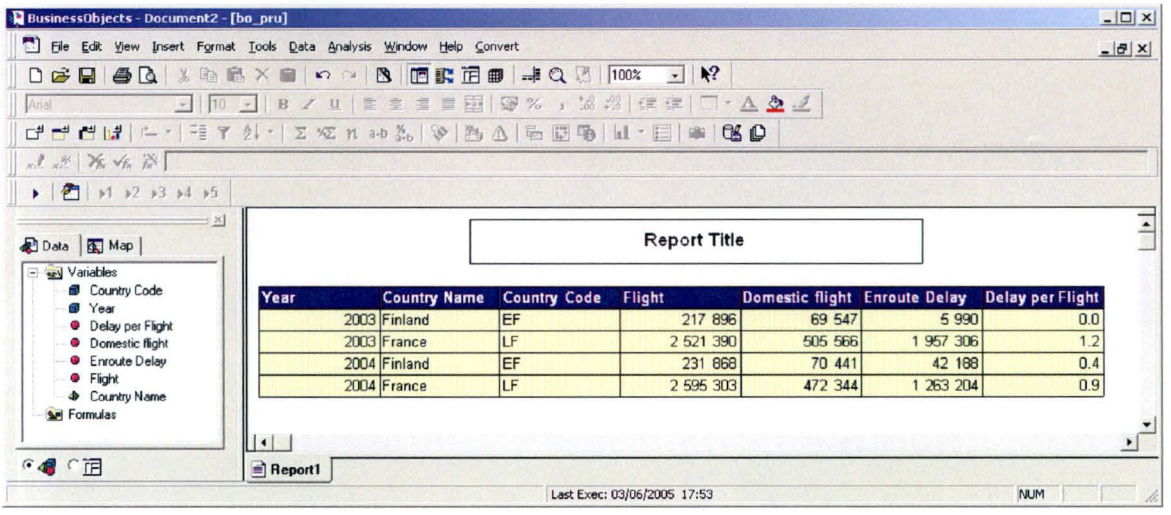


Figure 53: BO report

Figure 54 shows a BO report in drill down mode, clicking on the "month" button will allow the user to drill down one level in the Time hierarchy.

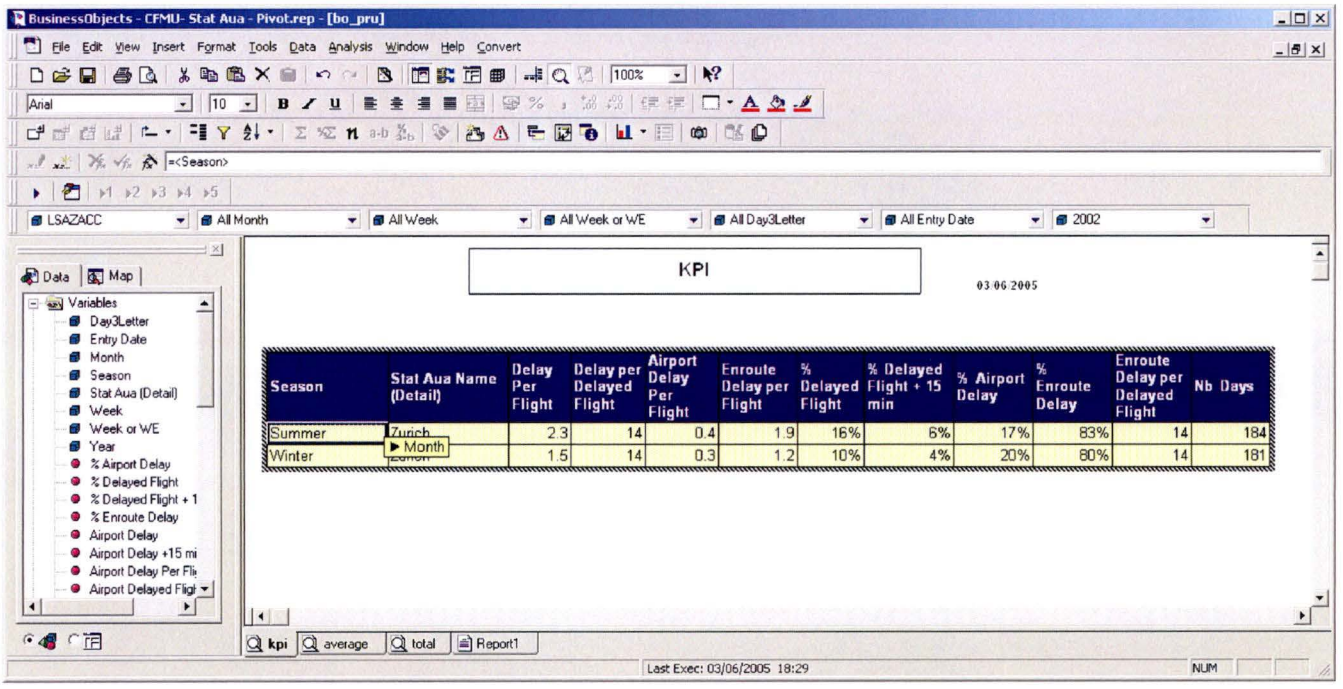


Figure 54: Example of BO report with drill down

Figure 55 shows a BO report as it appears when it is saved to be exported for the Web. The html report is static but it allows selecting for example a specific STAT-AUA or the time level (weekly or monthly graph)

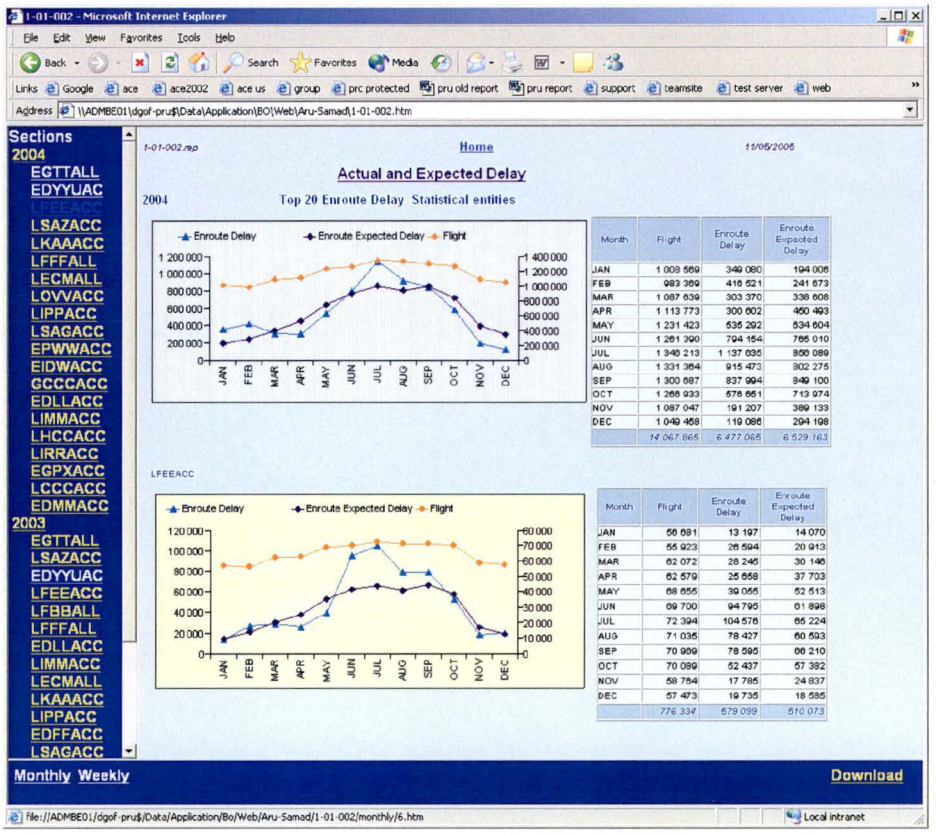


Figure 55: Example of BO report exported for the web

20.3 User interface development

The user interface has been developed by me, in close cooperation with the PRU's operational analyst. The interface uses terminology familiar to the analysts. The description of the dimensions and measures has also been agreed with the operational expert.

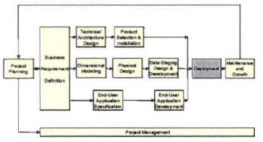
There is no Meta data that can be queried by the end user (for example through a web browser) that give information on the data selected (business rule, data source, quality of data,...). However, a short description of the dimensions and measures is available though BO.

This information is sufficient for the PRU analysts as they have a good knowledge of the domain but Meta data is certainly important for other users. (No Meta data tool is available for the moment from the EATMP team).

I have defined a first set of BO reports which mainly correspond to graphs which are presented each year in the annual reports of the PRU and which are accessible to the users from the BO repository.

At the request of the analysts, a few Bo reports have been designed to be exported in HTML. These reports allow the analysts to see the "current" situation (web reports are updated usually on a monthly base).

21 DEPLOYMENT



BO has been installed on the PC of five PRU analysts and access has been provided to the universes of the PRU and the CFMU.

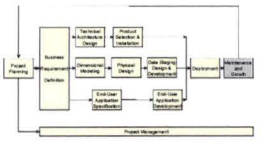
All the members of the PRU have access to the Web reports.

An introductory presentation has been given by the PRU team. It included:

- An introduction to the data warehousing and background project.
- A presentation of the CFMU and PRU universes: dimensions and facts
- A presentation of the predefined reports and web reports
- How to start using the data marts and access the predefined reports
- How to create their own report, graphs and navigate in the data (drill down,...)
- How to used BO with excel.

Two days hands-on training is also provided by EUROCONTROL.

22 MAINTENANCE AND GROWTH



The procedures for daily and monthly loading have been handled to the EATMP SAMAD Data Warehouse team in order to integrate it in their automatic scheduling. Some checks are performed in order to detect loading problems that may have occurred.

The manual update of the dimension is the responsibility of the PRU. A quality check is performed each month by the PRU by comparing data with the CFMU data marts.

Each time a difference is noticed with the CFMU data, an investigation is performed by the PRU to see the reason of this discrepancy. Some times it happens that data are not loaded correctly into the data warehouse. Following these experiences, extra checks have been put in place by the EATMP DW at the request of the PRU. During the PRU checks, quality of the data is also sometimes problematic (e.g. negative delay). As there are no common rules concerning these specific cases, the result can differ from the CFMU.

Every month, a procedure is run by the PRU to update the BO web reports

The update of the data mart tables and user interface following new requirements from the experts are the responsibility of the PRU.

If new data or indicators have to be added to the data warehouse on the request of the PRU, the PRU will sponsor the EATMP Data Warehouse Team to perform these new developments.

Technical support is also provided to PRU team by the EATMP Data Warehouse Team.

23 CO-ORDINATION WITH OTHER UNITS

Since the aggregation tables created by the PRU are of interest to other teams, discussions are ongoing to create similar aggregation at data warehouse level.

There is of course a need to coordinate with other units in order to agree on common measures and dimension (e.g. there are different ways to calculate the number of flights crossing an airspace).

For the moment, the other units of EATMP do not use Business Objects, so access to the PRU aggregation is provided either through Excel pivot tables which are refreshed by the PRU team, or direct access is given to the dimension and fact tables for more technical users.

Since no “easy to query” Meta data are available, it is often difficult for the analysts of the other teams to understand the exact definition and source of the information provided.

There are different initiatives being taken by the units of EATMP to create Data marts. There are also more and more data sources available, such as radar data, aircraft performance data etc (see Figure 56).

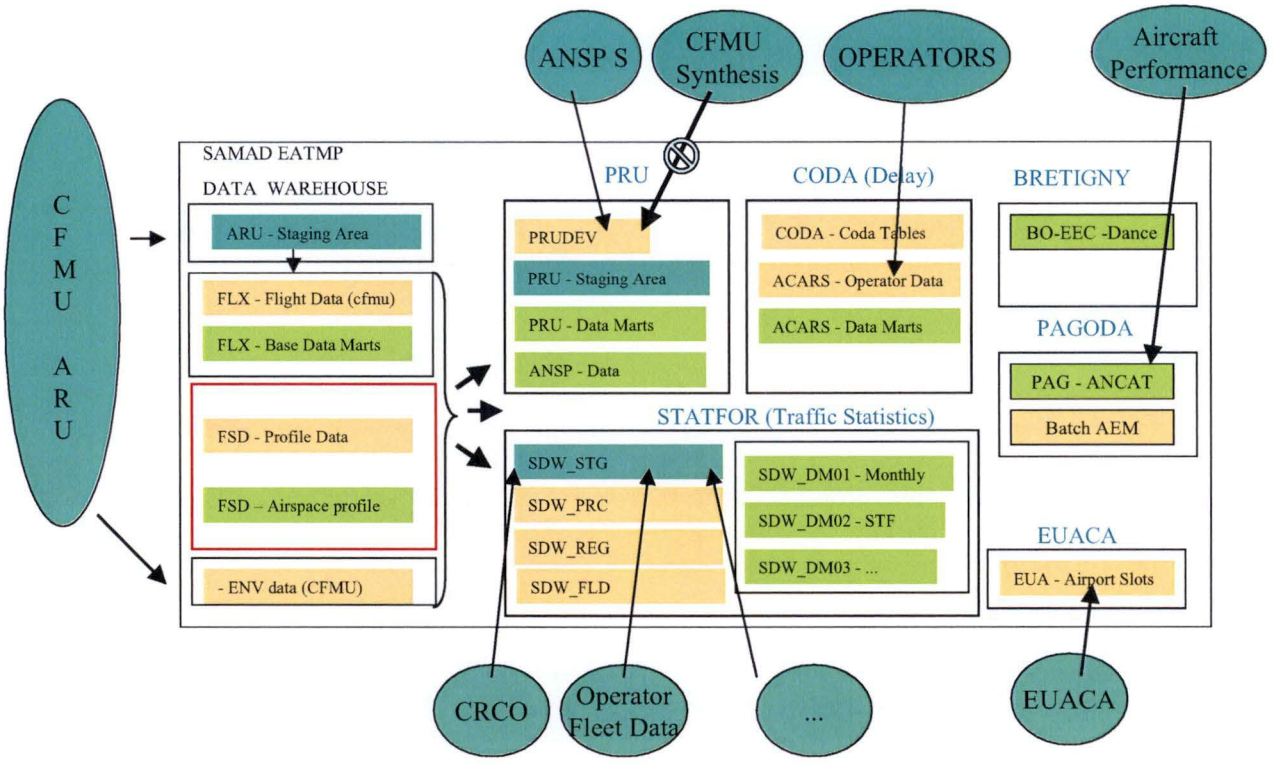


Figure 56: EATMP SAMAD Data Warehouse

The EATMP Data warehouse team organises meetings with all the units concerned in order to coordinate their work. Since no common rules have been defined (dimension, calculation methods...), it leads to a situation which is comparable to "independent data mart architecture" even if there is a common Data warehouse.

This coordination is necessary to avoid reverting to the previous situation of duplication of effort, and also the publication of different statistics by different units.

It is abundantly clear that in order to be able to use data managed by other units, there is a need to define common dimensions and also common or agreed calculation methods.

There is also a need to nominate data proponents (also called "stewards") for the maintenance of these dimensions. They should be the focal point for questions related to this information.

Above all, the most important item needed is to have a common Meta data tool (some tools are currently under investigation).

24 CONCLUSION

Section I of this paper described the main concepts of Data Warehousing. Section II described the PRU Data mart project.

From reviewing the literature on data warehousing, it can be concluded that there are no clear definitions or agreed methodologies concerning Data Warehousing.

There are arguments for and against the use of the Dimensional modelling or Relational modelling. The PRU found that Dimensional modelling was well-adapted to the project, which deals with well-defined queries on aggregated data. However, Dimensional modelling seems to be too rigid and too query-oriented to be used for the design of a central data warehouse.

The first task of the Data mart project was to define the project's data requirements. This was not a laborious task, as can often be the case in projects of this type, as there was a clear idea of the kind of analysis to be performed.

The extraction and transformation of data was more problematic. There was no clear documentation available which describes the contents, business rules, ETL or agreed definitions of the terminology. This resulted in time lost. From this experience it can readily be concluded that there is a compelling need to have clear documentation or, even better, a common Meta data repository. Such a repository should also contain information on the quality of the data available in order to avoid mis-interpretations.

Insofar as the outcome of the project is concerned, it should be noted that Business Objects does not correspond completely to the end-users needs. However, it has the major advantage that analysts can create their own queries through a graphical user interface by using business terms that are familiar to them without having to know SQL or the data structure behind the data. This also gives the analysts more freedom, as until then, they had to ask the PRU IT team to perform specific queries for them.

ANNEX I - GLOSSARY

TECHNICAL GLOSSARY[13]

Aggregation
Information stored in a data warehouse/ data mart in a summarized form

Data warehouse
A database where data is collected for the purpose of being analyzed. A data warehouse collects, organizes, and makes data available for the purpose of analysis - to give management the ability to access and analyze information about its business. This type of data can be called "informational data".

Data Warehousing
The process of visioning, planning, building, using, managing, maintaining, and enhancing data warehouses and/or data marts.

Data Mart
A database that has the same characteristics as a data warehouse, but is usually smaller and is focused on the data for one division or one workgroup within an enterprise

Dimensions
From a statistical point of view, dimensions describe the different possible states of an event (modalities or properties): e.g., departure airport of a flight. In SQL terminology, dimensions correspond to fields following "group by" clause.

Drill down/drill up
The ability to move between levels of the hierarchy when viewing data with an OLAP browser.

Decision Support System (DSS)
A computer system designed to assist an organization in making decisions.

Hierarchies
Ordered set of dimensions, logically put in a hierarchy: e.g. years, month, day are linked hierarchically from global to more detailed, as year dimension include all possible values of month dimension, including it self all possible values of day dimension.

Granularity
The level of detail of the facts stored in a data warehouse

Measures
From a statistical point of view, measures correspond to quantitative variables (continuous or discrete variables) describing the intensity of an event: e.g., number of flights, delays. They correspond to indicators measuring a given phenomenon or event.

Meta data
Data that describes the data in the data warehouse/mart.

MOLAP
OLAP that stores data and aggregations in a multidimensional database structures.
(Multidimensional Online Analytical Programming)

[13] Some definitions are extracted from " http://www.sdgcomputing.com/glossary.htm"

OLAP
"OLAP" is the most widely used term for multidimensional analysis software. The term "On-Line Analytical Processing" was developed to distinguish data warehousing activities from "On-Line Transaction Processing" - the use of computers to run the on-going operation of a business.
In its broadest usage the term "OLAP" is used as a synonym of "data warehousing". In a more narrow usage, the term OLAP is used to refer to the tools used for Multidimensional Analysis.

OLTP
The use of computers to run the on-going operation of a business (OnLine Transaction Processing).

ROLAP
OLAP that stores data and aggregations in a relational database. (Relational On-Line Analytical Processing)

Slice and Dice
The ability to move between different combinations of dimensions when viewing data with an OLAP browser.

Snowflake schema
It is a star schema on which normalization is applied to the dimension tables.

Star schema
Data model which is designed to provide data retrieval power, where a central fact table (detailed or aggregate table) is surrounded by and joined to multiple dimensions. Visually, model looks like a star. Star Schema. Technically, it is a a database design that consists of a fact table and one or more dimension tables. Each of the dimension tables has a single field primary key which has a one-to-many relationship with a foreign key in the fact table. The star schema is an intentional simplification of the database design that would be achieved by following the standard rules of normalization. The dimension tables are often flattened, to allow for more efficient querying

AVIATION RELATED GLOSSARY

Aircraft Operator
An aircraft operator is a commercial organisation or enterprise that engages in (or offers to engage in) aircraft operations, making use of the air traffic system to transport passengers and goods from one location to another.

Aircraft type
Aircraft type describes the operating characteristics of the most commonly used civil aircraft.

Airport
An airport is a defined area on land to be used for the arrival, departure and surface movement of aircraft.

ATFM delay
"ATFM delay" is defined as the duration between the last Take-Off time requested by the aircraft operator and the Take-Off slot given by the Central Flow Management Unit (CFMU).

ATC
An Air Traffic Control provides air traffic control services to controlled flights within its areas of jurisdiction

Airspace volume
An airspace volume is a generic term referring to various types of airspace volumes used in air navigation.

Airspace volume profile
The airspace volume profile describes the path (represented in four dimensions) that an aircraft is expected to follow between the departure and the arrival airport in terms of airspace volumes

Reference location
Reference locations are used for ATFM activities. They are the base reference for a traffic volume. Regulations applied to the traffic volume will use the defined capacity of the reference location as the basis for slot allocation.

Traffic volume profile
The traffic volume profile describes the path (represented in four dimensions) that an aircraft is expected to follow between the departure and the arrival airport in terms of traffic volumes that have been encountered.

Traffic Volume
Traffic volumes are the operational entities to which tactical ATFM measures are applied.

Regulation
A regulation describes an ATFM measure taken to try and relieve pressure on the air traffic system when a capacity threshold is reached. A regulation affects the time of departure of flights entering a defined traffic volume. A flight is submitted to the regulation when it follows the traffic volume and the estimated time of over flight/entry is between the regulation start and end time.

ANNEX II - BIBLIOGRAPHY

[IMHOFF,2003], Claudia Imhoff, Nicholas Galemmo, Jonathan G. Geiger, *Mastering data Warehouse Design : Relational and Dimensional Techniques* , John Wiley & Sons, USA, 2003

[INMON,1998], W. H. Inmon, *Building the data Warehouse,* John Wiley & Sons; USA, 2002, 3rd edition, 1998

[KIMBALL,1998], Ralph Kimball,Laura Reeves,Margy Ross, Warren Thornthwaite; *The data Warehouse Lifecycle Toolkit : Expert Methods for Designing, Developing, and Deploying data Warehouses;* John Wiley & Sons; USA, 1998

[BERKELEY,1997] UC BERKELEY, GROUP D I.S. 2006, An introduction to the data Warehouse, http://www.sims.berkeley.edu/courses/is206/f97/GroupD/ datawarehouse.html, 1997, (Date of access 10/04/2005)

[BERSON & SMITH,1997], Alex Berson and Stephen J. Smith, Components of a data warehouse, http://www.tdan.com/i003fe11.htm , data Administration Newsletter, TDAN.com Issue 3.0, December 1997 (Date of access 10/04/2005)

[CODD,1993], E.F. Codd & Associates, 'Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT", http://dev.hyperion.com/resource_library/white_papers/providing_olap_to_user_analysts_0.cfm, 1993, (Date of access 16/04/2005)

[DEMAREST,1995],Marc Demarest, A data Warehouse Evaluation Model, http://www.noumenal.com/marc/oracle7.html, April 1995, (Date of access 10/03/2005)

[DEMAREST,2001], Marc Demarest, two-tiered decision support systems architecture, http://www.noumenal.com/marc/waremart.pdf, November 2001, revision 6.2, (Date of access 10/03/2005)

[FIRESTONE,1998], Joseph M. Firestone,.Architectural Evolution in dataWarehousing and Distributed Knowledge Management Architecture, http://www.taborcommunications.com/, White Paper No. Eleven, July 1, 1998,

[FUNDP,2002]: laboratoire d'ingénierie des applications de base de données, TimeStamp project, understanding, developping, processing Temporal databases, Université de Namur (FUNDP), January 2002.

[INMON,1995], W.H. Inmon, What is a data Warehouse?", Prism, Volume 1, Number 1, 1995

[GUPTA, 2000], Vivek R. Gupta, "An introduction to data Warehousing", http:// datawarehouse.ittoolbox.com, May 17, 2000 , (Date of access 16/03/2005)

[HACKNEY,1998] Douglas Hackney, Warehouse Delivery: Who Are You? Part I, DM Review Magazine, February 1998 issue

[HACKNEY,2000] Douglas Hackney, Architectures and Approaches for Successful data Warehouses,, http:// datawarehouse.ittoolbox.com, FEB 2000 , (Date of access 14/04/2005)

[HAINAUT, 2001], Jean Luc Hainaut, Ingénerie des Bases de données, Volume1, 3ème édition, LIHD 2001-2002, Faculté Universitaire ND de la Paix, Institut d'Informatique, FUNDP, Septembre 2001

[LAMBERT,1996], Bob Lambert, data Warehousing Fundamentals: What You Need To Know To Succeed:

Special Feature, http://www.dmreview.com/article_sub.cfm?articleId=1313, 1996, (Date of access 16/03/2005)

[PENDSE,2005], *Nigel Pendse,* OLAP Market share analysis, 22 March 2005, http://www.olapreport.com/market.htm (Date of access 30/05/2005)

[POLENIS,2002], Shana Polenis, "Data Marts as management information delivery mechanisms", Magister Informationis Scientiae specialising in Information Science, University of Pretoria, November 2002

[ORR,2000] Ken Orr, data warehousing technology, http://www.kenorrinst.com/dwpaper.html, revised edition 2000, (Date of access 16/03/2005)

[STANFORD] The quotation attributed to Stanford University; published at http://www. datawarehousing.com

[TANRIKORUR,1998] Tulu Tanrikorur , "Enterprise DSS Architecture:A Hybrid Approach", DM Review Magazine, February 1998 Issue

[VAVOURAS,2002], Athanasios VAVOURAS, "A Meta data-Driven Approach for data Warehouse Refreshment", DISSERTATION DER WIRTSCHAFTSWISSENSCHAFTLICHEN FAKULTÄT DER UNIVERSITÄT ZÜRICH, February 2002