DATA WAREHOUSING

AT

AVONDALE COLLEGE

A Paper
Presented in partial fulfillment
of the requirements for
the Master of Computing

by
David Heise
23 December, 1996

Macquarie University

This document can be viewed online at http://www.andrews.edu/~dheise/DW/ACDWTOC.html

Table of Contents

List of Figures

List of Tables

Acknowledgments:

Used in this report:

Word Processing

Word processing was done using Microsoft Word 7.0a.

Graphics

All diagrams were drawn using Microsoft Word Draw. The diagrams were grouped and pasted into SmartDraw 3.0, a shareware program from SmartDraw Software Inc. (http://smartdraw.com/) and exported in Windows Meta File format and in GIF format. The WMF files were linked and embedded in the Word document. The GIF files were linked to the web version of the report.

Bibliography

A bibliographic database was maintained in Microsoft Access. Access reports were written to export this as HTML with hypertext links for the web references, and as Word formatted for inclusion in this report.

Web

The conversion to HTML was assisted using WebEdit PRO 2.0.1 for Windows 95, a shareware program from Nesbitt Software Corporation (http://www.nesbitt.com).

Used in the project:

CASE

Visible Analyst Workbench from Visible Systems was used for developing data models for the data warehouse. It was used for creating diagrammatic representations of the data warehouse, as well as for generating the relational database schema definition.

Relational Database

The data warehouse was built using Oracle RDB/SQL V6.1 on a DEC Vax Cluster running VMS V6.0.

Data Replication

Programs were written to extract data, and translate and load it into the data warehouse using the Cognos 4GL PowerHouse QTP.

Metadata

Data descriptions and meanings were extracted from the PowerHouse data repository using the Cognos 4GL PowerHouse QUIZ. A Microsoft Word macro then formatted this into user documentation. Data replication programs recorded update counts and times.

User Desktop Tools

For end user reporting and enquiry, Cognos Impromptu was used. For multi-dimensional analysis, Cognos PowerPlay was used.

I would like to thank Professor C.N.G. (Kit) Dampney from Macquarie University and Mr. Lyndon Harris, Director of Information Technology Services at Avondale College, for their encouragement and support in this project.

Executive Summary

In all spheres of business, including higher education, increasing emphasis is being placed on the role of Information Technology in becoming more competitive. More effort is being placed on aligning the goals of IT with the corporate strategic plan. Attention is being focused on the value of the investment in information resources and its potential, through improved decision support, to gain a competitive advantage.

Corporate data processing has traditionally had an operational focus, rather than a decision and analysis support focus. More and more information was available to executives via standard management reports, but there was no unified approach to placing analysis tools and supporting data in the hands of these decision makers. Data Warehousing is seen as the technology which attempts to address this problem.

Data Warehousing began to be a popular term in the mid 90's, even though the concepts had been practiced in some places for many years. The emergence of Data Warehousing as a so-called new technology came about for a variety of reasons.

- executive decision-making was becoming increasingly dependent on accurate up-to-date information, in order to remain competitive
- executives and their staffs were being drowned in a flood of detailed information from which it was difficult to plot strategies and analyze trends
- requests for ad hoc analysis and queries had to be programmed by computer professionals, already laboring with a lengthy backlog
- a multitude of factors prevented executives from performing their own analyses and creating their own enquiries on the data

This report documents a study into Data Warehousing and its application to providing management assistance at Avondale College, a small independent University in New South Wales, Australia. The report describes a pilot project at Avondale College which set out to build a data warehouse, develop decision support models and implement business intelligence desktop tools.

Broad Objectives

- provide an information system that is aligned with business goals
- provide a system that assists in clarifying goals and determining strategies
- integrate existing diverse data sources
- support decision making with enquiry, reporting, analysis and data mining tools
- support users with training, documentation and on-going support

After analyzing a broad spectrum of management needs, two factors were chosen from the academic administration area that were considered to be of paramount importance, that were not available from the current system, but that could be delivered via a data warehouse with appropriate business intelligence tools. The two factors were:

1. cohort retention rates
2. student performance

At the time of writing this report, the project is languishing somewhat, due in part to the Project Leader (the author) being transferred overseas. The project is reviewed, and conclusions and recommendations are given in Section 5.

1. Introduction

1.1 Background

This project began as an Executive Information System (EIS, sometimes known as an Enterprise Information System) project. Federal Government funding was offered from the National Priority (Reserve) Fund in 1994 for two categories of project – management improvement and course development. Avondale College was successful in obtaining a $20,000 grant in the category of "management improvement". In the proposal applying for the grant, the project was summarized as follows:

> "A corporate database will be designed with improved management information and decision support as the primary focus. Key administrators will be trained in the use of EIS tools to enable them to make the best use of the data resource.
>
> "Transfer programs will be written to automate the replication of data held in existing systems into the new corporate database. The database will be designed to be CASMAC[1] compliant, so that as legacy systems are replaced, they will operate on the new database."

Given that the initial impetus for this project came at least partly from a management improvement grant, it is appropriate to examine briefly the management and decision making processes at Avondale College. Owing to its small size, Avondale has a relatively simple management structure.

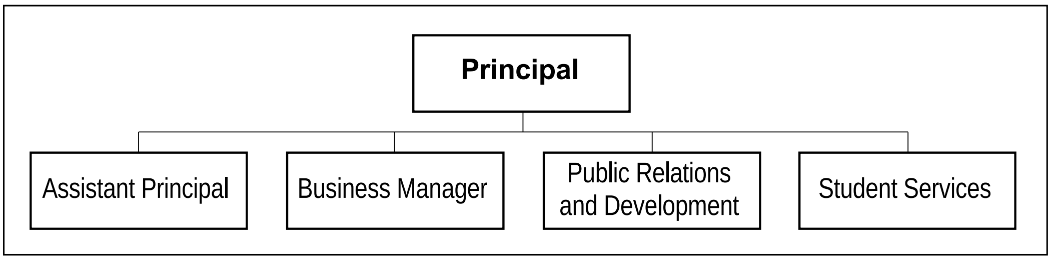


Figure 1. Management Structure

An administrative group consisting of the five people represented in this chart meets weekly. It is responsible for tactical and operational policies and procedures, as well as strategic planning and long term goal setting. It is supported by other management and advisory committees.

This administrative group and department heads reporting directly to them will be the major users this data warehousing system and its associated desktop tools.

[1] CASMAC is the Core Australian Specification for Management Computing.

1.2 Section Summaries

The report opens with an Executive Summary and this Introduction.

Section 2 lists some of the pressing reasons why Avondale College undertook this Data Warehousing pilot project.

Section 3 is a summary of Data Warehousing principles, particularly those that have relevance at Avondale.

Section 4 documents aspects of the project itself that were significant in terms of the time or effort they demanded, social or technical implications, or their overall importance to the success of the project.

Section 5 concludes the report with some observations on the pilot project and suggestions for the future.

The report also includes a comprehensive bibliography, some appendices and an index.

2. Motivation For This Project

Contents:

2.1 Decision-making Processes

Management processes at Avondale College have tended not to rely to a major extent on input from computer-based information systems. There are a number of reasons for this.

1. The financial software was developed in-house almost twenty years ago, and is primarily a record keeping system. The system does not produce management information.
2. Administrators and decision makers have not worked with systems that provide decision support, so they are not experienced with the kind of assistance that state-of-the-art management information systems can provide. As a result, their requests for information have been infrequent and relatively unsophisticated.
3. When the Computer Services Center was asked for decision support information, it was usually not able to provide it in a timely manner with the current systems, and that discouraged the administrators from making further requests.

2.2 Integrating Multiple Data Sources and Platforms

One of the major difficulties faced by decision makers and knowledge workers is that the corporate data resides in many different systems. Data is held in many different formats and on a variety of platforms.

The major information system components are as follows:

Component	**Platform**	**Data Format**
Student Administration	VAX	indexed flat files
General Ledger	VAX	Databus emulation
Budgeting, Financial Planning	PC	spreadsheets
DEET Reporting		
Student	VAX	indexed flat files
Staff	none	manual
Finance	VAX	Databus emulation
Banking	VAX	Databus emulation
Payroll	PC	Progress
Human Resources	PC	dBase
Accounts Payable	none	manual
Accounts Receivable	VAX	Databus emulation
Asset Management	VAX	Databus emulation
Recruiting, Marketing & Promotion	PC	Microsoft Access
Management Information	none	none
Ad hoc enquiry/analysis	none	none
EIS	PC	spreadsheets
Other	none	hardcopy

Table 1. Information System Components and Platforms

The same data item sometimes has different names or different attributes in each system. A simple example is the different ways that student ID can be treated in different systems.

```
STUDENT_ID                     PIC 9(10).
STUDENT_NUMBER                 PIC 9(5).
NAME_NUMBER                    PIC X(6).
  (NAME_NUMBER replaces STUDENT_NUMBER plus CHECK_DIGIT)
```

Table 2. Semantic Differences in STUDENT ID

2.3 IT Backlog

Because the data is not in a form suitable for user-generated analysis, enquiry and reporting, each request for decision support information requires special routines to be written by programmers. The IT department has a large backlog of projects and requests typical of the industry. These include requests for modifications, bug fixes, additional functionality and enhancements to existing systems, as well as new projects for replacing legacy systems and adding new systems, either through evaluating and implementing packaged software or in-house development.

The demands for accurate and timely management information to support ad hoc enquiries and reports, graphical and what-if analysis are increasing, and the IT backlog continues to grow larger. The answers to the questions are held in the data, but it is too difficult and time consuming to extract the information needed to find the answers. This results in questions not being asked, and in duplication of personal information systems, with inadequate controls over data integrity and accuracy.

2.4 End User Data Access

One solution to the problems caused by the IT backlog is to put user-oriented reporting and analysis tools in the hands of the users. This will assist the decision making process and it will help to realize more of the potential benefit of information systems in meeting the corporate goals.

The problem is that in its current format, corporate data is not accessible via end-user tools.

1. All of the financial data is processed using Databus programs via an emulator on the VAX. Software systems outside the emulator cannot access the financial data.
2. Student Administration data is accessible to other applications running on the VAX, but it is not held in a relational database. Hence, the SQL server used by most client/server EIS tools cannot access student data.
3. There is only limited personnel information in the Payroll system, but that is on a PC so even that limited information is not available to users on terminals or PC's networked to the VAX.

The long-term objective for Information Systems (known as the Administrative Software Project) is to develop and maintain a central corporate database in a form conducive to analysis via EIS tools. However, the problem here is that the timescale for completion is 5 to 8 years, so this does not address the urgent need for improved management information.

There is a need to:

- provide corporate data in a form that can be analyzed with end-user EIS tools,
- in the shortest possible time,
- in a way that is compliant with the CASMAC specification,
- and in a way that can be used as the foundation for the Administrative Software Project.

2.5 User Friendly Data Design

In order to provide end user access to data, users need to find it easy to navigate and use the database. Rather than wait until a fully integrated package can be purchased or developed in-house, a database can be designed to satisfy the requirements for ease-of-use by end users, with routines to populate it from existing data sources. This is essentially the concept of the Data Warehouse. Since this database will not be supporting operational data processing, its design can be optimized to support the needs of decision makers and knowledge workers. An integral part of make this data warehouse easy for user to navigate is the metadata. All data meanings, sources and refresh details will be held in the metadata, with procedures for maintaining them as the project develops.

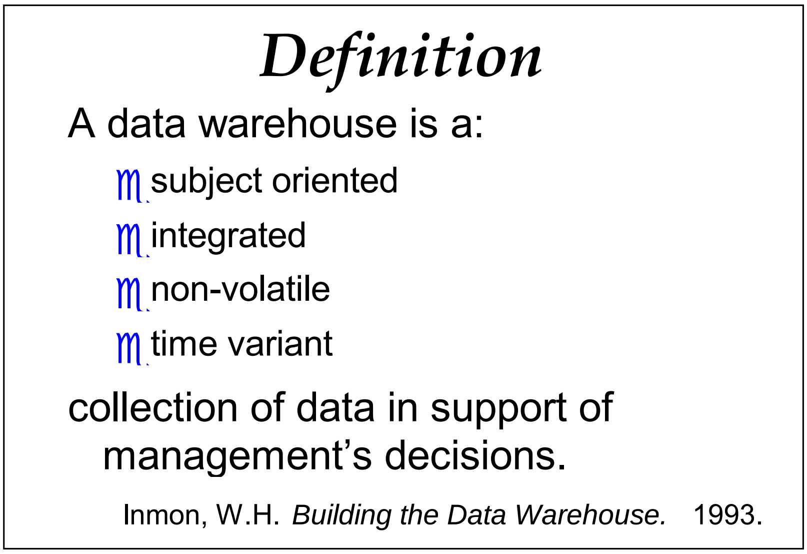
3. Data Warehousing Principles

Contents:

This section summarizes data warehousing concepts found in White Papers published on the web by vendors, consultants, practitioners and researchers, as well as material found in journals and reference books and from experience. Where possible, examples and illustrations are taken from experiences with the Avondale College data warehouse.

3.1 A Data Warehouse Defined

In 1993, the "father of data warehousing", Bill Inmon, gave this definition of a data warehouse:



Definition

A data warehouse is a:

- subject oriented
- integrated
- non-volatile
- time variant

collection of data in support of management's decisions.

Inmon, W.H. *Building the Data Warehouse.* 1993.

Figure 2. Definition of a Data Warehouse[2]

Many of the concepts and practices of data warehousing have existed for years, but it is only within the last few years that the term has acquired "buzz word" status. While it is true that software is available for automating some of the data warehouse processing, a data warehouse is not a product - it is not something that can be purchased from a vendor. Rather, it is a model of a corporation's data, put together in such a way that it answers the corporation's business questions.

According to this definition, a data warehouse is different from an operational database in four important ways.

Data Warehouse	Operational Database
subject oriented	application oriented
integrated	multiple diverse sources
time-variant	real-time, current
nonvolatile	updateable

Table 3. Comparing a Data Warehouse and an Operational Database

[2] Inmon, W.H. *Building the Data Warehouse.* p33.

An operational database is designed primarily to support day to day operations. A data warehouse is designed to support strategic decision making.

3.2 Different Data For Different Uses

One of the fundamental assumptions of data warehousing is that operational data needs to be stored separately in a different format in order to support data analysis. This diagram illustrates the fact that different sets of users access the data, using different sets of applications and for different purposes.

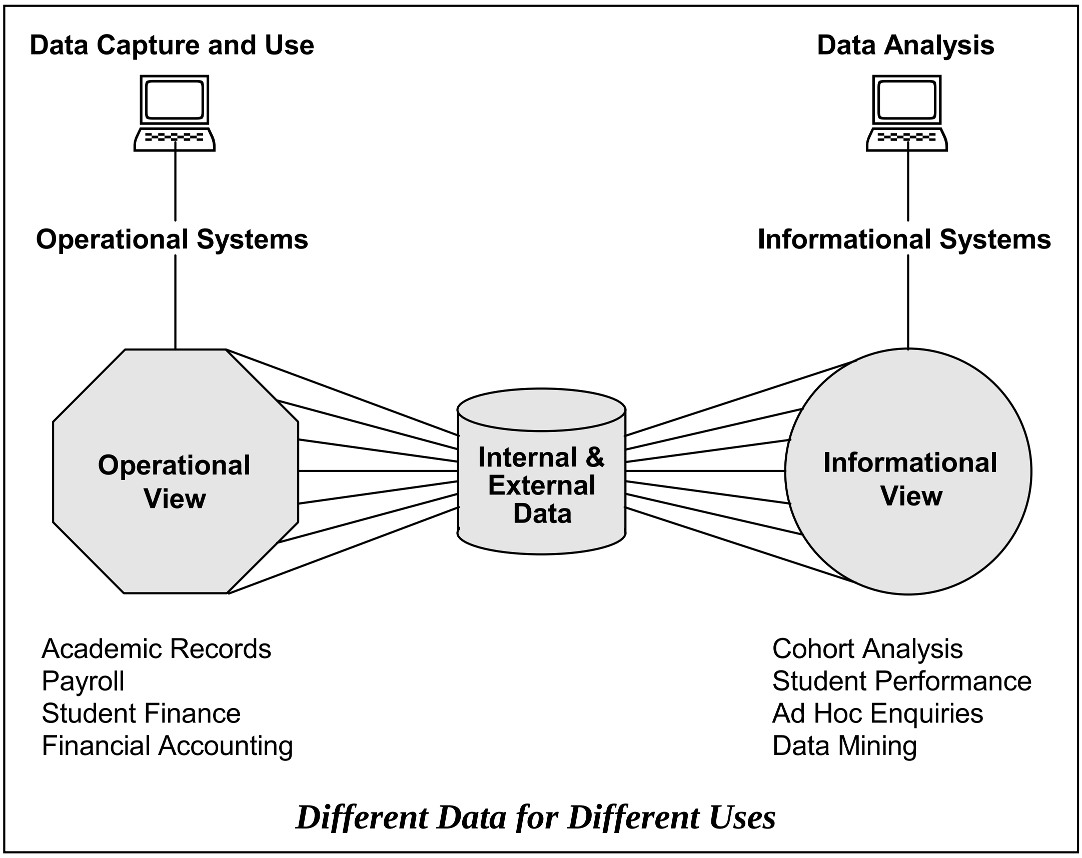


Figure 3. Operational and Informational Data[3]

A statement from *The IBM Information Warehouse Solution* describes the clear distinction that needs to be made between operational and informational data.

> "Most organizations need two different data environment, one optimized for operational applications and one optimized for informational applications. For example, operational applications and databases are typically optimized for fast response time and typically cannot tolerate the impact on response time created when access by an informational application. The two types of applications are fundamentally different. If the same data environment is used to support both, the performance, capability and benefit of both will be compromised."[4]

[3] Adapted from: IBM. *The IBM Information Warehouse Solution: A Data Warehouse Plus!* p3.

[4] IBM. *The IBM Information Warehouse Solution: A Data Warehouse Plus!* p3.

Within a data warehouse implementation itself, the following types of data will be required to support typical uses:

- Real-time Data
- Reconciled Data
- Derived Data
- Changed Data
- Metadata

These are described more fully in *The IBM Information Warehouse Solution*.

Oracle has published a useful table that compares three different types of database - OTLP, report/enquiry DSS and OLAP.

Characteristics	**RDBMS OLTP**	**RDBMS DSS**	**MDBMS OLAP**
Typical operation	Update	Report	Analyze
Level of analytical requirements	Low	Medium	High
Screens	Unchanging	User-defined	User-defined
Amount of data per transaction	Small	Small to large	Large
Data level	Detail	Detail to summary	Mostly summary
Age of data	Current	Historical and current	Historical, current and projected
Orientation	Records	Records	Arrays

Table 4. Comparing OLTP,DSS and OLAP Databases[5]

3.3 Need For Better Analysis Tools

As data processing systems record more and more data, it is not enough for IT to produce more and more management reports to be viewed by decision makers. In "Relational OLAP: Expectations & Reality", an Arbor Software White Paper, the author states that "organizations are attempting to maximize the business value of the data that are available in ever increasing volumes from operational systems, spreadsheets, external databases and business partners. It is not enough to simply view this data – business value comes from using it to make better informed decisions more quickly, and creating more realistic business plans. In the past, these decisions have often been made based on 'gut feel' and experience rather than solid data, analyses and tested hypotheses. With the flattening of management structures, re-engineered businesses and globalization, the need for better analysis tools is greater than ever."[6]

[5] Oracle. *Oracle OLAP Products: Adding Value to the Data Warehouse. An Oracle White Paper*. p4.

[6] Arbor Software. *Relational OLAP: Expectations & Reality*. p3.

3.4 Types of Data Warehousing Applications

Data warehousing systems target at least three different types of applications:

- personal productivity
- query and reporting
- planning and analysis

These are described well in a White Paper from Arbor Software entitled "*The Role of the Multidimensinal Database in a Data Warehousing Solution*".

> "As with all information systems, it is best to view data warehousing's core components against a framework that focuses not on technology, but on the business applications the system is designed to address. In general, the applications served by data warehousing can be placed in one of three main categories.
>
> **"Personal productivity applications** such as spreadsheets, statistical packages and graphics tools, are useful for manipulating and presenting data on individual PCs. Developed for a standalone environment, these tools address applications requiring only small volumes of warehouse data.
>
> **"Data query and reporting applications** deliver warehouse-wide data access through simple, list-oriented queries, and the generation of basic reports. These reports provide a view of historical data but do not address the enterprise need for in-depth analysis and planning.
>
> **"Planning and analysis applications** address such essential business requirements as budgeting, forecasting, product line and customer profitability, sales analysis, financial consolidations and manufacturing mix analysis--applications that use historical, projected and derived data.
>
> "These planning and analysis requirements, referred to as on-line analytical processing (OLAP) applications, share a set of user requirements that cannot be met by applying query tools against the historical data maintained in the warehouse repository. The planning and analysis function mandates that the organization look not only at past performance but, more importantly, at the future performance of the business. It is essential to create operational scenarios that are shaped by the past, yet also include planned and potential changes that will impact tomorrow's corporate performance. The combined analysis of historical data with future projections is critical to the success of today's corporation."[7]

[7] Arbor Software. *The Role of the Multidimensional Database in a Data Warehousing Solution.* p2.

3.5 Evolution of Data Across Three Server Platforms

The following table relates multidimensional databases (OLAP) to OLTP and relational database servers.

			DECISION SUPPORT SERVERS	
		OLTP Server	Relational Database (Warehouse repository)	Multi-dimensional Database (OLAP)
PURPOSE	System charter	Operational	Informational	Analytical
	Business significance	Mission critical	Informational critical	Management critical
ACCESS	Access type	Read/write	Read	Read/write
	Access mode	Singular, simple update queries	Singular, simple queries list oriented	Iterative, comparative, analytical investigation
	Access process	IS-supported queries	IS-assisted or preplanned queries	IS-independent ad hoc navigation and investigation, drill-down
	Response characteristics	Fast update, varied query response	Varied query response	Fast/consistent response
DATA	Content scope	Application specific • Actual/vertical • Limited historical	Cross-subject database • Actual/horizontal • Historical/archival data	Application specific • Actual/horizontal • Projected/"what if" data • Derived data
	Data detail level	Transaction detail	Cleansed & summarized	Summarized, aggregated & consolidated using complex computations (ratios, allocations, variance, time series and gross margin %)
	Data structure	Normalized	Denormalized	Multi-dimensional Hierarchical
	Data structure design goal	Update	Query	Analysis
	Data volumes	Gigabytes	Gigabytes/terabytes	Gigabytes
IMPLEMENTATION	Deployability	Slow(multi-month/yr.)	Slow(multi-month)	Fast(days/weeks)
	Adaptability	Limited	Low Significant resource	High Easily modified
	Computer hardware investment required	Extensive/high cost hardware	Moderate/medium cost hardware	Minimal/low cost hardware

Table 5. Evolution of data across three server platforms[8]

[8] Arbor Software. *The Role of the Multidimensional Database in a Data Warehousing Solution.* p3.

3.6 Benefits of Data Warehousing

Data warehousing is being hailed as one of the most strategically significant developments in information processing in recent times. One of the reasons for this is that it is seen as part of the answer to information overload.

Some of the benefits of data warehousing that were seen as relevant to the Avondale College project are listed here. The points highlighted in the Bill Inmon's definition give some of the reasons why data warehousing is regarded as important.

1. Has a subject area orientation
2. Integrates data from multiple, diverse sources
3. Allows for analysis of data over time
4. Adds *ad hoc* reporting and enquiry
5. Provides analysis capabilities to decision makers
6. Relieves the development burden on IT
7. Provides improved performance for complex analytical queries
8. Relieves processing burden on transaction oriented databases
9. Allows for a continuous planning process
10. Converts corporate data into strategic information

Table 6. Benefits of Data Warehousing

3.6.1 Has a subject area orientation

Data is arranged by subject rather than by application, and is more intuitive for users to navigate. This is closer in concept to the way decision makers think about their business.

3.6.2 Integrates data from multiple, diverse sources

One of the initial motivations for this data warehousing project at Avondale College was that we had multiple, diverse sources of data, and integrating them into a single administrative suite was known to be a long term project for which there was neither the staff, the funds nor the time. Many information needs were not being met, and the integration provided by a data warehouse was seen as very desirable.

Table 1 (*Information System Components and Platforms*) illustrates the diversity of data formats and platforms currently in use at Avondale College. Data that is currently stored in separate indexed flat file and Databus emulation systems and that is relevant to the pilot project has been integrated in the data warehouse. This is of immediate benefit to knowledge workers who are now able to formulate their own queries and write their own ad hoc reports. It is also used as the basis for the aggregation and summarization that makes up the multi-dimensional database.

3.6.3 Allows for analysis of data over time

The operational database provides detailed current information, often with last years data available for comparison, but analytical queries typically require much more. Plotting trends and looking for relationships over time, at all of the possible levels of aggregation, is not provided for in the operational database.

With a data warehouse, data snapshots taken at times that are significant to the decision making process make it possible to analyze trends over time. A typical example is to monitor the numbers of applications received, accepted and rejected, and the number who actually enroll. Snapshots can be taken at specified points in time over the months, weeks and days leading up to and immediately following registration. These can be plotted over recent years, and provide one way to assess the impact of promotional programs, changes in course offerings, changes in government regulations. This can readily indicate areas where further analysis is warranted.

3.6.4 Adds **ad hoc** *reporting and enquiry*

In addition to the points highlighted in the definition, a data warehouse is designed to be accessible with end-user tools, and this allows ad hoc reporting and analysis by end-users.

3.6.5 Provides analysis capabilities to decision makers

Until quite recently, management information meant hardcopy summaries and exception reports. Some of these were produced by automated processes running daily, but others required considerable manual preparation and could be delivered weeks or even months after the relevant time period had closed. A growing need to work interactively with the data has been identified, and is referred to generically as analysis.

> "Today's markets are much more competitive and dynamic than those in the past. Business enterprises prosper or fail according to the sophistication and speed of their information systems, and their ability to analyze and synthesize information using those systems. The numbers of individuals within an enterprise who have a need to perform more sophisticated analysis is growing."[9]

Knowledge workers have been attempting to meet this need using spreadsheets and simple, general purpose report writers based on data extracts prepared by the IT department.

> "Most notably lacking has been the ability to consolidate, view, and analyze data according to multiple dimensions, in ways that make sense to one or more specific enterprise analysts at any given point in time. This requirement is called 'multi-dimensional data analysis.' Perhaps a better and more generic name for this type of functionality is on-line analytical

[9] Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* p6.

> processing (OLAP), wherein multi-dimensional data analysis is but one of its characteristics."[10]
>
> …
>
> "OLAP is made up of numerous, speculative 'what-if' and/or 'why' data model scenarios executed within the context of some specific historical basis and perspective."[11]
>
> …
>
> "There are typically a number of different dimensions from which a given pool of data can be analyzed. This plural perspective, or Multi-Dimensional Conceptual View appears to be the way most business persons naturally view their enterprise."[12]

The multiple data dimensions correspond to data consolidation paths. This is what allows the end-user to chose to view certain data dimensions while aggregating over others. The choices of what dimensions to view and whether to "drill down" or "roll up" are made interactively, and may reveal new or unanticipated relationships in the data. This is referred to as dynamic data analysis.

> "Dynamic data analysis can provide an understanding of the changes occurring within a business enterprise, and may be used to identify candidate solutions to specific business challenges as they are uncovered, and to facilitate the development of future strategic and tactical formulae."[13]

[10] Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* p8.

[11] Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* p10.

[12] Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* p11.

[13] Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* p13.

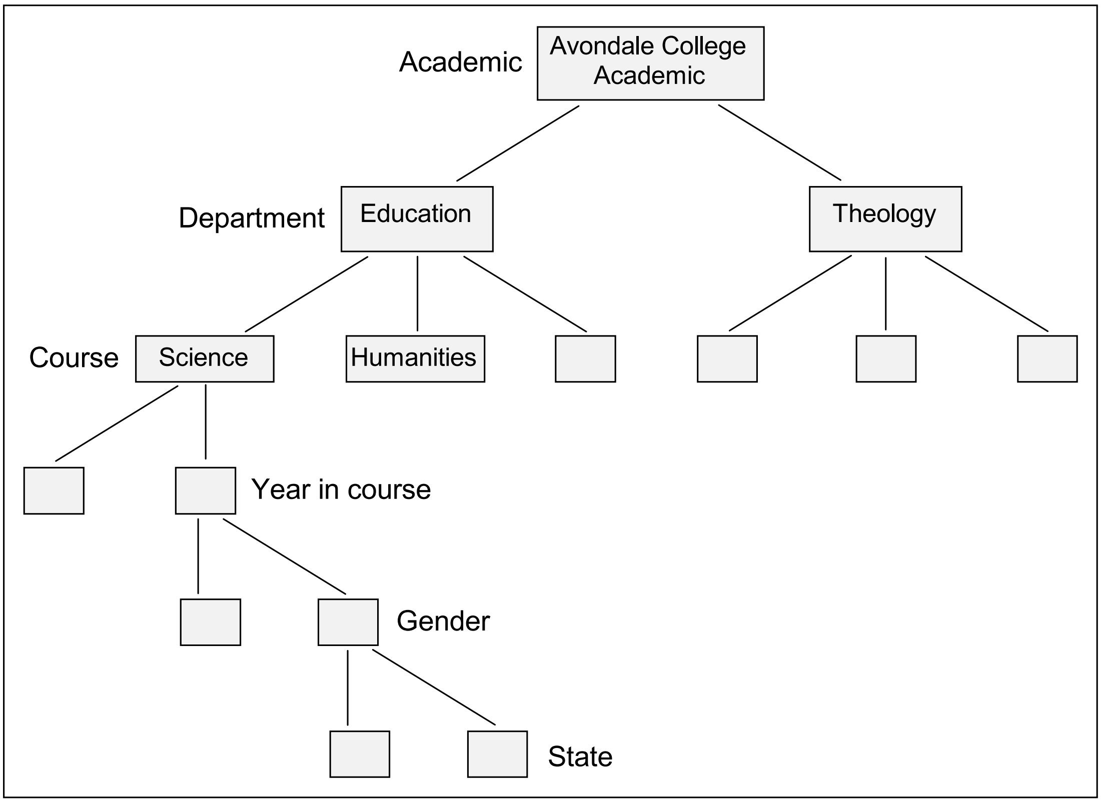


Figure 4. Sample Data Consolidation Paths[14]

Each level in the above diagram (State, Gender, Year in course, etc) correspond to dimensions in a model used to analyze facts such as student performance. While the diagram seems to imply a fixed hierarchy, this is actually not the case. Any combination of any number of dimensions can be analyzed and displayed in graphical or tabular form using typical MDA tools.

3.6.6 Relieves the development burden on IT

At Avondale, there was a wealth of corporate data that was virtually inaccessible to users because each request for information required code to be written by IT, which already had a large backlog of requests. With a data warehouse and users trained in the use of appropriate desktop tools, users can find answers to their own questions.

3.6.7 Provides improved performance for complex analytical queries

Online Transaction Processing (OLTP) systems are optimized for update and reporting with relatively simple cross-table joins. Multi-dimensional analysis (MDA) based on OLTP databases suffers in performance for two reasons. Firstly, the normalized nature of OLTP databases and the multi-dimensional nature of typical query and analysis operations means that many joins have to be done just to

[14] Adapted from Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* p12.

lookup or decode coded data such as course names and student acceptance status meanings. Secondly, and more importantly as noted in *Aggregate-Query Processing in Data Warehousing Environments*, "aggregate queries are frequent in decision support applications, where large history tables often are joined with other tables and aggregated. Because the tables are large, better optimization of aggregate queries has the potential to result in huge performance gains. Unfortunately, aggregation operators behave differently from standard relational operators like select, project, and join."[15]

A data warehouse provides improved performance for complex analytical queries by denormalization and aggregation. Frequently used aggregates are often precomputed and materialized in views commonly known as summary tables These materialized views provide fast access to integrated data, regardless of the original data sources.

Precomputing the data integration and aggregation is sometimes known as the *eager* or *in-advance* approach. The approach that must be taken if no data warehouse is available is referred to as *lazy* or *on-demand*. This is one of the fundamental benefits of data warehousing. Taking the eager approach:

"1. Information from each source that may be of interest is extracted in advance, translated and filtered as appropriate, merged with relevant information from other sources, and stored in a (logically) centralized repository.

2. When a query is posed, the query is evaluated directly at the repository, without accessing the original information sources."[16]

3.6.8 Relieves processing burden on transaction oriented databases

Not only do analytical queries give frustratingly poor response time for those who are performing them, they are also very demanding on resources needed for adequate transaction processing performance. This problem is not solved simply by installing bigger, faster database servers. The optimization required for one environment is fundamentally different from what is required in the other. A more effective solution to the performance problem for both OLTP and OLAP is to separate the data into independently designed structures, as a minimum step. In the case of very large data warehouses, it becomes necessary to use separate hardware as well as physically separate databases.

[15] Gupta A., Harionarayan V. and Quass D.: Stanford University. *Aggregate-Query Processing in Data Warehousing Environments.* p1.

[16] Widom J.: Stanford University. *Research Problems in Data Warehousing.* p1.

3.6.9 Allows for a continuous planning process

> "An intensively competitive and continuously changing business climate is driving (educational institutions) to more frequently evaluate business structures and the allocation of resources. In this environment, the once-a-year budget is being replaced by a continuous planning process. The rolling forecast and other *ad hoc* analyses are becoming the principal vehicles for providing management with the information needed to make fast-paced business decisions and to manage resource allocations."[17]

In the education marketplace, competition is increasing as universities develop more aggressive marketing strategies. Economic and political factors are having a marked affect on the numbers of students enrolling in institutions of higher learning, leading to reduced intakes at many universities. Demographics indicate that smaller family sizes now means that the number of potential students to offer places to is no longer able to support the growth that many universities are aiming to achieve. This has lead to "re-focusing" on core objectives and "re-engineering" of fundamental processes. In addition to these internal pressures, universities are facing increased competition from overseas.

Avondale College is actively pursuing the development of its marketing capacity. To avoid a threatened enrollment crisis, strategic planning exercises have been conducted with the help of consultants. The renewed focus on enrollment planning and management will depend very heavily on the Data Warehouse project described in this report.

The objective is to make planning a continuous process by providing powerful, integrated systems so that rolling forecasts and *ad hoc* analyses incorporate the level of detail, department head involvement and analytical richness of budgeting. The key to an integrated system for planning, analysis and reporting is the data warehouse and appropriate desktop tools.

3.6.10 Converts corporate data into strategic information

In a White Paper entitled "Multidimensional Analysis: Converting Corporate Data into Strategic Information", Arbor Software state:

> "Information is a strategic weapon in today's fast-paced global business environment. Companies are realizing that the key to successful competition and growth lies in their abilities to quickly obtain the right information for spotting trends, forecasting market changes, and analyzing performance. An in an effort to better manage the sheer volume of data available, they have invested heavily in information systems and technologies at both the corporate information system level and on individual's desktops."[18]

[17] Arbor Software. *An Enterprise Solution for Financial Planning & Analysis.* fpandac1.html p1.

[18] Arbor Software. *Multidimensional Analysis: Converting Corporate Data into Strategic Information.* multic0.html p1.

Spreadsheets	Relational Database Management System (RDBMS)
• single user tools only	• enterprise in scope
• limited to simple analyses, small data sets	• includes central repository, online transaction processing (OLTP) and large data sets
• slow consolidation	• inflexible in design, requiring extensive consulting or programming to obtain satisfactory performance
	• often restricted by proprietary structures and memory shortage requirements
	• slow for *ad hoc* analysis
	• require extensive training

Table 7. Limitations of Traditional Tools[19]

Through the use of data warehousing and appropriate desktop tools, the information overload can be reduced to meaningful strategic information, and the limitations of traditional tools can be overcome. The many dimensions of typical data can be summarized to just those of interest. Viewing other dimensions is accomplished simply by clicking its title button and dragging it into the viewing area.

3.7 Costs of Data Warehousing

There are down sides of course. Data warehouses can be extremely expensive to build and maintain, and have a high failure rate unless there is the right mix of high need, powerful sponsor, and reasonably short time scale. Also, the impact they can have on traditional views of data ownership and organizational structures can be quite disruptive.

Some of the more noticeable costs associated with data warehousing are listed in the following table, and are described below.

1. Time spent in careful analysis of measurable needs
2. Design and implementation effort
3. Hardware costs
4. Software costs
5. On-going support and maintenance
6. Resulting re-engineering effort

Table 8. Costs of Data Warehousing

[19] Arbor Software. *Multidimensional Analysis: Converting Corporate Data into Strategic Information.* multic1.html p2.

3.7.1 Time spent in careful analysis of measurable needs

Especially in the initial stages of the first pilot project, it can be very time-consuming and frustrating trying to establish needs that can be measured and that are important in decision making processes. End users will struggle to grasp what it is that they are being asked to do. A typical end user response is "give me what I say I want, then I can tell you what I really want."

Time will need to be spent demonstrating examples, where available, and showing prototypes of what is being developed. This process needs to be done repeatedly, and the system that is developed must be flexible. To be successful, the end user must be able to explore the possibilities. They need to be able to find answers to questions they had not thought of asking.

3.7.2 Design and implementation effort

One of the reasons why many data warehouse projects fail is that they attempt to do too much. The key is to choose a pilot project that has a high profile but is manageable in size. One rule of thumb suggests that a data warehouse should have its first model working within four months or the project has a high likelihood of failure.

Spending too much time on design and implementation can cost the success of the entire project. The most critical time is spent in choosing the actual measures to model.

3.7.3 Hardware costs

In very simple terms, a data warehouse requires additional disk space, and lots of it, but that is not all. Except in the case of small data warehouses, such as departmental data marts, significant processing power and very large amounts of disk storage are required to deliver acceptable performance to analysts. The data warehouse needs to be on hardware separate from the operational systems, in order the transaction processing is not adversely impacted. Large data warehouses require a major investment in hardware, typically for multiple CPU configurations, such as Symmetric Multi Processing (SMP) or Massively Parallel Processing (MPP) systems.

3.7.4 Software costs

Vendors of relational databases and other data related tools are rapidly developing a range of products to support data warehousing. This includes software to automate data extraction, cleanup, manipulation, aggregation, scheduling. It includes several different OLAP server technologies, as well as a variety of business intelligence tools for end user desktops. There are tools to assist with data design, change detection, automated maintenance for structural changes, automated maintenance and synchronization of metadata.

Whether solutions are purchased from vendors, or the work is done manually or with on-house developed code, it comes only at significant cost.

3.7.5 On-going support and maintenance

This cost was alluded to in the previous point. In addition to contracted software support costs, support staff will be needed to install updates, perform maintenance reprogram tools as requirements change.

3.7.6 Resulting re-engineering effort

One of the potential benefits of data warehousing is that it can reveal trends early enough to respond effectively. The response could be to take corrective action when a negative trend is detected, or to capitalize on actions that are causing a positive trend. However, the changes will very often involve changes to processes, work procedures and organizational structures. These can be very expensive in human terms because of the disruption that often accompanies these changes. This is one of the effects noted by C.N.G. Dampney when describing the "Growth of Impact" of an Information System on an organization. As Figure 5 shows, this impact can be described as a progression.

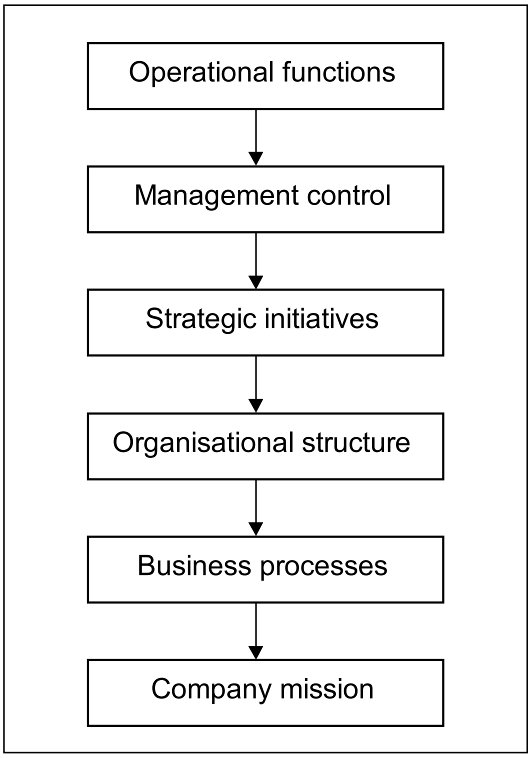


Figure 5. Growth of Impact[20]

Data warehousing starts with operational data, and looks at questions asked by management to meet corporate goals and develop strategic initiatives. Data that provides answers to these questions is what is placed in the data warehouse. One of the direct effects of such an approach is to impact organizational structure and business processes.

[20] CNG Dampney COMP820 *Notes for Session 3 Harnessing the Information Resource.* p23.

3.8 Data Warehousing Processes and Functions

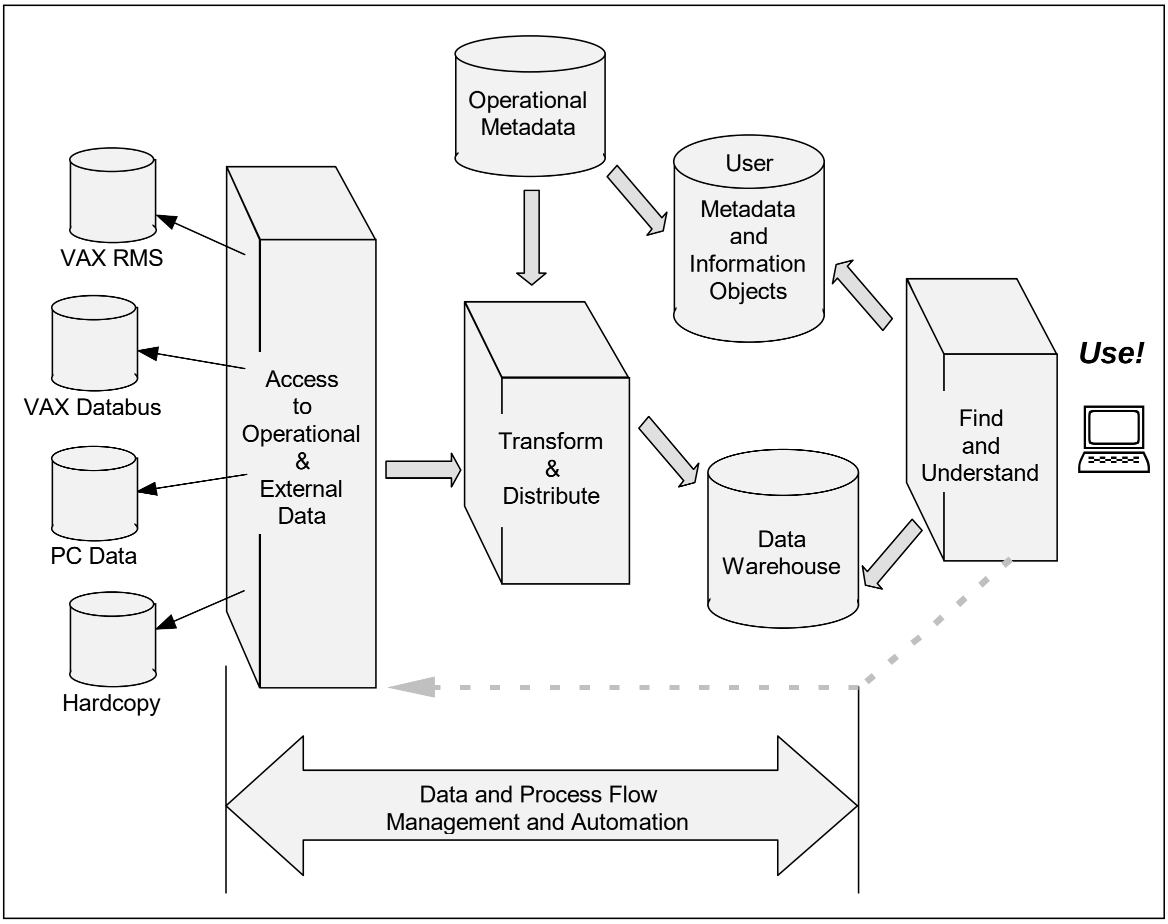

Operational Metadata
User
Metadata and Information Objects
VAX RMS
VAX Databus
PC Data
Hardcopy
Access to Operational & External Data
Transform & Distribute
Data Warehouse
Find and Understand
Use!
Data and Process Flow Management and Automation

Figure 6. Data Warehousing - A Process View[21]

This diagram gives a detailed view of the processes involved in managing and maintaining a data warehouse. The processes run from left to right, with a feedback loop from the users. One of the very clear lessons of data warehousing is that you don't build one in the way you build a house. Iteration and refinement is vital. The clue is to start small, and then evolve the data warehouse as the needs develop.

Flexibility and the ability to adapt to changing business needs are essential. Some vendors are beginning to talk about tools for automating maintenance. For this to happen, the management of the metadata needs to become more tightly integrated into the data warehousing process.

However, one of the fundamental assumptions of a data warehouse is that it is scaleable. All of the advice I have seen suggests starting small with a pilot project, and then letting it grow. In fact, I read where one consultant predicts that any data

[21] IBM. *The IBM Information Warehouse Solution: A Date Warehouse Plus!* p6.

warehouse that takes longer than 4 months to get its first model working has a high likelihood of failure.

The proprietary multi-dimensional databases work well for smaller data warehouses or departmental 'data marts' up to 4 or 5 gigabytes. Using the more open Relational OLAP database products, data warehouses into the terabytes in size have been built, running on multi-CPU platforms.

The actual design process for developing a data warehouse runs from right to left in this diagram.

1. talk to the users
2. determine their needs in terms that can be measured
3. design a database to support those needs
4. document the data descriptions and other attributes
   (this will ultimately include data sources, time stamps, data meanings that change over time, etc)
5. design the logic for translating data from various sources into an integrated data store
6. write the code for extracting data from the various sources and transforming it into the data warehouse, with updates to the metadata
7. and finally, package the procedures to handle scheduling, management and maintenance

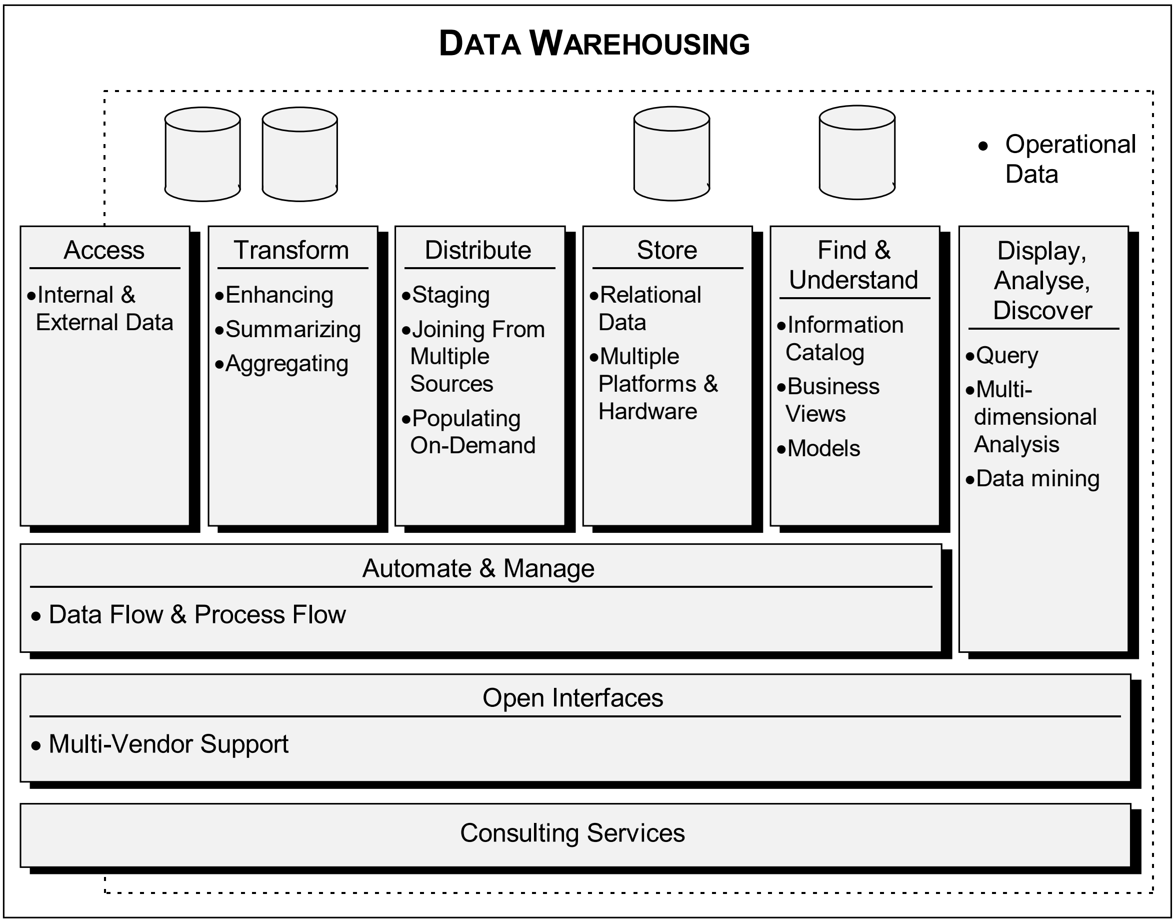


Figure 7. Data Warehousing Functions[22]

Functions that are desired as part of a data warehousing solution are shown in Figure 7. This illustrates the flow of data from originating sources to the user, and includes management and implementation aspects. It starts with access mechanisms for retrieving data from heterogeneous operational data sources. That data is replicated via a transformation model and stored in the data warehouse. The definition of data elements in the data warehouse and in the data sources, and the transformation rules that relate them, are referred to as 'metadata'. Metadata is the means by which the end-user finds and understands the data in the warehouse. The data transformation and movement processes are executed whenever an update to the warehouse data is desired. Different parts of the warehouse may require updates at different times, some at regular intervals such as weekly or monthly, and some on specified dates. There should be a capability to manage and automate the processes required to perform these functions. Particularly in a multi-vendor environment, adopting an architecture with open interfaces would facilitate the integration of the products that implement these functions. Quality consulting services can be an important factor in assuring a successful and cost effective implementation.

[22] IBM. *The IBM Information Warehouse Solution: A Date Warehouse Plus!* p5.

3.9 Important Data Warehousing Concepts

These are just a few of the terms and concepts used when describing data warehouses. More complete glossaries are given in the Appendices. The terms given here relate to storage and processing technologies, and multi-dimensional analysis (MDA). Some of the definitions are taken from page 2 of an IBM White Paper entitled *Multi-Dimensional Analysis: Extending the Information Warehouse Framework.*

3.9.1 OLTP: OnLine Transaction Processing

OLTP is the traditional data processing area, now dominated by relational databases, which have matured into products optimized for transaction throughput.

3.9.2 OLAP: OnLine Analytical Processing

The case for OLAP is very well put in a white paper by E. F. Codd & Associates. OLAP requires the ability to consolidate, view, pivot and rotate, and analyze data according to its multiple dimensions. This requirement is called "multi-dimensional analysis" or MDA.

3.9.3 MDD: Multi-Dimensional Database

An analysts view of the enterprises' universe is typically multi-dimensional in nature. For instance, they will want to analyze a given student population with regard to course, department, school, year in course, gender, age, etc. These could be some of the dimensions for analyzing a retention rate model. The effect the dimensions have on retention rate is observed interactively by an operation know as 'slice and dice'. Consolidation paths can be followed up or down using 'roll-up' or 'drill-down'.

There are a number of proprietary multi-dimensional database formats, one of which was developed by Cognos and is used in their PowerPlay product. The multi-dimensional attributes of this data model - also known as a Hypercube - are designed into the storage technology of the database and the desktop tool that sits on top of the database. All of the various levels of summarization and cross-tabulation are pre-computed and stored using sparse matrix technology.

3.9.4 ROLAP: Relational OLAP

ROLAP or Relational OLAP is the answer to MDD being proposed by vendors of traditional RDBMS. They argue that the multi-dimensionality of data is merely an attribute of the way the data is viewed and made available to user applications. The actual storage technology used to store the views can be treated separately. However, multi-dimensional analysis (MDA) based on OLTP databases suffers in performance for two reasons. Current optimizing algorithms are inappropriate for the resulting complex joins, often spanning large history tables. Also, aggregate queries are frequent in decision support applications. New optimization strategies are needed, with MDA-supporting index types and aggregation operators.

3.9.5 Facts and Dimensions

There are two types of tables in a data warehouse. These are designated as "fact" tables and "dimension" tables. A fact table holds the information that is the subject of the analysis. Facts are metrics which describe the results of business activity. They are scaleable and provide the measurement data on which business decisions are based. They describe the magnitude of business performance from the business strategies and operational activities. Examples of facts are: sales revenue, percent of store sales, and cost of goods sold.

Dimensions are different points of view about the facts. Dimensions describe what facts are. For example, sales revenue is examined for a certain period of time, by a set of markets, and for specific product brands. In this example, period, market, and product are dimensions of the facts.

For different sets of values in the dimension tables, the fact table will hold a different value. A database that is designed for data warehousing will use what is known as a star schema. The star schema supports analysis of facts by any combination of dimensional data.

3.9.6 Drill-down and Roll-up

Drill-down is the repetitive selection and analysis of summarized facts, with each repetition of data selection occurring at a lower level of summarization. An example of drill-down is a multiple-step process where sales revenue is first analyzed by year, then by quarter, and finally by month. Each iteration of drill-down returns sales revenue at a lower level of aggregation along the period dimension.

Roll-up is the opposite of drill-down. Roll-up is the repetitive selection and analysis of summarized facts with each repetition of data selection occurring at a higher level of summarization.

MDA is performed through software which enables repetitive drill-down and roll-up of facts along varying combinations of dimensions.

3.9.7 Aggregation and Granularity

Aggregation is a key attribute of a data warehouse. Summarization and consolidation are other words that used to convey the same meaning. In the case of a multi-dimensional database, the summarizations are pre-computed for all the various combinations of the dimensions. This allows for very fast response to slice and dice operations at any level of drill-down, and also allows for fast drill-down and roll-up operations.

Granularity is a term that is used to describe the level below which no supporting details are stored, only the summaries. Good judgment needs to be exercised to determine granularity. If the granularity is set to be too fine, unused data will be stored, wasting processing time during replication steps, and wasting disk space to hold it. On the other hand, if the granularity is set too coarse, the detailed data will not be available if it is needed at some point in the future.

3.10 Multi-dimensional Models - Data Cubes and Star Schema

Traditional data analysis has usually been presented in the form of two-dimensional tables. For example, numbers of students classified by home state and course, or numbers of students by gender and year in course are two-dimensional.

Decision makers analyzing data today formulate more complex queries across multiple dimensions. A three dimensional model, for example, would help answer the question, "how many female students from Victoria are enrolled in a science degree program?" The concept of answering across multiple dimensions can readily be extended, and its value is immediately apparent.

In a three dimensional model, the intersection of the three axes is depicted. This intersection is called a fact. Some multi-dimensional vendors store facts in proprietary formats, often called multi-dimensional databases. Relational database vendors are adding OLAP capabilities to their products, use the relational database storage technology to hold multi-dimensional data in a star schema model or cube.[23]

3.10.1 The Data Cube Model

Representing data in a multidimensional structure makes it possible for business professionals to retrieve information and analyze it using terminology that they understand, and in ways that make sense to them.

In Figure 8, a three dimensional data cube is used to model student performance over three measures or dimensions - age group, gender and department/course. The dimensional axes hold the metrics to be analyzed. In this case it is student performance, represented as the number of classes completed with a satisfactory grade as a fraction of the number attempted. The views of the metrics are called *dimensions*. An individual student performance metric is given by the intersection of the three axes, and is referred to as a *fact*. In this particular representation, the facts are for a given year, aggregated over all students in that age group, gender and course. Cohort year and year in course are typically other dimensions in this model.

Slicing and dicing takes place along any of the three axes. This makes it possible to relate the effect of age group and course on student performance, or age and gender, or all three dimensions. Graphical representations make it easy to spot apparent relationships for further analysis.

The model contains all the precomputed aggregations, making response fast. For instance, Slice A contains values of student performance for both genders and all courses for the 20 to 29 age group. In slice B, the values for all age groups and all course for female students. This is what makes it possible to drill-down and roll-up.

[23] From IBM. *Decision Support Solutions: IBM's Strategy*. p10.

The analysis can start with Department (Business or Education), then drill down to actual courses.

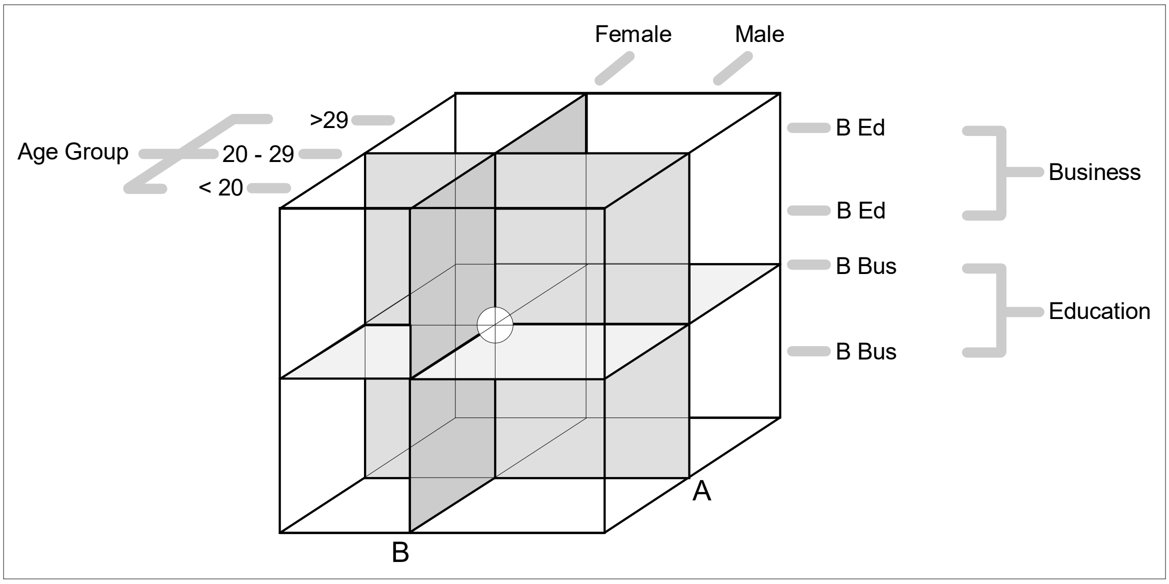


Figure 8. Data Cube Example

3.10.2 Star Schema

Traditional OLTP RDBMSs depend on schema definitions that focus on defining tables that map very efficiently to operational requests while minimizing contention for access to individual records. This maximizes concurrency while optimizing insert/update/delete performance.

The analytical processing performed on data warehouses places very different demands on the RDBMS. OLAP involves queries that are large, complex, ad hoc and data-intensive. A fundamentally different approach to defining the database schema is required.

A Red Brick White Paper entitled *STARjoin Technology*[24] illustrates this point very well with a simple purchase order example.

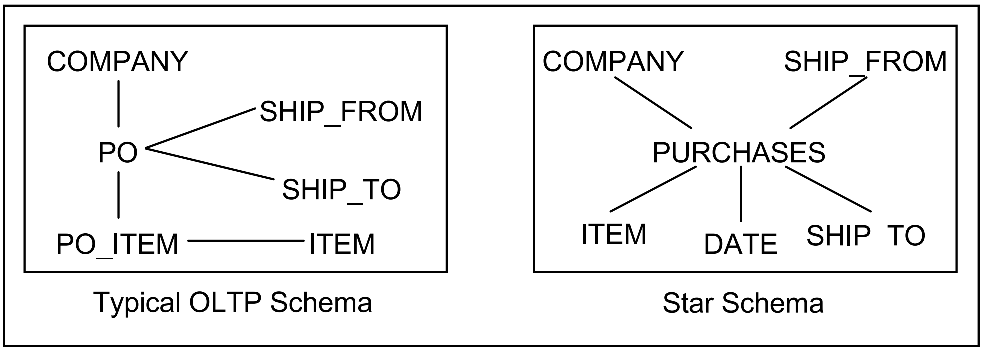


Figure 9. Star Schema Example

A query to list company name, cost of goods, source location and destination location would require the analyst to non-intuitively navigate the PO table in an OLTP database. Asking the same business question against the same data represented in a Star Schema is much more straight forward, because we are looking up specific facts (PURCHASES) through a set of dimensions (SHIP_FROM, SHIP_TO, ITEM). Because star schemas represent data intuitively, they are far better suited to data warehousing than traditional OLTP schemas. When querying an OLTP schema, analysts usually spend inordinate amounts of time navigating a maze of interrelated tables with cryptic names. When they query a star schema, they can go directly to the key facts, using a set of familiar dimensions.[25]

[24] Red Brick Systems. *Star Schemas and STARjoin Technology*. p2.
[25] Red Brick Systems. *Star Schemas and STARjoin Technology*. p2.

3.10.3 Evaluation Rules for Multi-dimensional Analysis Tools

In *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate* on page 18, Codd & Associates describe 12 rules for evaluating MDA tools. They are summarized in the following table, and described more fully in **Appendix C: *OLAP Product Evaluation Rules***.

1. Multi-Dimensional Conceptual View
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 Dimensions and Aggregation Levels

Table 9. Evaluation Rules for Multi-dimensional Analysis Tools[26]

3.11 Metadata

The simplest definition of the term *metadata* is "data about data". A data warehouse unlocks the data held in corporate databases but only if business users are able to find out about the data and information objects (queries, analyses, reports, etc) that are stored there. Such as facility is known as metadata or an information directory.

An information directory needs to hold more than the names of the tables with their elements and data types. As Figure 10 shows, it needs to hold information for technical tasks as well as for business tasks.

Much of the data needed by technical users will exist in a variety of places, such as program libraries, DBMS system catalogs, CASE tools, etc. Some of these places will include information of interest to business users as well, such as data descriptions and meanings. As Colin White points out, "one key objective of an information directory is to be able to integrate this diverse set of metadata, and then provide easy access to it for data warehouse developers, administrators, and business users."[27]

[26] Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* p18.

[27] White C. *Data Warehousing: The Role of the Information Directory*. p5.

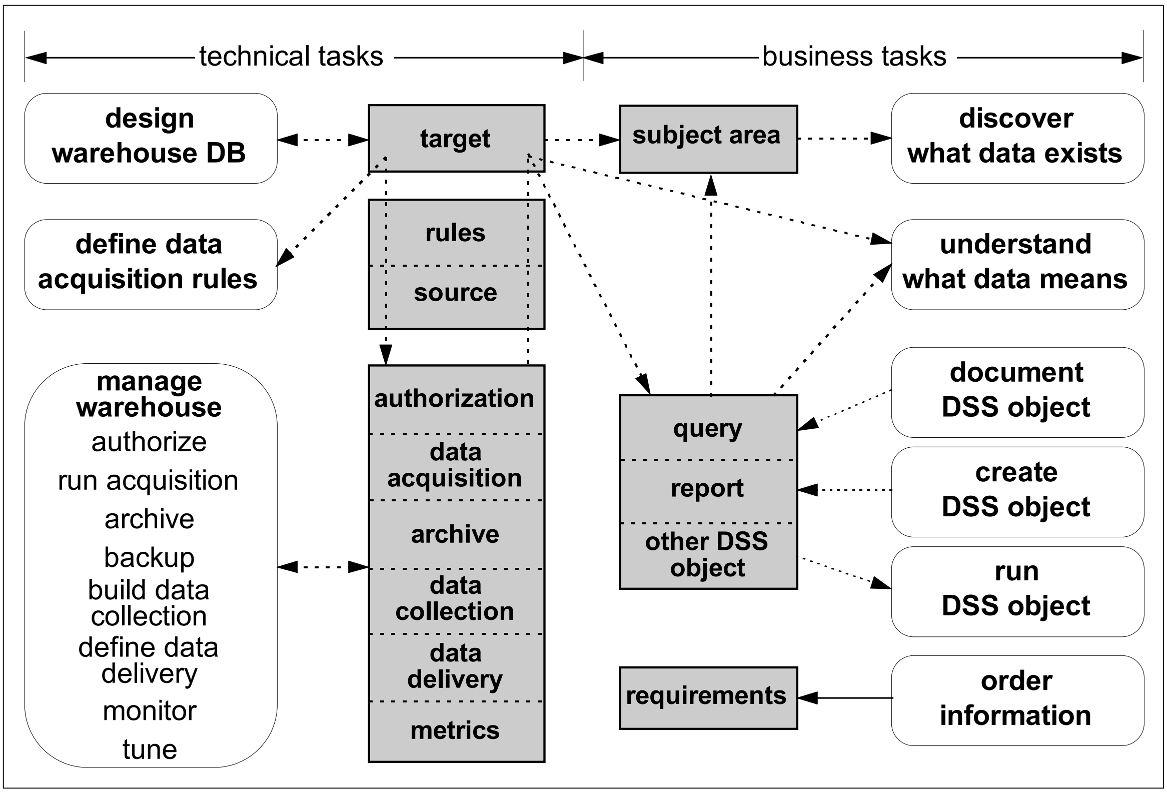


Figure 10. Tasks and Metadata[28]

The information directory holds information about the design of the data warehouse, including a history of changes that occur over time. It includes all the information necessary for administering the data warehouse - authorization, archiving, backups, building data collections, etc. It also specifies the data acquisition rules, including scheduling, sources, transformations and cleanups, etc. And it maintains a log of data collection as well as data access operations. An important component of a complete data warehousing solution is one that integrates and synchronizes the various sources of metadata.

3.12 Replication

The term *replication* is used to describe the process of managing copies of data, and has traditionally been applied in the field of distributed databases and in client/server environments. Its application in data warehousing is relatively recent and somewhat specialized, due in part to the aggregation and denormalization that occurs. Also, a data warehouse will contain multiple instances of the same set of data elements as snapshots, each with a different timestamp.

The more difficult replication problems in data warehousing occur in tables where the current values are merely to be updated if they have changed. In a research paper written at Stanford University, entitled *Efficient Snapshot Differential*

[28] White C. *Data Warehousing: The Role of the Information Directory*. p8.

Algorithms for Data Warehousing, the authors state that "detecting and extracting modifications from information sources is an integral part of data warehousing." They report that "there are essentially three ways to detect and extract modifications:

> 1. The application running on top of the source is altered to send the modifications to the warehouse.
> 2. A system log file is parsed to obtain the relevant modifications to the application.
> 3. The modifications are inferred by comparing a current source snapshot with an earlier one. We call the problem of detecting differences between two sources snapshots the *snapshot differential* problem; it is the problem we address in this paper."[29]

Replication tools are available for use with the major databases. However, Avondale's legacy systems are not accessible by any of these tools. Instead, we have written our own tools to extract, transform and load data into the data warehouse, using the Cognos 4GL PowerHouse. This is described further in Section 4.6.3 *Develop the Data Warehouse Load Routines*.

An important part of replication is to determine the schedule – what parts of the data warehouse get written or refreshed when? Some parts of the data warehouse will get updated with a regular frequency, such as monthly or weekly. Other parts will get new snapshots written according to a calendar of significant dates, such as around census and registration dates. The replication schedule needs to be automated, and documented in the information directory as well as in the system manual.

One of the benefits of data warehousing is consistency – analyzing the same data in different ways always accumulates to the same totals. This is not possible with operational systems since real-time updates to the data can take place between doing the first analysis and doing a later one. However, in planning the replication schedule, special care needs to be taken to ensure that the operational data is in a stable, non-volatile state during the replication process. Synchronizing the various parts of the data warehouse is important, or inconsistencies can still arise.

3.13 Maintainability

It is a fact of life that systems change, and as changes are made in the operational systems, they must be reflected in the data warehouse. The Stanford University Database Group is conducting research into making views self-maintainable in data warehouses. "A data warehouse stores materialized views over data from one or more sources in order to provide fast access to the integrated data, regardless of the availability of the data sources. Warehouse views need to be maintained in response to changes in the base data in the sources… A view is a derived relation defined in terms of base relations. A view is said to be materialized when it is stored in the database, rather than computed from the base relations in response to

[29] Labio W. and Garcia-Molina H.: Stanford University. *Efficient Snapshot Differential Algorithms for Data Warehousing*. p1.

queries."[30] The paper presents an algorithm for determining what are called *auxiliary views*. The objective is to find a minimal set of auxiliary views sufficient to maintain a view, and ultimately, to maintain a set of views.

In addition to the typical tasks associated with database maintenance, there are two additional factors to take into account when a data warehouse is involved:

- history
- metadata

One of the benefits of a data warehouse is that it is time-variant, allowing data to be analyzed over time. When a change occurs in the operational database, it will be necessary to propagate the change back through all the history records. This will often mean recomputing aggregations, but the detail will not always be available to do this. Even when it is available, it will not always make sense to recompute. For instance, if academic restructuring leads to a department of one school being moved to a different school, should school-based aggregations prior to the move be recomputed or should they stay the way they are?

Whatever choice is made in this case, it needs to be recorded in the metadata. This is the other complicating factor when maintaining a data dictionary - all changes need to be recorded in the information directory.

[30] Quass D., Gupta A., Mumick I. and Widom J.: Stanford University. *Making Views Self-Maintainable for Data Warehousing.* p1.

3.14 Commercially Available Data Warehousing Tools

3.14.1 The Data Warehousing Institute "Roadmap"

The Data Warehousing Institute has created what they call a "New Roadmap To Data Warehousing" in the form of a poster. It is a comprehensive although not exhaustive compilation of data warehousing components and tools organized into categories of use. The different categories are summarized in the following diagram.

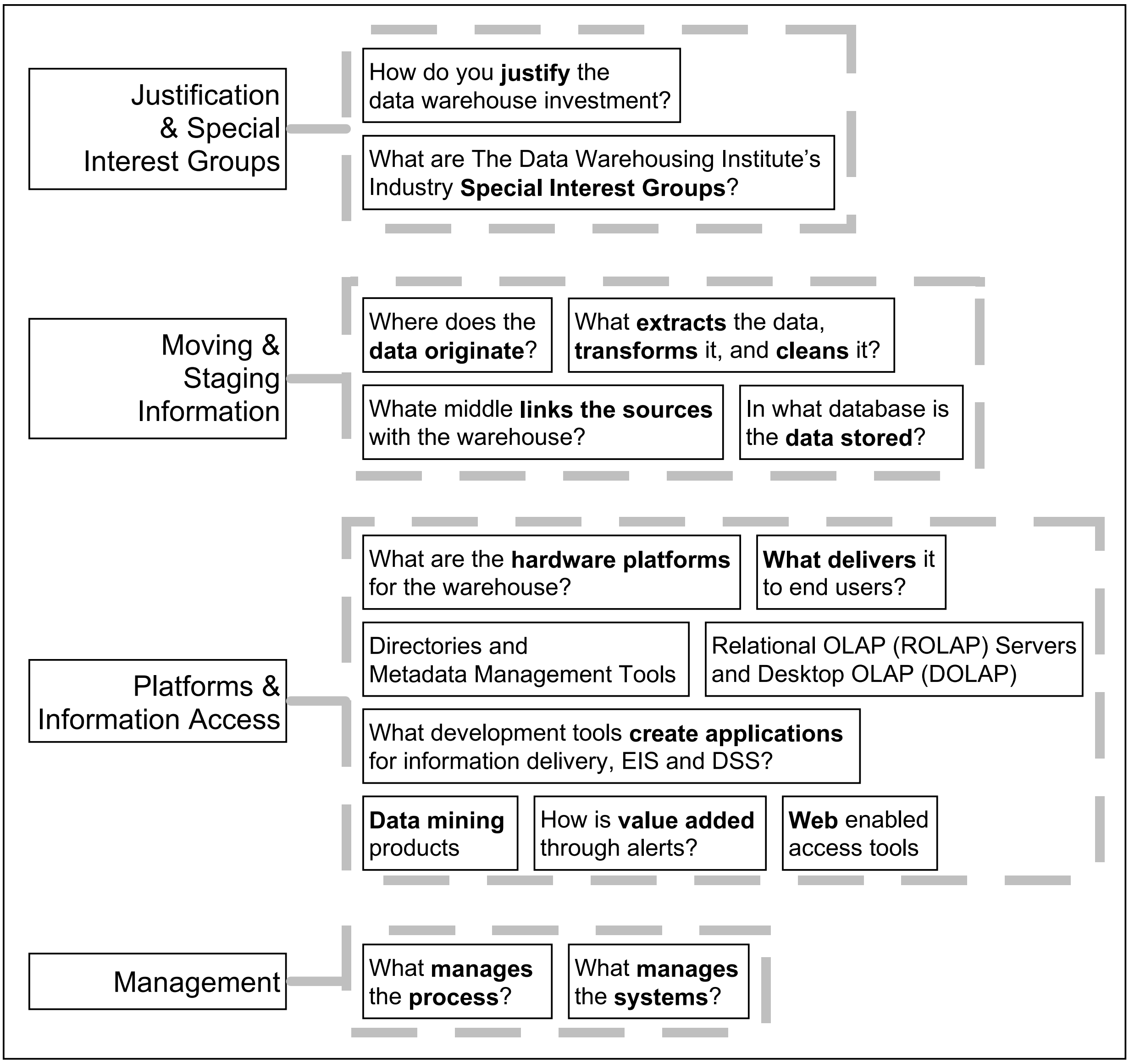

Justification & Special Interest Groups
How do you **justify** the data warehouse investment?
What are The Data Warehousing Institute's Industry **Special Interest Groups**?
Moving & Staging Information
Where does the **data originate**?
What **extracts** the data, **transforms** it, and **cleans** it?
Whate middle **links the sources** with the warehouse?
In what database is the **data stored**?
Platforms & Information Access
What are the **hardware platforms** for the warehouse?
What delivers it to end users?
Directories and Metadata Management Tools
Relational OLAP (ROLAP) Servers and Desktop OLAP (DOLAP)
What development tools **create applications** for information delivery, EIS and DSS?
Data mining products
How is **value added** through alerts?
Web enabled access tools
Management
What **manages** the **process**?
What **manages** the **systems**?

Figure 11. "New Roadmap To Data Warehousing"[31]

[31] The Data Warehousing Institute, 9158 Rothbury Drive, #200 Gaithersburg, MD 20879 USA

3.14.2 The "Data Warehousing Information Center"

Larry Greenfield has compiled and organized links on the world wide web to a comprehensive list of data warehousing resources, at URL:

http://pwp.starnetinc.com/larryg/index.html

There are listings for vendors of end user tools as well as for infrastructure technology, as shown in the following list of headings.

End User Tool Vendors

- Report and Query
- OLAP / Multidimensional Databases
- Financial, Marketing, and Supply Chain Analysis
- Executive Information Systems
- Data Mining
- Document Retrieval
- Geographic Information Systems
- Decision Analysis
- Statistics
- Process Modeling
- Information Filtering
- Industry Specific Tools
- Other End User Decision Support Tools

Infrastructure Technology Vendors

- Data Extraction, Cleaning, Loading
- Information Catalogs
- Databases for Data Warehousing
- Query and Load Accelerators
- Middleware
- Other Database Tools
- Hardware

4. Data Warehousing At Avondale

Contents:

This section describes what was done at Avondale College, how we went about it, the choices we made and why we made them Where our approach was different from common practice, the reasons are given. At times, general data warehousing principles are amplified here before describing the particular situation at Avondale College.

Issues or problem areas that we faced are highlighted. Many of these issues relate closely to the actual steps in the process of building a data warehouse. While such a list would vary in detail from one installation to another, there would be some commonality.

4.1 Decision Support Stages

An important element in creating a decision support environment is to break it down into its key components. Different sets of user tools are available for different aspects of decision support. Studies have shown that these tools fall into four categories:

1. Data Enquiry - basic query and reporting. The premise for ad hoc query is to verify some known or semi-known hypothesis.
2. Data Interpretation - more advanced analysis. The application provides statistical models for forecasting and trend analysis.

3. Multi-dimensional Analysis - These applications provide a multi-dimensional view of data presented in an aggregated form to demonstrate the inter-relationships of different dimensions.
4. Information Discovery - more popularly known as data mining, these applications have the capabilities to discover previously unknown information by examining a large amount of data.[32]

This is summarized in the following table:

Sophistication Level	**Computer-based tool**
data enquiry	traditional data processing applications
data interpretation	'point-and-click' reporting tools
multi-dimensional analysis	'slice and dice', 'drill-down' analytical tools
information discovery	intelligent agents

Table 10. Levels of Decision Support Tools

Each of the four levels of sophistication depicted in the following diagram corresponds to a different set of computer-based tools.

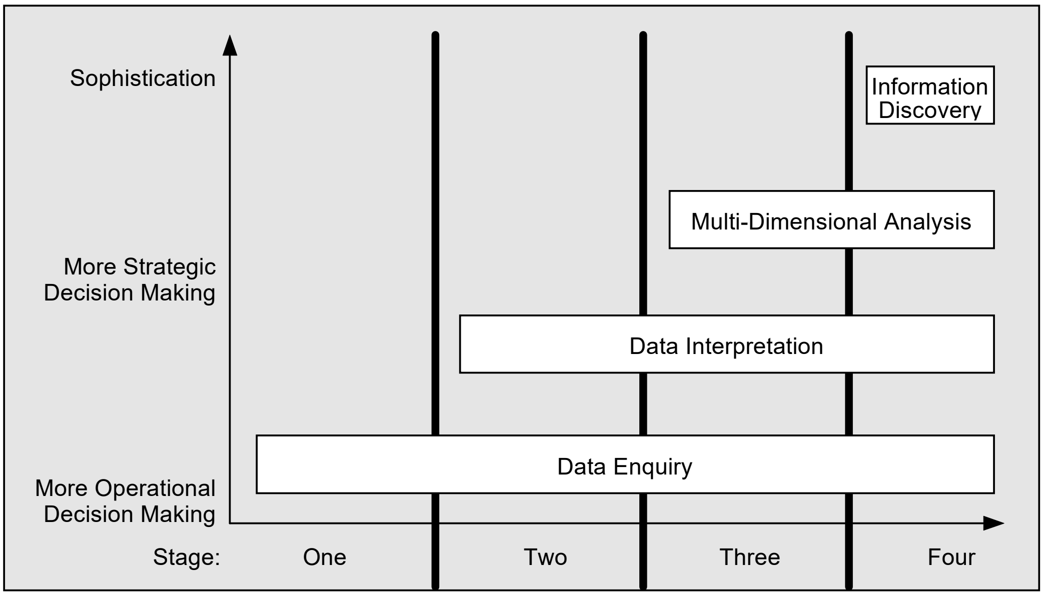

Sophistication
More Strategic Decision Making
More Operational Decision Making
Stage:
One
Two
Three
Four
Information Discovery
Multi-Dimensional Analysis
Data Interpretation
Data Enquiry

Figure 12. Decision Support Stages[33]

Avondale has been operating at Stage One in this representation. Canned reports are written by IT to support operational reporting and enquiry. The data warehouse will extend our decision support capabilities along to stages 2 and 3.

[32] IBM. *Decision Support Solutions: IBM's Strategy.* p5.
[33] IBM. *Decision Support Solutions: IBM's Strategy.* p5.

Users will be able to perform *data enquiry* using a ‘point and click’ reporting tool like Impromptu without having to wait for special programs to be written by IT. As further questions are raised in response to this, users will be able to refine their questioning with further enquiries, and will be better able to *interpret* the data.

Decision makers will be able to perform strategic analysis on *multi-dimensional* data models using business intelligence tools like PowerPlay.

It is important to note that as an organization moves through the stages of decision support, and achieves higher levels of sophistication in their use of data, the lower levels are not made redundant. There will always be a place for standard operational reports. Knowledge workers will always benefit from having easy-to-use reporting tools, and so on.

4.2 Establishing Needs

This is where a data warehousing project begins, but it proved to be very elusive at Avondale. Administrators found it very difficult to be precise about what they felt would enhance the quality of their decision making. We had to start with very general ideas, and then over a period of time and after several meetings, we refined the needs to a set of key areas for which we could provide or derive the data or dimensions that would be used in the data analysis.

For instance, after several rounds of meetings, we approached the Principal immediately after he had returned from 2 weeks vacation and asked him what information he would most dearly have wished he had to assist with the decisions he had to make on his first day back. The reality was that he was unable to describe to us something he had never had and had never seen.

For a Data Warehouse to be successful, it must satisfy the real needs of decision makers. These needs can be found by listing key topics or subject areas for each of the management areas. Key topics are those that have a critical effect on achieving the strategic plan. The strategic plan comes out of the corporate goals and objectives, which are determined by the mission statement.

This progression is summarized in the following checklist.

1. State Mission Statement
2. State Goals and Objectives
3. State Strategic Plan
4. List Management Areas
5. List Key Topics or Subject Areas
6. Choose High Impact Management Areas
7. Identify and Define Measures

Table 11. Checklist for Data Warehousing Success

Arriving at Step 7 (Identify and Define Measures) in the above table is also very important to the success of a data warehousing project. In this regard, the Goal Question Measure (GQM) approach can be very helpful. GQM starts with the goals and strategic plans. Factors critical to the success of those plans are identified. Questions are asked about the critical success factors and these are decomposed iteratively until the questions can be answered with measurements. These measures are described as Key Performance Indicators (KPIs).

Representatives from each of the management areas shown in Figure 1 (*Management Structure*) were invited to separate meetings where presentations were made describing the proposed Executive Information System (EIS) and Data Warehouse. At these meetings, the participants were interviewed to determine the most pressing needs of each group. We began by listing the key topics for each management area.

Management Area	**Key Topics**
Principal	• numerous ad hoc requests
Assistant Principal	• class size • teaching load/teaching capacity • staffing • course development • staff study projections • management reports
Business Manager	• budget forecasting and modeling • management reports • board proposals
Assistant Business Manager	• course costing • enrollment/fee modeling • financial planning • cafeteria/residence costing & fees • teaching load/staffing • fleet management
Director of Public Relations and College Development	• recruiting/applications from school leavers • alumni
Registrar	• class sizes • resource scheduling - classes, lecturers, students, rooms, buildings • staff loads • retention rates by course and other fields • pass/fail by course and other fields • changes in application and enrollment patterns • distribution patterns of TER's • various statistical summaries and analyses as required by the Principal

Table 12. Information requirements for specific management areas

On the basis of these meetings, a pilot group was chosen, consisting of the Principal, Assistant Principal and Registrar. Following successful implementation with the pilot group, the project will be extended to include the Business Manager, Marketing, and Public Relations and Development.

Having chosen a pilot group, the project leaders from Avondale's Computer Services Center met with the group to refine the specification of critical needs. A brainstorming technique know as a Joint Application Development (JAD) session was used to identify key attributes that could be measured.

At this JAD session, it became apparent that there were two sets of information access requirements, and later in the project, a third type of access requirement emerged.

User Type	Access Type
decision maker	drill down, slice and dice, multi-dimensional analysis
knowledge worker	point and click enquiry and reporting
casual enquirer	world wide web access

Table 13. Data Warehouse User and Access Types

Following the JAD session, critical factors that could be measured from data already being recorded were selected as dimensions to model for analysis. Some of the requirements were expressed a little vaguely at this point in the project, but meanings and definitions became clearer as the project progressed.

4.3 Possible Subject Areas

4.3.1 Retention Rates and Cohort Analysis

It was felt that it would be very useful to be able to analyze retention rates by a variety of factors, such as course, school and discipline, year in course, Tertiary Entrance Score, academic progress and performance, age, gender and other factors.

The initial attempts to specify how to measure retention rate met with difficulty in finding a precise definition of the term retention. One way to measure retention was to include all students who completed any course, even when the course was different from the one they started. An alternative was to consider only those students who completed the course they started as being retained.[34]

Contact with other Australian universities revealed the fact that some of these measurement issues had been studied elsewhere. A series of indicators of institutional performance had been proposed by the Australian Vice-Chancellor's Committee (Directors and Principals in Advanced Education (1988)) and Linke

[34] In the nomenclature used at Avondale College, a unit of study is called a **subject**. A program of study leading to an award is called a **course**.

(1991). The specific indicators of student performance and progress proposed in these two studies reveal a common thrust:

AVCC/ACDP	LINKE
Second year retention	Student progress rate
Major sequence retention	Program completion rate
Completion rate in minimum time	Mean completion time
Eventual completion rate	

Table 14. Indicators of Student Performance and Progress[35]

The key to finding a measurable definition of retention was to adopt the traditional concept of a cohort. Because of small class sizes, Avondale lecturers are often able to know personally all the students in their teaching area. This means that it is possible to have a reasonably good feel for the number of students dropping out of courses, changing to different courses, or resuming after an absence. However, competitive pressures are focusing much more pressure on some of these issues, and more accurate and more responsive measures are being sought. Hence, the concept of a cohort as a group of students commencing a course in any one year has been adopted.

A software package called COHORT was developed in 1991 by Bardsley at the Institutional Research office at Curtin University of Technology. It uses the standard DEET Student Submission files for Enrollment and Past Course Completions to monitor the progress of any commencing cohort of students. Using this terminology, the members of a cohort can be divided, at increments of one year, into three groups – completed, lapsed or continuing.

[35] Peter Manass & Robyn Peutherer, *Approaches to Student Progression Analysis Based On National Data Collections In Australia.* Australasian Association for Institutional Research Conference. Aair_pap.doc p2.

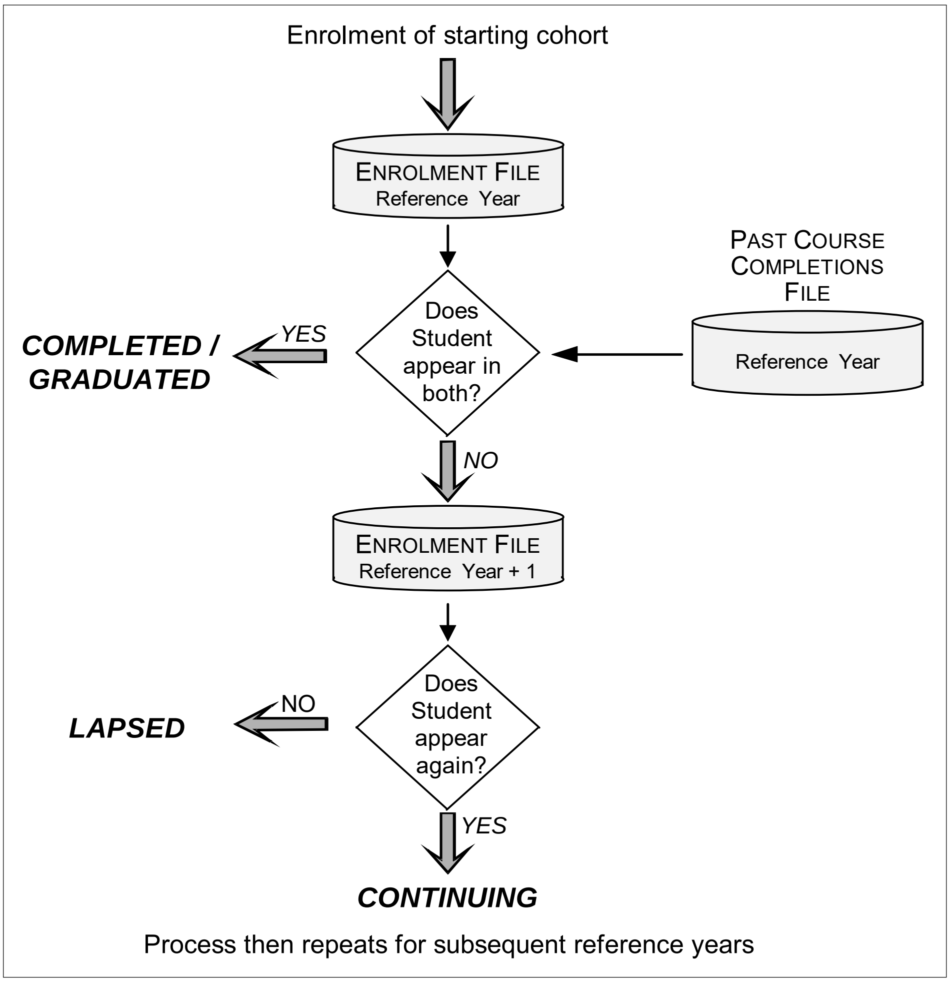

Enrolment of starting cohort
ENROLMENT FILE
Reference Year
PAST COURSE COMPLETIONS FILE
Reference Year
Does Student appear in both?
YES
COMPLETED / GRADUATED
NO
ENROLMENT FILE
Reference Year + 1
Does Student appear again?
NO
LAPSED
YES
CONTINUING
Process then repeats for subsequent reference years

Figure 13. DEET Files & COHORT Methodology[36]

4.3.2 Average time to graduate

Some of the indicators mentioned in Table 14 (*Indicators of Student Performance and Progress*) will be useful in gaining an understanding of how to measure average time to graduate. One area to resolve was what date to use as the starting point. There were several alternatives, all of which could be used to reveal useful information:

- date started the course being graduated from
- date started a degree course (any degree)
- date started any course at Avondale College

[36] Peter Manass & Robyn Peutherer, *Approaches to Student Progression Analysis Based On National Data Collections In Australia*. Australasian Association for Institutional Research Conference. Aair_pap.doc p6.

There are other questions still to be resolved.

- how to handle periods of time spent on study leave
- how to handle time spent earning credit transferred in from another institution or another course

4.3.3 Student Performance Indicators

The Student Progress Unit (SPU) was first proposed as a student performance indicator by Linke (1991), the first practical applications of its use were reported by Dobson & Sharma (1992).

$$\boxed{\text{STUDENT PROGRESS UNIT (SPU)} = \frac{\sum \text{EFTSU (where completion status = passed)}}{\sum \text{All EFTSU attempted}}}$$

Figure 14. Derivation of Student Progress Unit[37]

The element *completion status* can take one of four values:

- Withdrawn without penalty
- Failed
- Successfully completed (the value identified by SPU)
- Incomplete

In previous work using this indicator of student performance, data elements available in the DEET Student Load file and Student Enrollment file were used. While most analysis would be based on time units of whole academic years, we felt we needed to design the data warehouse to support analysis down to the level of semester or half year. Approximately 20 per cent of all graduates complete courses that start and finish in the middle of an academic year. Since our student enrollment and graduating class size are so small, we felt that this could distort some analysis based on whole academic years. Therefore we are using data from our internally developed student administration system rather than from the DEET files to compute SPU.

[37] Manass, P. & Peutherer, R. *Approaches to Student Progression Analysis Based On National Data Collections In Australia.* Australasian Association for Institutional Research Conference, November 1995. Aair_pap.doc p8.

4.3.4 Demographic Data

For many of the indicators that were being requested, there was a need to perform analyses based of various types of demographic data. One of these was the origin or source of students:

- for students applying from overseas - country of residence when they applied
- for students applying from within Australia - postcode of place of residence when they applied

One of the projected uses of this information was for followup on the effectiveness of promotional campaigns. However, the data to support this are not held in the required form in the current system. The data held as home and term address are dynamic, and are updated as a student moves from one address to another. A new data item needs to be created, and the data that are captured should be held permanently without further change.

For students coming from within Australia, the postcode could be used as a very fine indicator of location. This data is available in electronic form, and was obtained for use in this project. Maps of Australia showing postcodes have been used to group postcodes into regions of interest, since the analysis is usually by region. There can be multiple mapping of postcodes into regions, depending on the purpose of the analysis.

4.3.5 Enrollment Analysis

There are a number of attributes that can be used to yield useful information throughout the enrollment process.

- number of applicants
- number of acceptances – in total as well as by type of acceptance
    - full acceptance
    - provisional acceptance
    - not accepted, etc
- number accepted by not enrolled
- school where Year 12 was completed
- number from Seventh-day Adventist schools
- religion
- for Seventh-day Adventists, number baptized and not baptized

4.3.6 Other Indicators

The Avondale College Mission Statement states:

> "Avondale College is a community with a Christian world view committed to excellence in education, the development of the whole person with a love for life and learning, and the preparation of skilled professionals to serve in the workplace and in society."

Since Avondale College is a Christian institution, and spiritual development is an important part of its mission, some indicators of its effectiveness in this area were suggested. Religious affiliation and baptismal status are held in the current database, but they are updated dynamically, as changes occur. For this kind of analysis to be possible, additional data items need to be held to record this information. It needs to be recorded initially at time of enrollment, on leaving Avondale College, and later on as part of a graduate survey.

4.4 Dimensions for Pilot Project

With this background, our next task was to map goals to measurable indicators of performance. The first so-called measures that were proposed were very vague, and in fact were not capable of being measured. As our understanding of what we were trying to do developed, some indicators were proposed that were easy to measure and build into models, but were not particularly useful in assisting decision making. Others were found that were highly desirable and specific, but important parts of the data were not available.

In the end, we settled on measurable indicators of ***cohort retention rate*** and ***student performance***. Each of these were developed into PowerPlay models, with a range of relevant dimensions, including course, year in course, department, classes taken, class completion status, age, gender, and others.

The ***cohort*** analysis model has the following dimensions:

1. cohort year (year a group of students started the course)
2. analysis year (year being analyzed)
3. AOU (Academic Organizational Unit)
4. Course Type (a subtype of AOU)
5. Gender
6. Age Group
7. Age (a subtype of Age Group)
8. Course Status (Continuing, Completed, Lapsed)
9. SPU Annual (Student Progress Unit - see Section 4.3.3)
10. SPU Cumulative

Table 15. Cohort Analysis Model Dimensions

4.5 Data Warehousing Architectures

4.5.1 Simplified Data View

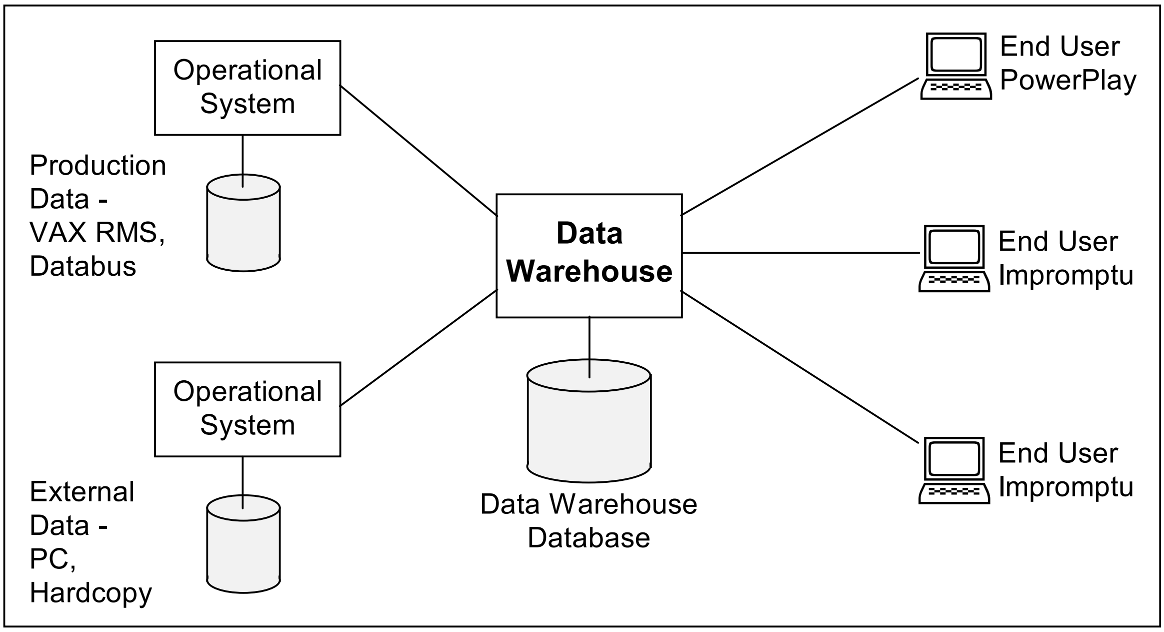


Figure 15. A Simplified Data View

This is a simplified view of how data from the various sources is taken into the data warehouse, and is then accessible to end users for reporting and analysis, and corresponds to a model we had proposed early in our project. We later learned that a database to support multi-dimensional analysis was quite different from one supporting knowledge workers writing their own *ad hoc* reports and enquiries.

A data warehouse suitable for multi-dimensional analysis is denormalized in several ways:

- coded data is replaced with values or meanings,
- the complex joins are done ahead of time and stored as tables

(this requires a good understanding of requirements, including anticipated queries, as well as the ability to adapt and evolve as the requirements develop)

- significant aggregation and summarization occurs

4.5.2 Data for Query and Analysis

The data warehouse we are building consists of two parts:

- there is an integrated relational view of all operational data, with history, for enquiry and reporting, which is sometimes referred to as an ***Operational Data Store***
- there are summarized and pre-joined data to support multi-dimensional analysis

It is not really important whether the "two parts" of this data warehouse are physically separate or only logically separate. In the pilot project at Avondale, they are physically separate, as in the following diagram.

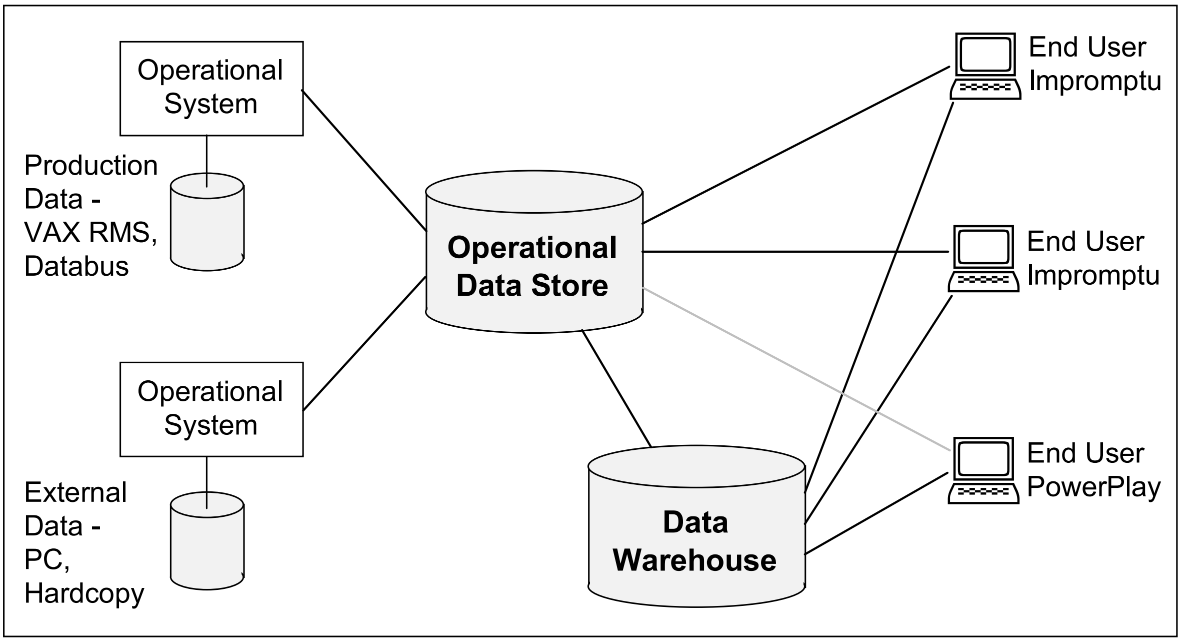


Figure 16. Relational View of Operational Data, With Data Warehouse

The Operational Data Store is a relational representation of the operational data, which is held in a variety systems, formats and platforms (see Table 1. *Information System Components and Platforms*). Procedures are executed every night that refresh the tables in this database from the production sources. In addition, this database holds detailed historical information - all data is indexed by year.

Using Impromptu, users are able to create their own queries and reports, directly accessing the data held in the Operational Data Store. Impromptu uses ODBC (Open DataBase Connectivity) drivers to access both of these relational databases.

The Data Warehouse contains only the data needed to support the multi-dimensional analysis models that have been developed in PowerPlay. The Transformer utility is used to convert this data into the proprietary multi-dimensional format used by PowerPlay. It is in this step that the aggregation and indexing occurs, that allows rapid analysis through slicing and dicing, and drill-down.

The following diagram is a variation on the previous two diagrams that illustrates the flow of data from the various operational systems to end-user data analysts:

- operational data source
- 2-dimensional relational database (rows and columns only)
- PowerPlay multidimensional cube
- end-user desktops

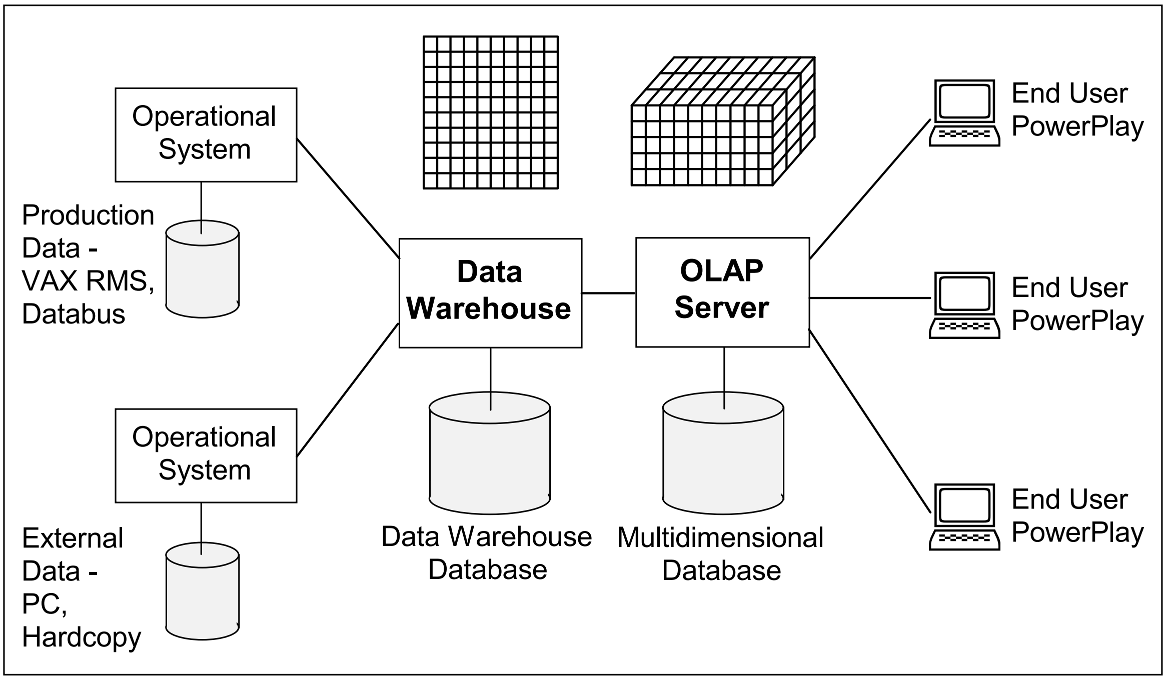


Figure 17. Architecture For Analysis

4.6 Data Warehouse Construction

4.6.1 Document Current System

Quite a bit of work had to be done as preparation for designing the data warehouse. Parts of the existing data structures were defined in one or both of the data dictionaries we were using. These two data dictionaries were the Data Repository in the data modeling tool we were using (Visible Analyst Workbench) and the PowerHouse Data Dictionary for the 4GL we were using. We had developed a number of useful reports based on the PowerHouse Data Dictionary so we decided to make that our primary source for data definitions.

We created definitions for all files, records and elements that were not already defined in the Data Dictionary. In addition, we added descriptive comments to all items defined in the dictionary. A relational database called an Operational Data

Store (ODS) was built from these definitions with very little changes being made. Samples of the reports produced from the data dictionary are shown in Appendix E: Data Dictionary - Files & Elements and Appendix G: Data Dictionary - Elements Alphabetically. As the sample reports show, the tables are fully normalized and contain many items that will never be used in data analysis or for decision making purposes. They are included in the ODS since it is a relational image of operational data, with history, for easy reporting via desktop reporting tools.

A high level Entity Relationship Diagram for the ODS is shown in Figure 18. In a typical data modeling project, the main objectives are to establish logical business data objects or entities, and to specify relationships between entities. Entity attributes are chosen that are consistent with the rules of normalization, with foreign keys to support the relationships, and a minimum of alternate keys to reduce processing overheads for add, update and delete transactions. However, since both the ODS and the Data Warehouse itself are used for read only operations, there is no need for foreign keys or referential integrity constraints. Removing them speeds up the data replication processes. However, in their place, it is common to find additional alternate keys to improve the performance of enquiries, as well as various levels of summarization.

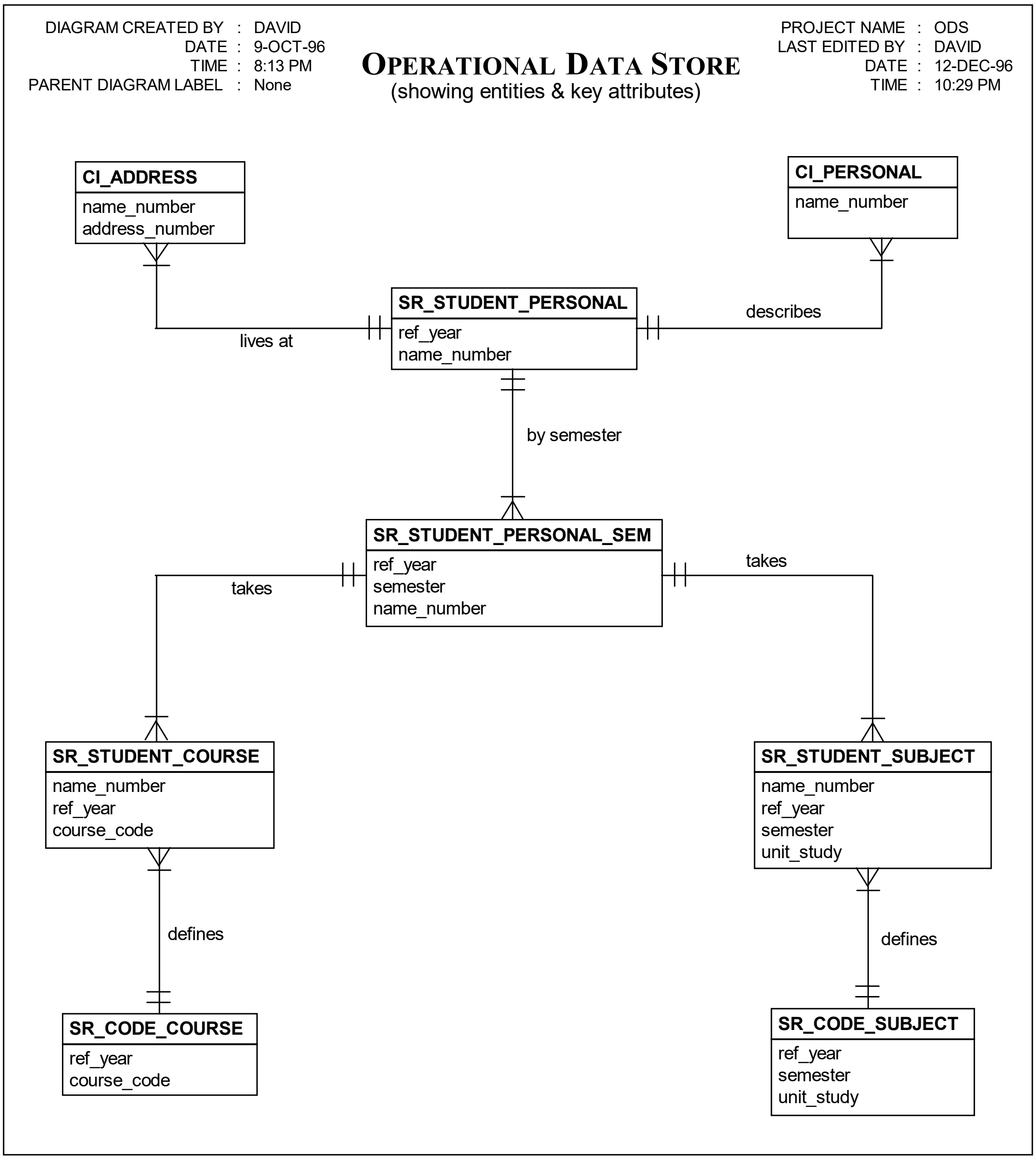


Figure 18. Entity Relationship Diagram for ODS

4.6.2 Design Data Warehouse for Pilot Project

The important thing to keep in mind when designing a data warehouse is to have a small number of well defined subject areas and to design a database that supports them. There will be numerous other possibilities for data to include in the data warehouse, but one of the most common reasons for failure in building a data warehouse is trying to start too big. The key is to choose a high impact subject area, with a powerful, committed sponsor, and go for a small, fast implementation. As the users become more sophisticated in their queries and analysis, and as the pilot project spreads to other users and subject areas, the database will pass through many iterations as it evolves.

The database that is used as the source for the PowerPlay cube has only three tables:

- `DW_STUDENT`
- `DW_SUBJECT`
- `DW_PP_COHORT`

The simple structure of this database is illustrated in Figure 19. The database definition for the DW database is given in Appendix F: Data Warehouse Schema Definition.

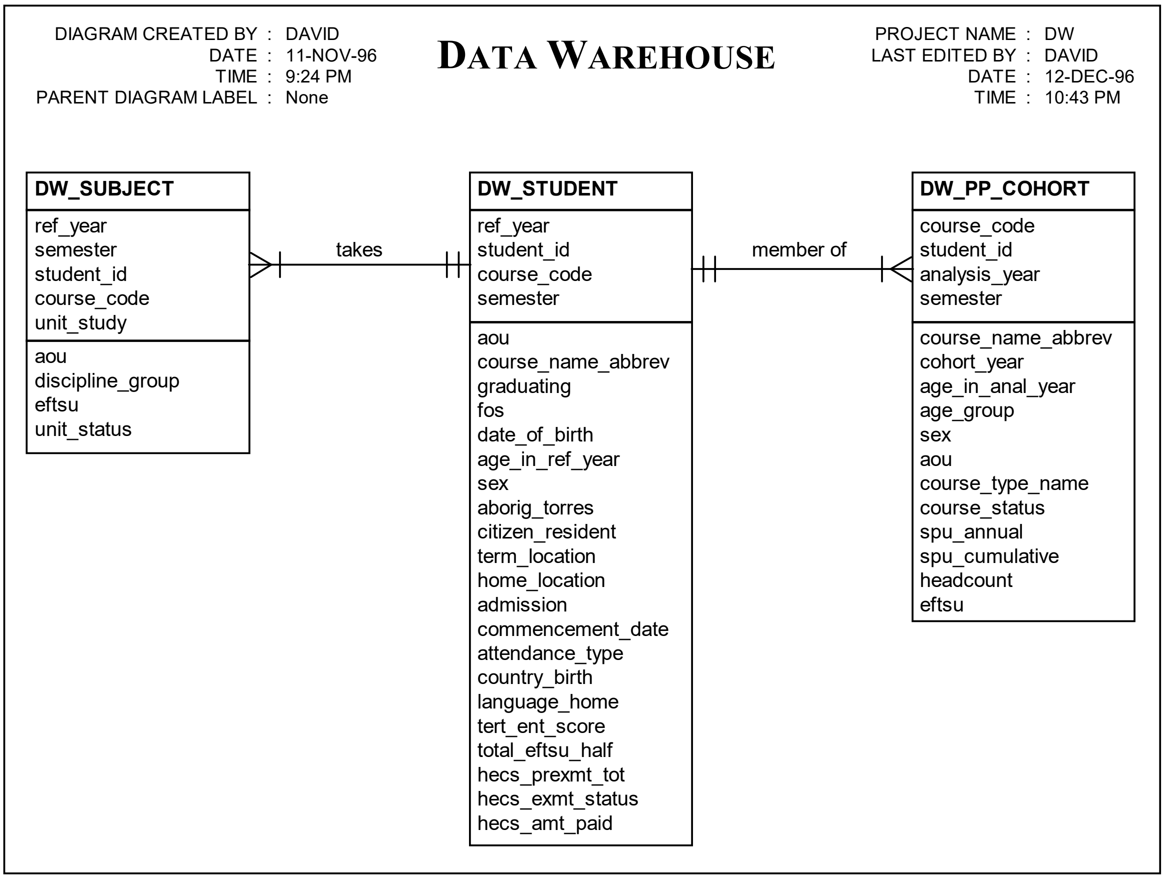


Figure 19. Entity Relationship Diagram for DW

4.6.3 Develop the Data Warehouse Load Routines

We used Oracle Rdb for the data warehouse, and the PowerHouse 4GL QTP from Cognos for extracting, transforming and loading the data. Since most of the administrative software we use has been written in-house, we had already completed the integration of name and address under a single person ID. QTP has proved to be quick and easy to write and maintain, and is powerful and efficient in its operation.

```
DISPLAY 'Updating SPU_ANNUAL fields'

ACCESS *DW_DATA:RAWSPUA ALIAS RAW_SPU

DEFINE D_SPU NUM = (T_PASS / T_ATTEMPT) * 100
DEFINE D_SPU_GROUP NUM = 10 IF D_SPU EQ 100                            &
                    ELSE  8 IF D_SPU GE  80                            &
                    ELSE  6 IF D_SPU GE  60                            &
                    ELSE  4 IF D_SPU GE  40                            &
                    ELSE  2 IF D_SPU GE  20                            &
                    ELSE  0

OUTPUT DW_PP_COHORT IN DW UPDATE ADD                                   &
  VIA COURSE_CODE, STUDENT_ID, ANALYSIS_YEAR, SEMESTER                 &
  USING SUBJECT_CODE OF RAW_SPU,                                       &
        NAME_NUMBER OF RAW_SPU,                                        &
        YEAR_YYYY OF RAW_SPU,                                          &
        SEMESTER OF RAW_SPU

  ITEM SPU_ANNUAL = D_SPU_GROUP
```

Figure 20. Sample QTP source code

The above sample of QTP code illustrates some of the features that are useful in a data warehouse load procedure.

- The `ACCESS` statement specifies the data source. This can be the output of a previous procedure, a single table, or a join across multiple tables.
- The `DEFINE` statement is used to derive new fields from fields in the data source.
- **`OUTPUT`** `<table>` **`UPDATE ADD`** is the powerful statement that efficiently inserts new records into the data warehouse and updates existing records if the data source has changed. No logic needs to be coded to read the old record and determine if a refresh is required. This means that the same program can be used to perform initial loads as well as performing the scheduled replications. In larger data warehouses, change detection becomes vital, to ensure that only altered records are refreshed.

The objectives in this step are as follows:

- select data needed to support analysis needs from data source
- extract source data and derive additional data fields where necessary
- perform summarization, discarding detail to accepted level of granularity
- perform data integration, decoding and denormalization

We were initially going to load data and analyze it by year, but because of small class sizes and courses starting in second semester as well as first semester, we decided we needed to load data by semester. The code for doing the load by semester was not completed when I moved away from Avondale College. In its present form, each student has records created unconditionally in both semesters. This creates unwanted records for students who start in second semester, or who finish in first semester.

All of the data in both the Operational Data Store (ODS) and the Data Warehouse (DW) is recoverable from production data and archive directories. The code that loads the semester data needs to be completed so that it has the proper logic for dealing with semesters. Then the ODS and DW can be deleted, regenerated, and the data can be reloaded.

4.6.4 Establish Granularity and a Replication Schedule

A replication schedule needs to be determined, put in place and automated. We have nothing sophisticated for achieving this at present, but we have an automated procedure for refreshing the Operational Data Store on a daily basis. We are simply using batch jobs that resubmit themselves appropriately. This is working quite satisfactorily for our size of data warehouse (5-10MB), which is really only in the small data mart size category.

```
$ submit daily.com -
        /restart -
        /queue=ACVSA_SLOW$BATCH -
        /after="TOMORROW+00:05:00" -
        /log=od_prog:daily.log
$ submit CI_ALL.COM /queue=acvsa$batch /log=od_prog:CI_ALL.log
```

Figure 21. Sample JCL for scheduling - `daily.com`

4.6.5 Develop MDD Models

In the debate over MDD versus ROLAP, for our size of data warehouse, MDD provides all the functionality and performance that we can use. The data warehouse we have built is used by the Cognos Transformer tool to populate the multidimensional database used by PowerPlay.

Most of the preliminary work to build the cohort analysis and student performance tracking models has been completed. The unwanted semester records, described in the previous section, prevented the models from becoming operational. As soon as the coding and data reload described in the previous section are completed, the models can be completed and put into operation.

The dimensions for the cohort analysis model are as listed in Table 15.

4.6.6 Security

Fortunately, many of today's database and business intelligence desktop tools include comprehensive security provisions. Access rights can be granted down to the level of data items and even data values, with templates for different roles and sets of users..

However, designing a security strategy that all parties will agree to, then implementing and maintaining it, can be quite a task.

4.7 Implementation Issues

4.7.1 Data Ownership and Responsibility

At Avondale, name and address information had traditionally been maintained separately in the finance, the academic and the alumni records. We had actually achieved some integration of this data in the production systems as a preliminary step to setting up the data warehouse. When alumni mailouts turned up surname or address corrections, there was some initial conflict between the academic office and the alumni office over who owned the data and therefore who had the right to update it.

The reverse problem also occurred in some instances. Once some of the initial problems of data ownership were resolved and users became accustomed to the concept of distributed maintenance of some of the data, it became evident that we needed to identify "data custodians" to be responsible for different portions of the data. This was important for maintaining standards in the operational systems for data entry and update procedures.

4.7.2 Data Integration and Cleanup

A range of problems were found on trying to integrate data from various sources into the data warehouse, as summarized in the following table.

1. semantics
2. integrating mismatched data types
3. integrating mismatched coded values
4. invalid typed data
5. orphan child data entries
6. duplicated IDs
7. out of date surnames and addresses
8. data restructuring for array items, etc

Table 16. Data Integration & Cleanup Issues

Some fields were known by different names or had different data types in different systems, or were represented with different sets of coded values. Integrating these proved to be quite a challenge. **STUDENT_ID** is one example of this.

```
STUDENT_ID                    PIC 9(10).
STUDENT_NUMBER                PIC 9(5).
NAME_NUMBER                   PIC X(6).
  (NAME_NUMBER replaces STUDENT_NUMBER plus CHECK_DIGIT)
```

Table 17. Semantic Differences in STUDENT ID

Other examples of elements with semantic or data type differences include:

ACCOUNT	account, student_account, gl_account
YEAR	subject_year, acad_year, ref_year
SEMESTER	subject_semester, current_semester, requested_semester, submission

Some fields contained data that was invalid for the data type, such as date fields with invalid dates or numeric fields containing non-numeric data. These data errors were not detected in indexed file system, but were rejected by the RDBMS when being loaded into the relational ODS or DW.

In some of the earlier data models that we built, we defined relationships between tables with foreign keys. This highlighted other types of errors in the source data, since it did not obey the referential integrity rules we were trying to impose. We corrected these errors in the source data. They occurred in such cases as where a student had been enrolled in a class which was subsequently canceled and deleted from the classes offered file. The student enrollment remained, making it an orphan record.

The most difficult data cleanup problem was when duplicate records occurred for the same person, with different IDs. This happened even within the same system. Correcting all the records with the wrong ID was a time-consuming process, and required special care and concentration to ensure that it was done correctly. It also had to be repeated in each archive directory where the same errors were found, and in the DEET submission files. Even then, in many cases it was not possible to determine if two possibly duplicate records where really for two different people or not.

Another problem with the validity of the data was out of date surnames and addresses. This is a perpetual problem and has no easy solution.

Some restructuring of the data was necessary in going from the operational systems to the ODS and DW. For instance, the year had to be added to the key where it was not already present to allow for holding of historical information. There were several cases were COBOL-style group items were used. These had to be eliminated. Also, it was necessary to handle information held by semester, which was indicated by a suffix in field name or a subscripted array. These fields were properly denormalized.

A major data and program conversion project had been completed prior to the commencement of the data warehouse project. This was to convert the 5-digit `NAME_NUMBER` and `CHECK_DIGIT` fields to a single 6-character `NAME_NUMBER` field.

5. Pilot Review & Conclusions

Contents:

5.1 Review & Conclusions

I should point out again that at the time of writing this report, the project is languishing somewhat, due in part to the Project Leader (the author) being transferred overseas. Although I have moved to a new country and have had to learn a new job, my involvement in the Avondale Data Warehouse Project has continued from a distance. The other member of the technical team was appointed as my successor at Avondale and has also had to learn a new job, so the project understandably has lost a bit of momentum.

In writing this report, we had hoped to be able to make some observations on the impact of the project on decision making processes at Avondale, but since it has not yet been implemented, that is not possible.

Nevertheless, we believe we can claim some credit for helping to focus attention on planning and decision making at Avondale. It is clear that the models chosen for the pilot project will provide vital strategic information, and other areas are being considered for extending the project. In addition, knowledge workers will begin to get answers to questions they had given up asking because of the IT backlog.

In Chapter 10. (A Data Warehouse Design Review Checklist) of Bill Inmon's book *Building the Data Warehouse*, the author states that "the attendees at the design review include anyone who has a stake in the development, operation, or usage of the DSS subject area being reviewed."[38] This normally includes the following, and of this group, the most important attendees are the end users and the DSS analysts.

- Data Administrator (DA)
- Database Administrator (DBA)
- programmers
- DSS analysts
- other end users
- operations

[38] Inmon, W.H. *Building the Data Warehouse.* p295

- systems support
- auditing
- management

A review questionnaire was distributed to members of the pilot project group, and is reported in full along with responses (in quoted italics) in **Appendix H: *Review Questionnaire***. The quoted responses are from a senior administrator. Parts of the survey and responses are included here for comment.

1. The EIS/Data Warehousing concept

1. Were the concepts sufficiently well presented?

"I thought concepts were reasonably well presented."

2. How did management regard the importance of the project?

"It was probably not regarded with sufficient importance by Senior Management. The middle levels would probably be more involved in the actual use of the system - at least if we had a broader middle level. We are too small."

3. What level of commitment did management give to the project?

Part 2 in the above response suggests that the role of a data warehouse in supporting management decision making was not fully appreciated by Senior Management. The potential benefit of PowerPlay in data analysis and decision support was overshadowed by the very attractive gains offered by Impromptu as an enquiry and reporting tool for middle management. The shortage of technical resources and the lack of adequate management commitment made it difficult to keep the project focused on "management improvement", which was an integral part of its initial justification.

4. Training

1. Was training relevant, appropriate, timely, useful?

"Training was unfortunately irrelevant and at the wrong time. I think the day we spent trying to learn the software was wasted. This part of the exercise should have come after the data warehouse was complete and in place so that we could get an introduction to its use and then move straight into its use."

The training that was delivered by the consulting firm was indeed irrelevant, although if the implementation had proceeded according to schedule, the timing would have been appropriate. The trainer who was sent to provide the PowerPlay training was very new in the job and had a superb technical understanding of the Cognos programming languages, but lacked an understanding of what academic administrators might need to support their decision making. He was able to describe what each menu and command button did, but did not show the users how the product could be used to solve their business problems.

5. Consulting

1. Please comment on your impressions value of the consultants to the project, from your own experiences with them.

"I thought the consulting process was adequate but we probably did not have a clear strategic plan in place to help us zero in on the critical performance indicators. We are still wrestling with the problem of trying to develop a strategic plan."

Late in 1996, Avondale College went through a Strategic Planning exercise with a consultant. Lack of a clear plan may have slowed the initial progress of the project, but the subject areas that were chosen were regarded as valuable and achievable, and quite appropriate for a pilot project. The Strategic Planning that has taking place since the project started should prove invaluable for extending the project, once the pilot is implemented.

No comments could be made on many parts of the questionnaire since the project did not proceed far enough. When time and resources are available and the project can be restarted, it is clear that the training will need to be repeated.

5.2 Recommendations

5.2.1 Complete the Pilot Project

The immediate tasks which I have been able to identify for finishing the pilot project are as follows:

1. complete coding to handle data load by semester
2. complete creation of PowerPlay models
3. determine replication frequency for each subject area
4. repeat the PowerPlay training
5. deliver the Impromptu training

5.2.2 Extend to Other Subject Areas

A number of other possible subject areas were identified earlier in the project and are reported in Section 4.3 of this report. In addition to retention rates, cohort analysis and student performance indicators, there were:

- average time to graduate
- demographics
- enrollment analysis
- other indicators

After the recent Strategic Planning exercise conducted at Avondale College in late 1996, it would be useful to repeat the work reported in Section 4.2 on **Establishing Needs**. This could be expected to reveal new information requirements and priorities.

5.3 The Future of Data Warehousing

5.3.1 Integrated Tools

Currently, the tools for data warehousing as categorized on The Data Warehousing Institutes "New Roadmap To Data Warehousing" poster are only loosely connected or not connected at all. Single vendor solutions are limited in one way or another in terms of a total data warehousing solution. This almost always means tools from a variety of vendors must be used. Data warehousing standards are needed for things such as:

- extracting, transforming and cleansing source data
- automating data warehouse maintenance
- metadata management
- process management and scheduling

5.3.2 The World Wide Web

Application development for the web is seen as having great potential as a truly platform-independent environment. Web browsers are becoming the environments in which people work, and is an avenue for extending the reach of decision support front-ends to existing client software. The data warehousing environment of a read-only database on a separate hardware platform lends itself well to Internet access. New products are appearing at a rapid rate that simplify linking data to web sites. The advent of Java offers an alternative to, or at least variation on, client-server processing for the masses.

For a review of the current state of data warehousing on the web, see the Technology Viewpoint article published by Aberdeen Group. It is available on the web at http://www.aberdeen.com/secure/viewpnts/v9n6/v9n6.htm.

5.3.3 OLAP constructs in RDBMS

A relational database designed for OLTP will not serve well as a database for data analysis. Optimization techniques such as aggregating fact tables, partitioning fact tables, and denormalizing relation tables all provide significant improvements in performance. However, the RDBMS itself and especially its optimizing algorithm need to do things differently for OLAP.

Some of the deficiencies in existing OLTP are described in an article by Neil Raden that was published in Information Week (March 18, 1996, Issue: 571, Section: OpenLabs).

> "Systems designed for transaction processing are overwhelmed by the volume of data [involved in data warehousing]. A data warehouse also needs special tools that eliminate the extraneous overhead of transaction logging, rollback/commit, and incremental referential integrity checking. In addition, a common approach in data warehousing is to store aggregated information in the database, avoiding slow and costly 'group by' operations by individual

> queries. Some critical components to look for in a DBMS are **auto-aggregation** at load time, **aggregation scheme advisers**, and **aggregate navigators** to direct the query optimizer to always use the highest level of aggregation available.
>
> "To facilitate the types of queries common in decision support, the indexes in data warehousing are larger and more complex. Rebuilding these indexes can be time-consuming, resulting in update cycles that are too long. To get the best performance from a data warehouse, look for DBMSs that have **new index types** (bitmaps, join indexes) and **index building processes**, including incremental indexing–using the existing index rather than the base tables as a starting point–and parallelizing the index processing."[39]

5.3.4 No Future Without Data Warehousing

The world of higher education as well as business in general is becoming increasingly competitive. Those institutions and businesses that realize the potential benefit of the information resource first will gain a competitive advantage. As stated in the closing statement of the White Paper by E. F. Codd & Associates entitled *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate*, "The quality of strategic business decisions made as a result of OLAP is significantly higher and more timely than those made traditionally. Ultimately, an enterprise's ability to compete successfully and to grow and prosper will be in direct correlation to the quality, efficiency, effectiveness and pervasiveness of its OLAP capability. It is, therefore, incumbent upon IT organizations within enterprises of all sizes, to prepare for and to provide rigorous OLAP support for their organizations".[40]

[39] Raden, N. Technology Tutorial, Part 1 – Maximizing Your Warehouse. Information Week. March 18, 1996.

[40] Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* p31.

Bibliography

References to books, papers and journals

Australian Vice-Chancellors' Committee. *Report of the AVCC/ACDP Working Party on Performance Indicators*. 1988.

Bardsley, N. Student Tracking using COHORT v2.0. *Paper issued at Third Annual Conference of the Australasian Association for Institutional Research*. Auckland, New Zealand, November 1992.

Inmon, W.H. *Building the Data Warehouse*. John Wiley & Sons, Inc. 1993.

Kimball, R. *The Data Warehouse Toolkit*. John Wiley & Sons, Inc. 1996.

Linke, R. *Performance Indicators in Higher Education*. AGPS Canberra. 1991.

Manass, P. University of Technology, Sydney. *Australian Universities - Degrees By PowerPlay*. Cog_nauc.doc

Manass, P. University of Technology, Sydney. *Student Cohort Tracking With PowerPlay*. Visions.doc

Manass, P. & Peutherer, R. *Approaches to Student Progression Analysis Based On National Data Collections In Australia*. Australasian Association for Institutional Research Conference, November 1995. Aair_pap.doc.

Raden, N. Modelling a Data Warehouse. Information Week. January 29, 1996.

Raden, N. Modelling a Data Warehouse. Information Week. March 18, 1996.

References to material available on-line on the World Wide Web

The style described by Michael B. Quinion is used for citing online sources, with the variations of showing the title in italics.

Articles

Aberdeen Group, Inc.
Data Warehouse Query Tools: Evolving to Relational OLAP
http://www.aberdeen.com/secure/viewpnts/v8n8/v8n8.htm
Last updated: May 13, 1996. Accessed: December 9, 1996.

Aberdeen Group, Inc.
Web Warehouses: DSS For The Masses
http://www.aberdeen.com/secure/viewpnts/v9n6/v9n6.htm
Last updated: May 13, 1996. Accessed: December 9, 1996.

Arbor Software
An Enterprise Solution for Financial Planning & Analysis
http://www.arborsoft.com/essbase/wht_ppr/fpandaTOC.html
Last updated: December 6, 1996. Accessed: December 9, 1996.

Arbor Software
Multidimensional Analysis: Converting Corporate Data into Strategic Information
http://www.arborsoft.com/papers/multiTOC.html
Last updated: September 21, 1995. Accessed: October 8, 1996.

Arbor Software
Relational OLAP: Expectations & Reality
http://www.arborsoft.com/essbase/wht_ppr/rolapTOC.html
Last updated: December 6, 1996. Accessed: December 9, 1996.

Arbor Software
Sales and Marketing - Planning and Analysis for the Enterprise
http://www.arborsoft.com/essbase/wht_ppr/sandmTOC.html
Last updated: December 6, 1996. Accessed: December 9, 1996.

Arbor Software
The Role of the Multidimensional Database in a Data Warehousing Solution
http://www.arborsoft.com/essbase/wht_ppr/wareTOC.html
Last updated: December 6, 1996. Accessed: December 9, 1996.

Boar B.: NCR Corporation
Understanding Data Warehousing Strategically
http://www.tekptnr.com/tpi/tdwi/review/bboar1.htm
Last updated: June 14, 1996. Accessed: December 9, 1996.

Brown T.C.: Action Inquiry Network (ActNet)
Action Inquiry, Problem Types
http://enhanced-designs.com/actnet/problem.htm
Last updated: December 5, 1996. Accessed: December 9, 1996.

Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates
Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate
http://www.arborsoft.com/essbase/wht_ppr/coddTOC.html
Last updated: December 6, 1996. Accessed: December 9, 1996.

Creative Data, Inc.
Data Warehouse Terminology
http://www.std.com/CreativeData/credata/termin.html
Last updated: November 20, 1996. Accessed: December 9, 1996.

Enterprise Solutions, Inc.
Data Warehouse Program Initiation
http://www.infocat.com/dwplan1.htm
Last updated: April 6, 1996. Accessed: December 9, 1996.

Finkelstein R.: Performance Computing, Inc.
Understanding the Need for On-Line Analytical Servers
http://www.arborsoft.com/essbase/wht_ppr/finkTOC.html
Last updated: December 6, 1996. Accessed: December 9, 1996.

Gupta A., Harionarayan V. and Quass D.: Stanford University
Aggregate-Query Processing in Data Warehousing Environments
http://www-db.stanford.edu/pub/papers/vldb.ps
Last updated: June 27, 1996. Accessed: December 9, 1996.

Hammer J., Garcia-Molina H., Widom J., Labio W. and Zhuge Y.: Stanford University
The Stanford Data Warehousing Project
http://www-db.stanford.edu/pub/papers/whips-overview.ps
Last updated: July 1, 1996. Accessed: December 9, 1996.

IBM
Data Replication: Data on the Go for Businesses on the Move
http://www.software.ibm.com/data/dbtools/drswp.ps
Last updated: January 2, 1996. Accessed: December 9, 1996.

IBM
Decision Support Solutions: IBM's Strategy
http://www.software.ibm.com/data/dbtools/dssstrat.ps
Last updated: August 13, 1996. Accessed: December 9, 1996.

IBM
Multi-Dimensional Analysis: Extending the Information Warehouse Framework
http://www.software.ibm.com/data/dbtools/mdaandiw.ps
Last updated: May 16, 1996. Accessed: December 9, 1996.

IBM
The IBM Information Warehouse Solution: A Date Warehouse Plus!
http://www.software.ibm.com/data/dbtools/iwsolu.ps
Last updated: December 6, 1995. Accessed: December 9, 1996.

Information Advantage
Multi-Level Security
http://www.infoadvan.com/1b_5ts.4.html
Last updated: November 26, 1996. Accessed: December 9, 1996.

Informix
Data Warehousing with Informix
http://www.informix.com/informix/solution/warehous/intro.htm
Last updated: . Accessed: December 9, 1996.

Kimball R.: DBMS Magazine Online - Book Excerpt
The Data Warehouse Toolkit - Practical Techniques for Building Dimensional Data Warehouses
http://www.dbmsmag.com/dwtlkit.html
Last updated: November 9, 1996. Accessed: December 9, 1996.

Labio W. and Garcia-Molina H.: Stanford University
Efficient Snapshot Differential Algorithms for Data Warehousing
http://www-db.stanford.edu/pub/papers/window-short.ps
Last updated: June 26, 1996. Accessed: December 9, 1996.

Mattison R.: CIO Magazine
State of the Art - Warehousing Wherewithall
http://www.cio.com/CIO/040196_soa.html
Last updated: . Accessed: December 9, 1996.

McGuff F.: Frank McGuff
Task Index
http://members.aol.com/fmcguff/spiral/FULLNDX.HTM
Last updated: October 29, 1996. Accessed: December 9, 1996.

McGuff F.: Frank McGuff
Translating user requirements to product requirements
http://members.aol.com/fmcguff/userreqs/userreqs.htm
Last updated: October 18, 1996. Accessed: December 9, 1996.

MicroStrategy, Inc.
Relational OLAP: An Enterprise-Wide Data Delivery Architecture
http://www.strategy.com/wp_a_i1.htm
Last updated: November 19, 1996. Accessed: December 9, 1996.

MicroStrategy, Inc.
The Case For Relational OLAP
http://www.strategy.com/dwf/wp_b_a1.htm
Last updated: November 19, 1996. Accessed: December 9, 1996.

Oracle
Oracle OLAP Products: Adding Value to the Data Warehouse. An Oracle White Paper
http://www.oracle.com/products/olap/collatrl/olapwp.pdf
Last updated: September 1, 1995. Accessed: December 9, 1996.

Paller A.: The Data Warehousing Institute
TDWI Issues Online: Top Six Trends for 1996
http://www.tekptnr.com/tpi/tdwi/issues/top6trnd.htm
Last updated: June 14, 1996. Accessed: December 9, 1996.

Perkins A.: Information Engineering Systems Corporation
Developing a Data Warehouse - The IES Approach
http://www.ozemail.com.au/~ieinfo/dw.htm
Last updated: October 20, 1996. Accessed: December 9, 1996.

PLATINUM technology, inc.
Data Warehousing: PLATINUM technology Approach
http://www.platinum.com/products/gleason.htm
Last updated: . Accessed: December 9, 1996.

Quass D., Gupta A., Mumick I. and Widom J.: Stanford University
Making Views Self-Maintainable for Data Warehousing
http://www-db.stanford.edu/pub/papers/self-maint.ps
Last updated: September 27, 1996. Accessed: December 9, 1996.

Quinion M.B.
Citing online sources
http://clever.net/quinion/words/citation.htm
Last updated: September 20, 1996. Accessed: December 9, 1996.

Raden N.: Archer Decision Sciences, Inc.
Data, Data Everywhere
http://netmar.com/~nraden/iw_mct01.htm
Last updated: February 20, 1996. Accessed: December 9, 1996.

Raden N.: Archer Decision Sciences, Inc.
Modeling the Data Warehouse
http://netmar.com/~nraden/iw0196_1.htm
Last updated: February 20, 1996. Accessed: December 9, 1996.

Raden N.: Archer Decision Sciences, Inc.
Star Schema 101 , Table of Contents
http://netmar.com/~nraden/str101.htm
Last updated: February 20, 1996. Accessed: December 9, 1996.

Red Brick Systems
Star Schemas and STARjoin Technology
http://www.redbrick.com/rbs/whitepapers/star_wp.html
Last updated: October 22, 1995. Accessed: December 9, 1996.

Red Brick Systems
The Data Warehouse: The Competitive Advantage for the 1990s
http://www.redbrick.com/rbs/whitepapers/datawh_wp.html
Last updated: April 30, 1996. Accessed: December 9, 1996.

SAS Institute Inc.
A SAS Institute White Paper: Data Warehousing Methodology
http://www.sas.com/solutions/papers/dw_method.html
Last updated: June 21, 1996. Accessed: December 9, 1996.

SAS Institute Inc.
A SAS Institute White Paper: The SAS Data Warehouse
http://www.sas.com/solutions/papers/dw_gen.html
Last updated: June 21, 1996. Accessed: December 9, 1996.

SAS Institute Inc.
Data Warehousing Checklists for Success
http://www.sas.com/new/checklists.html
Last updated: July 13, 1996. Accessed: December 9, 1996.

Saylor M., Acharya M.G. and Trenkamp R.G.: MicroStrategy, Inc.
True Relational OLAP: The Future of Decision Support
http://www.strategy.com/tro_dbj.htm
Last updated: November 26, 1996. Accessed: December 9, 1996.

Shah A.D. and Milstein B.M.
Data Warehousing: Practical Tips for Successful Implementation
http://www.sw-expo.com/perfdev/data-warehousing.html
Last updated: April 18, 1996. Accessed: December 9, 1996.

The Data Warehousing Institute
TDWI: Ten Mistakes To Avoid
http://www.tekptnr.com/tpi/tdwi/papers/10mistks.htm
Last updated: June 14, 1996. Accessed: December 9, 1996.

The OLAP Council
OLAP and OLAP Server Definitions
http://www.access.digex.net/~grimes/olap/glossary.html
Last updated: July 28, 1995. Accessed: December 9, 1996.

White C.: IBM
Data Warehousing: The Role of the Information Directory
http://www.software.ibm.com/data/dataguide/dguide.ps
Last updated: August 26, 1996. Accessed: December 9, 1996.

Widom J.: Stanford University
Research Problems in Data Warehousing
http://www-db.stanford.edu/pub/papers/warehouse-research.ps
Last updated: June 27, 1996. Accessed: December 9, 1996.

Zhuge Y., Garcia-Molina H., Hammer J. and Widom J.: Stanford University
View Maintenance in a Warehousing Environment
http://www-db.stanford.edu/pub/papers/anomaly-short.ps
Last updated: June 26, 1996. Accessed: December 9, 1996.

Link Pages

Arbor Software
Arbor Software White Papers
http://www.arborsoft.com/essbase/wht_ppr.html
Last updated: November 27, 1996. Accessed: December 9, 1996.

Archer Decision Sciences, Inc.
Archer Decision Sciences Home Page
http://netmar.com/~nraden/index.htm
Last updated: August 29, 1996. Accessed: December 9, 1996.

Brown T.C.: Action Inquiry Network (ActNet)
Action Inquiry, Various Aspects
http://enhanced-designs.com/actnet/index.htm
Last updated: December 5, 1996.   Accessed: December 9, 1996.

Greenfield L.
The Data Warehousing Information Center
http://pwp.starnetinc.com/larryg/index.html
Last updated: December 9, 1996.   Accessed: December 9, 1996.

IBM
Data Management White Papers
http://www.software.ibm.com/data/dbtools/dbsmwp.html
Last updated: .   Accessed: December 9, 1996.

International Data Warehousing Association
International Data Warehousing Association
http://www.idwa.org/
Last updated: September 24, 1996.   Accessed: December 9, 1996.

McGuff F.: Frank McGuff
Frank McGuff Home Page
http://members.aol.com/fmcguff/index.htm
Last updated: December 4, 1996.   Accessed: December 9, 1996.

MicroStrategy, Inc.
Data Warehousing Forum
http://www.strategy.com/msi_dwf1.htm
Last updated: December 9, 1996.   Accessed: December 9, 1996.

Nova Southeastern University
Qualitative Research Web Sites
http://www.nova.edu/ssss/QR/web.html
Last updated: September 26, 1996.   Accessed: December 9, 1996.

Stanford University
Data Warehousing Publications
http://www-db.stanford.edu/warehousing/publications.html
Last updated: December 2, 1996.   Accessed: December 9, 1996.

The Data Warehousing Institute
The Data Warehousing Institute
http://www.tekptnr.com/tpi/tdwi/
Last updated: November 20, 1996.   Accessed: December 9, 1996.

Appendices

Contents:

Appendix A: OLAP and OLAP SERVER DEFINITIONS

Available at http://www.access.digex.net/~grimes/olap/glossary.html

OLAP: ON-LINE ANALYTICAL PROCESSING

On-Line Analytical Processing (OLAP) is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

OLAP functionality is characterized by dynamic multi-dimensional analysis of consolidated enterprise data supporting end user analytical and navigational activities including:

- calculations and modeling applied across dimensions, through hierarchies and/or across members
- trend analysis over sequential time periods
- slicing subsets for on-screen viewing
- drill-down to deeper levels of consolidation
- reach-through to underlying detail data
- rotation to new dimensional comparisons in the viewing area

OLAP is implemented in a multi-user client/server mode and offers consistently rapid response to queries, regardless of database size and complexity. OLAP helps the user synthesize enterprise information through comparative, personalized viewing, as well as through analysis of historical and projected data in various "what-if" data model scenarios. This is achieved through use of an OLAP Server.

OLAP SERVER

An OLAP server is a high-capacity, multi-user data manipulation engine specifically designed to support and operate on multi-dimensional data structures. A multi-dimensional structure is arranged so that every data item is located and accessed based on the intersection of the dimension members which define that item. The design of the server and the structure of the data is optimized for rapid ad-hoc information retrieval in any orientation, as well as for fast, flexible calculation and transformation of raw data based on formulaic relationships. The OLAP Server may either physically stage the processed multi-dimensional information to deliver consistent and rapid response times to end users, or it may populate its data structures in real-time from relational or other databases, or offer a choice of both. Given the current state of technology and the end user requirement for consistent and rapid response times, staging the multi-dimensional data in the OLAP Server is often the preferred method.

OLAP GLOSSARY

Defined terms:

Aggregate
Analysis, multi-dimensional
Array, multi-dimensional
Calculated member
Cell
Children
Column dimension
Consolidate
Cube
Dense
Derived data
Derived members
Detail member
Dimension
Drill down/up
Formula
Formula, cross-dimensional
Generation, hierarchical
Hierarchical relationships
Horizontal dimension
Hypercube
Input members
Level, hierarchical
Member, dimension
Member combination
Missing data, missing value
Multi-dimensional data structure
Multi-dimensional query language
Navigation
Nesting (of multi-dimensional columns and rows)
Non-missing data
OLAP client
Page dimension
Page display
Parent
Pivot
Pre-calculated/pre-consolidated data
Reach through
Roll-up
Rotate
Row dimension
Scoping
Selection
Slice
Slice and dice
Sparse
Vertical dimension

Definitions:

AGGREGATE

See: Consolidate

ANALYSIS, MULTI-DIMENSIONAL

The objective of multi-dimensional analysis is for end users to gain insight into the meaning contained in databases. The multi-dimensional approach to analysis aligns the data content with the analyst's mental model, hence reducing confusion and lowering the incidence of erroneous interpretations. It also eases navigating the database, screening for a particular subset of data, asking for the data in a particular orientation and defining analytical calculations. Furthermore, because the data is physically stored in a multi- dimensional structure, the speed of these operations is many times faster and more consistent than is possible in other database structures. This combination of simplicity and speed is one of the key benefits of multi-dimensional analysis.

ARRAY, MULTI-DIMENSIONAL

A group of data cells arranged by the dimensions of the data. For example, a spreadsheet exemplifies a two-dimensional array with the data cells arranged in rows and columns, each being a dimension. A three-dimensional array can be visualized as a cube with each dimension forming a side of the cube, including any slice parallel with that side. Higher dimensional arrays have no physical metaphor, but they organize the data in the way users think of their enterprise. Typical enterprise dimensions are time, measures, products, geographical regions, sales channels, etc.
Synonyms: Multi-dimensional Structure, Cube, Hypercube

CALCULATED MEMBER

A calculated member is a member of a dimension whose value is determined from other members' values (e.g., by application of a mathematical or logical operation). Calculated members may be part of the OLAP server database or may have been specified by the user during an interactive session. A calculated member is any member that is not an input member.

CELL

A single datapoint that occurs at the intersection defined by selecting one member from each dimension in a multi-dimensional array. For example, if the dimensions are measures, time, product and geography, then the dimension members: Sales, January 1994, Candy Bars and United States specify a precise intersection along all dimensions that uniquely identifies a single data cell, which contains the value of candy bar sales in the United States for the month of January 1994.
See: Member Combination

CHILDREN

Members of a dimension that are included in a calculation to produce a consolidated total for a parent member. Children may themselves be consolidated levels, which requires that they have children. A member may be a child for more than one parent, and a child's multiple parents may not necessarily be at the same hierarchical level, thereby allowing complex, multiple hierarchical aggregations within any dimension.

COLUMN DIMENSION

See: Page Display

CONSOLIDATE

Multi-dimensional databases generally have hierarchies or formula-based relationships of data within each dimension. Consolidation involves computing all of these data relationships for one or more dimensions, for example, adding up all Departments to get Total Division data. While such relationships are normally summations, any type of computational relationship or formula might be defined.
Synonyms: Roll-up, Aggregate
See: Formula, Hierarchical Relationships, Children, Parents

CUBE

See: Array, Multi-dimensional

DENSE

A multi-dimensional database is dense if a relatively high percentage of the possible combinations of its dimension members contain data values. This is the opposite of sparse.

DERIVED DATA

Derived data is produced by applying calculations to input data at the time the request for that data is made, i.e., the data has not been pre-computed and stored on the database. The purpose of using derived data is to save storage space and calculation time, particularly for calculated data that may be infrequently called for or that is susceptible to a high degree of interactive personalization by the user. The tradeoff is slower retrievals.
See: Pre-calculated Data

DERIVED MEMBERS

Derived members are members whose associated data is derived data.

DETAIL MEMBER

A detail member of a dimension is the lowest level number in its hierarchy.
See: Level

DIMENSION

A dimension is a structural attribute of a cube that is a list of members, all of which are of a similar type in the user's perception of the data. For example, all months, quarters, years, etc., make up a time dimension; likewise all cities, regions, countries, etc., make up a geography dimension. A dimension acts as an index for identifying values within a multi-dimensional array. If one member of the dimension is selected, then the remaining dimensions in which a range of members (or all members) are selected defines a sub-cube. If all but two dimensions have a single member selected, the remaining two dimensions define a spreadsheet (or a "slice" or a "page"). If all dimensions have a single member selected, then a single cell is defined. Dimensions offer a very concise, intuitive way of organizing and selecting data for retrieval, exploration and analysis.

DRILL DOWN/UP

Drilling down or up is a specific analytical technique whereby the user navigates among levels of data ranging from the most summarized (up) to the most detailed (down). The drilling paths may be defined by the hierarchies within dimensions or other relationships that may be dynamic within or between dimensions. For example, when viewing sales data for North America, a drill-down operation in the Region dimension would then display Canada, the eastern United States and the Western United States. A further drill- down on Canada might display Toronto, Vancouver, Montreal, etc.

FORMULA

A formula is a database object, which is a calculation, rule or other expression for manipulating the data within a multi-dimensional database. Formulae define relationships among members. Formulae are used by OLAP database builders to provide great richness of content to the server database. Formulae are used by end users to model enterprise relationships and to personalize the data for greater visualization and insight.

FORMULA, CROSS-DIMENSIONAL

Formulae with all operands within a dimension are common, even in non-OLAP systems: e.g., Profit = Sales - Expense might appear in a simple spreadsheet product. In an OLAP system, such a calculation rule would normally calculate Profit for all combinations of the other dimensions in the cube (e.g., for all Products, for all Regions, for all Time Periods, etc.) using the respective Revenue and Expense data from those same dimensions. Part of the power of an OLAP system is the extensive multi-dimensional application of such a simply stated rule, which could be specified by the OLAP application builder or created by the end user in an interactive session. The true analytical power of an OLAP server, however, is evidenced in its ability to evaluate formulae where there are members from more than one dimension. An example is a multi-dimensional allocation rule used in business unit profitability applications. If, for example, a company has a Business Unit dimension and one of the business units (XYZ) is funding a special advertising campaign for Product A, and the other business units which also sell Product A are willing to share the advertising costs in proportion to their sales of the product, then the formula would be:

ADVERTISING EXPENSE = (PRODUCT A SALES/TOTAL CORPORATION PRODUCT A SALES) * ADVERTISING EXPENSE FOR PRODUCT A FOR BUSINESS UNIT XYZ
Here, Advertising is from the Measures dimension wherever it intersects with other dimensions (e.g., Business Unit, Product), but Product A Sales is more specific; it is Sales from the Measures dimension restricted to the Product A member from the Product dimension. The Advertising Expense to be shared is the Advertising Expense for Product A spent by Business Unit XYZ that the business units which have non-zero sales of Product A agreed to share. These references to several dimensions within the same rule make it a Cross-Dimensional Formula.

GENERATION, HIERARCHICAL

Two members of a hierarchy have the same generation if they have the same number of ancestors leading to the top. For example, the top member of a dimension is from Generation 1. There may be two or more members in Generation 1 if there are multiple hierarchies in the dimension.

NOTE: The terms generation and level are both necessary to describe sub-groups of dimension members, since, for example, although two siblings share the same parent and are therefore of the same generation, they won't be from the same level if one of the siblings has a child and the other doesn't.
Synonyms: Peer, Sibling
See: Level, Hierarchical Relationships, Parent, Children

HIERARCHICAL RELATIONSHIPS

Any dimension's members may be organized based on parent-child relationships, typically where a parent member represents the consolidation of the members which are its children. The result is a hierarchy, and the parent/child relationships are hierarchical relationships.

HORIZONTAL DIMENSION

See: Page Display

HYPERCUBE

See: Cube, Array, Multi-dimensional

INPUT MEMBERS

Input members have values that are loaded directly from either manual entry or by accessing another computer-based data source, as opposed to being calculated from the raw data.

LEVEL, HIERARCHICAL

Members of a dimension with hierarchies are at the same level if, within their hierarchy, they have the same maximum number of descendants in any single path below. For example, in an Accounts dimension which consists of general ledger accounts, all of the detail accounts are Level 0 members. The accounts one level higher are Level 1, their parents are Level 2, etc. It can happen that a parent has two or more children which are different levels, in which case the parent's level is defined as one higher than the level of the child with the highest level.
See: Generation, Hierarchical

MEMBER, DIMENSION

A dimension member is a discrete name or identifier used to identify a data item's position and description within a dimension. For example, January 1989 or 1Qtr93 are typical examples of members of a Time dimension. Wholesale, Retail, etc., are typical examples of members of a Distribution Channel dimension.
Synonyms: Position, Item, Attribute

MEMBER COMBINATION

A member combination is an exact description of a unique cell in a multi-dimensional array, consisting of a specific member selection in each dimension of the array.
See: Cell

MISSING DATA, MISSING VALUE

A special data item which indicates that the data in this cell does not exist. This may be because the member combination is not meaningful (e.g., snowmobiles may not be sold in Miami) or has never been entered. Missing data is similar to a null value or N/A, but is not the same as a zero value.

MULTI-DIMENSIONAL DATA STRUCTURE

See: Array, Multi-dimensional

MULTI-DIMENSIONAL QUERY LANGUAGE

A computer language that allows one to specify which data to retrieve out of a cube. The user process for this type of query is usually called slicing and dicing. The result of a multi-dimensional query is either a cell, a two-dimensional slice, or a multi-dimensional sub-cube.

NAVIGATION

Navigation is a term used to describe the processes employed by users to explore a cube interactively by drilling, rotating and screening, usually using a graphical OLAP client connected to an OLAP server.

NESTING (OF MULTI-DIMENSIONAL COLUMNS AND ROWS)

Nesting is a display technique used to show the results of a multi-dimensional query that returns a sub-cube, i.e., more than a two-dimensional slice or page. The column/row labels will display the extra dimensionality of the output by nesting the labels describing the members of each dimension. For example, the display's columns may be:

January				February				March			
Actual		Budget		Actual		Budget		Actual		Budget	
Prod A	Prod B	Prod A	Prod B	Prod A	Prod B	Prod A	Prod B	Prod A	Prod B	Prod A	Prod B

These columns contain three dimensions, nested in the user's preferred arrangement. Likewise, a report's rows may contain nested dimensions:

Chocolate Bars	Unit Sales	xxxx	xxxx	xxxx
	Revenue	xxxx	xxxx	xxxx
	Margin	xxxx	xxxx	xxxx
Fruit Bars	Unit Sales	xxxx	xxxx	xxxx
	Revenue	xxxx	xxxx	xxxx
	Margin	xxxx	xxxx	xxxx

NON-MISSING DATA

Data which exists and has values, as opposed to null or missing data.

OLAP CLIENT

End user applications that can request slices from OLAP servers and provide two-dimensional or multi-dimensional displays, user modifications, selections, ranking, calculations, etc., for visualization and navigation purposes. OLAP clients may be as simple as a spreadsheet program retrieving a slice for further work by a spreadsheet- literate user or as high-functioned as a financial modeling or sales analysis application.

PAGE DIMENSION

A page dimension is generally used to describe a dimension which is not one of the two dimensions of the page being displayed, but for which a member has been selected to define the specific page requested for display. All page dimensions must have a specific member chosen in order to define the appropriate page for display.

PAGE DISPLAY

The page display is the current orientation for viewing a multi-dimensional slice. The horizontal dimension(s) run across the display, defining the column dimension(s). The vertical dimension(s) run down the display, defining the contents of the row dimension(s). The page dimension-member selections define which page is currently displayed. A page is much like a spreadsheet, and may in fact have been delivered to a spreadsheet product where each cell can be further modified by the user.

PARENT

The member that is one level up in a hierarchy from another member. The parent value is usually a consolidation of all of its children's values.
See: Children

PIVOT

See: Rotate

PRE-CALCULATED/PRE-CONSOLIDATED DATA

Pre-calculated data is data in output member cells that are computed prior to, and in anticipation of, ad-hoc requests. Pre-calculation usually results in faster response to queries at the expense of storage. Data that is not pre-calculated must be calculated at query time.
See: Derived Data/Members, Output Data

REACH THROUGH

Reach through is a means of extending the data accessible to the end user beyond that which is stored in the OLAP server. A reach through is performed when the OLAP server recognizes that it needs additional data and automatically queries and retrieves the data from a data warehouse or OLTP system.

ROLL-UP

See: Consolidate

ROTATE

To change the dimensional orientation of a report or page display. For example, rotating may consist of swapping the rows and columns, or moving one of the row dimensions into the column dimension, or swapping an off-spreadsheet dimension with one of the dimensions in the page display (either to become one of the new rows or columns), etc. A specific example of the first case would be taking a report that has Time across (the columns) and Products down (the rows) and rotating it into a report that has Product across and Time down. An example of the second case would be to change a report which has Measures and Products down and Time across into a report with Measures down and Time over Products across. An example of the third case would be taking a report that has Time across and Product down and changing it into a report that has Time across and Geography down.

Synonym: Pivot

ROW DIMENSION

See: Page Display

SCOPING

Restricting the view of database objects to a specified subset. Further operations, such as update or retrieve, will affect only the cells in the specified subset. For example, scoping allows users to retrieve or update only the sales data values for the first quarter in the east region, if that is the only data they wish to receive.

SELECTION

A selection is a process whereby a criterion is evaluated against the data or members of a dimension in order to restrict the set of data retrieved. Examples of selections include the top ten salespersons by revenue, data from the east region only and all products with margins greater than 20 percent.
Synonyms: Condition, Screen, Filter

SLICE

A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset. For example, if the member Actuals is selected from the Scenario dimension, then the sub-cube of all the remaining dimensions is the slice that is specified. The data omitted from this slice would be any data associated with the non-selected members of the Scenario dimension, for example Budget, Variance, Forecast, etc. From an end user perspective, the term slice most often refers to a two- dimensional page selected from the cube.

SLICE AND DICE

The user-initiated process of navigating by calling for page displays interactively, through the specification of slices via rotations and drill down/up.

SPARSE

A multi-dimensional data set is sparse if a relatively high percentage of the possible combinations (intersections) of the members from the data set's dimensions contain missing data. The total possible number of intersections can be computed by multiplying together the number of members in each dimension. Data sets containing one percent, .01 percent, or even smaller percentages of the possible data exist and are quite common.
See: Dense

VERTICAL DIMENSION

See: Page Display

Appendix B: Data Warehouse Terminology

Available at http://www.std.com/CreativeData/credata/termin.html

Bitmapped Indexing

An family of advanced indexing algorithms that optimize RDBMS query performance by maximizing the search capability of the index per unit of memory and per CPU instruction. Properly implemented, bitmapped indices eliminate all table scans in query and join processing.

Business Model

An object-oriented model that captures the kinds of things in a business or a business area and the relationships associated with those things (and sometimes associated business rules, too). Note that a business model exists independently of any data or database. A data warehouse should be designed to match the underlying business models or else no tools will fully unlock the data in the warehouse.

Corporate Data

All the databases of the company. This includes legacy systems, old and new transaction systems, general business systems, client/server databases, data warehouses and data marts.

Data Dictionary

A collection of **Meta Data**. Many kinds of products in the data warehouse arena use a data dictionary, including database management systems, modeling tools, middleware, and query tools.

Data Mart

A subset of a data warehouse that focuses on one or more specific subject areas. The data usually is extracted from the data warehouse and further denormalized and indexed to support intense usage by targeted customers.

Data Mining

Techniques for finding patterns and trends in large data sets. See also **Data Visualization**.

Data Model

The road map to the data in a database. This includes the source of tables and columns, the meanings of the keys, and the relationships between the tables.

Data Visualization

Techniques for turning data into information by using the high capacity of the human brain to recognize visually recognize patterns and trends. There are many specialized techniques designed to make particular kinds of visualization easy.

Data Warehouse

A database built to support information access. Typically a data warehouse is fed from one or more transaction databases. The data needs to be cleaned and restructured to support queries, summaries, and analyses.

Decision Support
Data access targeted to provide the information needed by business decision makers. Examples include pricing, purchasing, human resources, management, manufacturing, etc.

Decision Support System (DSS)
Database(s), warehouse(s), and/or mart(s) in conjunction with reporting and analysis software optimized to support timely business decision making.

Meta Data
Literally, "data about data." More usefully, descriptions of what kind of information is stored where, how it is encoded, how it is related to other information, where it comes from, and how it is related to your business. A hot topic right now is standardizing meta data across products from different vendors.

Methodology
The steps followed to guarantee repeatability of success. A good methodology is built on top of real world experience. For example, see **The Hughes-Vollum Methodology**.

Middleware
Hardware and software used to connect clients and servers, to move and structure data, and/or to pre-summarize data for use by queries and reports.

Multi-dimensional database (MDD)
A DBMS optimized to support multi-dimensional data. The best systems support standard RDBMS functionality and add high-bandwith support for multi-dimensional data and queries. Users that need a lot of slices and dices might appreciate a multi-dimensional database.

Object Oriented Analysis (OOA)
A process of abstracting a problem by identifying the kinds of entities in the problem domain, the is-a relationships between the kinds (kinds are known as classes, is-a relationships as subtype/supertype, subclass/superclass, or less commonly, specialization/generalization), and the has-a relationships between the classes. Also identified for each class are its attributes (e.g. class **Person** has attribute Hair Color) and its conventional relationships to other classes(e.g. class **Order** has a relationship Customer to class **Customer**.)

Object Oriented Design (OOD)
A design methodology that uses Object Oriented Analysis to promote object reusability and interface clarity.

OLAP
An acronym for **On Line Analytical Processing**.

On Line Analytical Processing (OLAP)
A common use of a data warehouse that involves real time access and analysis of multi-dimensional data such as order information.

Performance

Data, summaries, and analyses need to be delivered in a timely fashion. Performance is often a key issue with data warehouses: the right answer isn't worth much if it shows up after the decisions have been made.

Rapid Application Development (RAD)

Part of a methodology that specifies incremental development with constant feedback from the customers. The point is to keep projects focused on delivering value and to keep clear and open lines of communication. English is not adequate for specification of computer systems, even small ones. RAD overcomes the limitations of language by minimizing the time between concept and implementation.

Relational On-Line Analytic Processing (ROLAP)

OLAP based on conventional relational databases rather than specialized multi-dimensional databases.

Replication

A standard technique in data warehousing. For performance and reliability several independent copies are often created of each data warehouse. Even data marts can require replication on multiple servers to meet performance and reliability standards.

Replicator

Any of a class of product that supports **replication**. Often these tools use special load and unload database APIs and have scripting languages that support automation.

Report

A repeatable, formatted, nonatomic request for information from a database. Usually a report formats and combines several related **queries**.

Reporting Strategy

A top down collection of methodology, products, plans, and teams that ensure business people can get information reliably, accurately, and understandably. It includes choosing tools matched to the organization's particular needs and existing infrastructure, capturing the business models used by the business people, finding source data, integrating all the above into a data warehouse and/or data marts as needed.

Query

A specific atomic request for information from a database.

Security

The right data for the right person. Note that a business analyst may need access to summaries of data s/he should not see. Security systems need to make this easy to implement while making sure outsiders or rogue employees do not see data they should not see.

Snowflake Schema

A layering of **Star Schema** that scales that technique to handle an entire warehouse.

Star Schema

A standard technique for designing the summary tables of a data warehouse. "Fact" tables each join to a larger number of independent "dimension" tables. The tables may be partially denormalized for performance, but most queries will still need to join in one or more of the star tables.

Copyright 1996, Creative Data, Inc.

Appendix C: OLAP Product Evaluation Rules

This Appendix is reproduced from "Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate", a paper by E.F. Codd, S.B. Codd and C.T. Salley[41]

The twelve rules for evaluating OLAP products are:

1. Multi-Dimensional Conceptual View
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 Dimensions and Aggregation Levels

1. Multi-Dimensional Conceptual View

A user-analyst's view of the enterprise's universe is multi-dimensional in nature. Accordingly, the user-analyst's conceptual view of OLAP models should be multi-dimensional in nature. This multi-dimensional conceptual schema or user view facilitates model design and analysis, as well as inter and intra dimensional calculations through a more intuitive analytical model. Accordingly user-analysts are able to manipulate such multi-dimensional data models more easily and intuitively than is the case with single dimensional models. For instance, the need to "slice and dice," or pivot and rotate consolidation paths within a model is common. Multi-dimensional models make these manipulations easily, whereas achieving a like result with older approaches requires significantly more time and effort.

2. Transparency

Whether OLAP is or is not part of the user's customary front-end (e. g., spreadsheet or graphics package) product, that fact should be transparent to the user. If OLAP is provided within the context of a client-server architecture, then this fact should be transparent to the user-analyst as well. OLAP should be provided within the context of a true open systems architecture, allowing the analytical tool to be embedded anywhere the user-analyst desires, without adversely impacting the functionality of the host tool.

[41] Codd E.F., Codd S.B. and Salley C.T.: E. F. Codd & Associates. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* p18.

Transparency is crucial to preserving the user's existing productivity and proficiency with the customary front-end, providing the appropriate level of function, and assuring that needless complexity is in no way introduced or otherwise increased.

Additionally, it should be transparent to the user as to whether or not the enterprise data input to the OLAP tool comes from a homogeneous or heterogeneous database environment.

3. Accessibility

The OLAP user-analyst must be able to perform analysis based upon a common conceptual schema composed of enterprise data in relational DBMS, as well as data under control of the old legacy DBMS, access methods, and other non-relational data stores at the same time as the basis of a common analytical model. That is to say that the OLAP tool must map its own logical schema to heterogeneous physical data stores, access the data, and perform any conversions necessary to present a single, coherent and consistent user view. Moreover, the tool and not the end-user analyst must be concerned about where or from which type of systems the physical data is actually coming. The OLAP system should access only the data actually required to perform the indicated analysis and not take the common "kitchen sink" approach which brings in unnecessary input.

4. Consistent Reporting Performance

As the number of dimensions or the size of the database increases, the OLAP user-analyst should not perceive any significant degradation in reporting performance. Consistent reporting performance is critical to maintaining the ease-of-use and lack of complexity required in bringing OLAP to the end-user.

If the user-analyst were able to perceive any significant difference in reporting performance relating to the number of dimensions requested, there would very likely be compensating strategies developed, such as asking for information to be presented in ways other than those really desired. Spending one's time in devising ways of circumventing the system in order to compensate for its inadequacies is not what end-user products are about.

5. Client-Server Architecture

Most data currently requiring on-line analytical processing is stored on mainframe systems and accessed via personal computers. It is therefore mandatory that the OLAP products be capable of operating in a client-server environment. To this end, it is imperative that the server component of OLAP tools be sufficiently intelligent such that various clients can be attached with minimum effort and integration programming.

The intelligent server must be capable of performing the mapping and consolidation between disparate logical and physical enterprise database schema necessary to effect transparency and to build a common conceptual, logical and physical schema.

6. Generic Dimensionality

Every data dimension must be equivalent in both its structure and operational capabilities. Additional operational capabilities may be granted to selected dimensions, but since dimensions are symmetric, a given additional function may be granted to any dimension. The basic data structure, formulae, and reporting formats should not be biased toward any one data dimension.

7. Dynamic Sparse Matrix Handling

The OLAP tools' physical schema must adapt fully to the specific analytical model being created to provide optimal sparse matrix handling. For any given sparse matrix, there exists one and only one optimum physical schema. This optimal schema provides both maximum memory efficiency and matrix operability unless of course, the entire data set can be cached in memory. The OLAP tool's basic physical data unit must be configurable to any subset of the available dimensions, in any order, for practical operations within large analytical models. The physical access methods must also be dynamically changeable and should contain different types of mechanisms such as:

1. direct calculation;
2. B-trees and derivatives,
3. hashing;
4. the ability to combine these techniques where advantageous.

Sparseness (missing cells as a percentage of possible cells) is but one of the characteristics of data distribution. The inability to adjust (morph) to the data set's data distribution can make fast, efficient operation unobtainable. If the OLAP tool cannot adjust according to the distribution of values of the data to be analyzed, models which appear to be practical, based upon the number of consolidation paths and dimensions, or the size of the enterprise source data, may be needlessly large and/or hopelessly slow in actuality. Access speed should be consistent regardless of the order of cell access and should remain fairly constant across models containing different numbers of data dimensions or varying sizes of data sets.

For example, given a set of input data from the enterprise database which is perfectly dense (every possible input combination contains a value, no nulls), it is possible to predict the size of the resulting data set after consolidation across all modeled data dimensions.

For example, in a particular five-dimensional analytical model, let us suppose that the physical schema size after model consolidation is two-and-one-half times the size of the input data from the enterprise database.

However, if the enterprise data is sparse, and has certain distribution characteristics, then the resulting physical schema might be one-hundred times the size of the enterprise data input. But, given the same size data set, and the same degree of sparseness, but with different data distribution, the size of the resulting physical schema might be only two and-one-half times the size of the enterprise data input as in the case of the perfectly dense example. Or, we could experience anything in between these two extremes. “Eyeballing” the data in an attempt to form an educated guess is as hopeless as is using conventional statistical analysis tools to obtain crosstabs of the data.

Because conventional statistical analysis tools always compare only one dimension against one other dimension, without regard for the other, perhaps numerous, data dimensions, they are unsuitable to multi-dimensional data analysis. Even if such tools could compare all dimensions at once (which they can’t), the resulting crosstab would be the size of the product of all the data dimensions, which would be the maximum size of the physical schema itself.

By adapting its physical data schema to the specific analytical model, OLAP tools can empower user-analysts to easily perform types of analysis which previously have been avoided because of their perceived complexity. The extreme unpredictability and volatility in the behavior of multi-dimensional data models precludes the successful use of tools which rely upon a static physical schema and whose basic unit of data storage has fixed dimensionality (e.g., cell, record or two-dimensional sheet). A fixed, physical schema which is optimal for one analytical model, will typically be impractical for most others. Rather than basing a physical schema upon cells, records, two dimensional sheets, or some other similar structure, OLAP tools must dynamically adapt the model’s physical schema to the indicated dimensionality and especially to the data distribution of each specific model.

8. Multi-User Support

Oftentimes, several user-analyst’s have a requirement to work concurrently with either the same analytical model or to create different models from the same enterprise data. To be regarded as strategic, OLAP tools must provide concurrent access (retrieval and update), integrity, and security.

9. Unrestricted Cross-Dimensional Operations

The various roll-up levels within consolidation paths, due to their inherent hierarchical nature, represent in outline form, the majority of 1:1, 1:M, and dependent relationships in an OLAP model or application. Accordingly, the tool itself should infer the associated calculations and not require the user-analyst to explicitly define these inherent calculations. Calculations not resulting from these inherent relationships require the definition of various formulae according to some language which of course must be computationally complete.

Such a language must allow calculation and data manipulation across any number of data dimensions and must not restrict or inhibit any relationship between data cells regardless of the number of common data attributes each cell contains.

For example, consider the difference between a single dimensional calculation and a cross-dimensional calculation. The single dimensional calculation: *Contribution = Revenue - Variable Cost* defines a relationship between attributes in only one data dimension, which we shall call *D_ACCOUNTS*. Upon calculation, what occurs is that the relationship is calculated for all cells of all data dimensions in the data model which possess the attribute *Contribution*

A cross-dimensional relationship and the associated calculations provide additional challenges. For example, given the following simple five-dimensional outline:

- **D[42]_Accounts**
  - Sales
  - Overhead
  - InterestRate
  - et cetera
- **D_Corporate**
  - United Kingdom
    - London
    - York
    - et cetera
  - France
    - Paris
    - Cannes
    - et cetera
- **D_FiscalYear**
  - Quarter1
    - January
    - February
    - March
  - Quarter2
    - April
    - May
    - June
    - et cetera
- **D_Products**
  - Audio
  - Video
  - et cetera
- **D_Scenario**

[42] "D_" is used each time to indicate that this top most aggregation level is the dimension.

Budgeted
Actual
Variance
et cetera

Sample Five-Dimensional Outline Structure

The formula to allocate corporate overhead to parts of the organization such as local offices (Paris, Cannes, et cetera) based upon their respective contributions to overall company sales might appear thus:

Overhead equals the percentage of total sales represented by the sales of each individual local office multiplied by total corporate overhead

Here is another example of necessary cross-dimensional calculations. Suppose that the user-analyst desires to specify that for all French cities, the variable *InterestRate* which is used in subsequent calculations, should be set to the value of the *BUDGETED MARCH INTERESTRATE* for the city of Paris for all months, across all data dimensions. Had the user-analyst not specified the city, month and scenario, the attributes would alter and stay consistent with the month attributes of the data cell being calculated when the analytical model is animated. The described calculation could be expressed as :

If *the value within the designated cell appears within the consolidation path D_Corporate, beneath the consolidation level France,* **then** *the global interest rate becomes the value of the interest rate for the month of March which is budgeted for the city of Paris*

10. Intuitive Data Manipulation

Consolidation path re-orientation, drilling down across columns or rows, zooming out, and other manipulation inherent in the consolidation path outlines should be accomplished via direct action upon the cells of the analytical model, and should neither require the use of a menu nor multiple trips across the user interface. The user-analyst's view of the dimensions defined in the analytical model should contain all information necessary to effect these inherent actions.

11. Flexible Reporting

Analysis and presentation of data is simpler when rows, columns, and cells of data which are to be visually compared are arranged in proximity or by some logical grouping occur ring naturally in the enterprise. Reporting must be capable of presenting data to be synthesized, or information resulting from animation of the data model according to any possible orientation. This means that the rows, columns, or page headings must each be capable of containing/displaying from 0 to N dimensions each, where N is the number of dimensions in the entire analytical model.

Additionally, each dimension contained/displayed in one of these rows, columns, or page headings must itself be capable of containing/displaying any subset of the members, in any order, and provide a means of showing the inter-consolidation path relationships between the members of that subset such as indentation.

12. Unlimited Dimensions and Aggregation Levels

Research into the number of dimensions possibly required by analytical models indicates that as many as nineteen concurrent data dimensions (this was an actuarial model) may be needed. Thus the strong recommendation that any serious OLAP tool should be able to accommodate at least fifteen and preferably twenty data dimensions within a common analytical model.

Furthermore, each of these generic dimensions must allow an essentially unlimited number of user-analyst defined aggregation levels within any given consolidation path.

Appendix D: Case Studies and Action Research

This data warehousing project was conducted in the form of a case study, and many of the methods of action inquiry or research are applicable. A statement from the Action Inquiry Network (ActNet) makes the point that "*the action inquiry model focuses primarily on identifying and resolving difficult, complex, real-life problems critical to organizations and society. This includes the formidable challenges of leadership, innovation, informed participation, and prejudice.*" [43]

Although the action research model is applied more commonly in the fields of education (teaching and learning), health services, social sciences and economics, this statement is a very close fit for the issues that are confronted when developing a data warehouse. The success of the project is very dependent on the support and input of the senior leadership. Their informed participation is crucial. The project is primarily for the purpose of improving the decision making process. There also needs to be a willing open-mindedness to consider and suggest new ways of doing things, including organizational and social changes. An attitude of pre-judgment and prejudice is almost certain to deprive the organization of some of the best ideas for improvement.

> "A crucial issue in action inquiry is whether a problem is routine or non-routine, trivial or difficult. The difference between routine/trivial and non-routine/difficult problems is, admittedly, not always easy to discern. Nevertheless, the action inquiry model focuses primarily on identifying and resolving difficult, complex, real-life problems critical to organizations and society. This includes the formidable challenges of leadership, innovation, informed participation, and prejudice. Bewildering problems often develop when members of a group try to formulate and carry out a new strategy and evaluate their work together.
>
> "Effective groups can resolve difficult problems by taking innovative action relatively soon. As the participants question their underlying programs or the credibility of ideas, they maintain high levels of interpersonal openness. They accept that while openness is actually or potentially embarrassing, threatening, or frustrating, it is simultaneously necessary to increase trust and individuality in their group. They may deny these difficulties but rarely do such instances not get noticed, challenged, and corrected. As they improve action inquiry skills, their minimally defensive interpersonal and group relations enable them to innovate and become productive."[44]

[43] Brown T.C.: Action Inquiry Network (ActNet). *Action Inquiry, Problem Types*

[44] Brown T.C.: Action Inquiry Network (ActNet). *Action Inquiry, Problem Types*

Appendix E: Data Dictionary - Files & Elements (selected entries)

CI_ADDRESS

Addresses held as part of Common Information about persons

Item	Type	Size	Dec			Description
name_number	C	6		CHAR	6	Computerised record number assigned to this person
address_number	C	1		CHAR	1	Address number of this address record
date_address_modified	D	8		ZONED-U	8	Date this address was modified
good_address	C	1		CHAR	1	Indicates the validity of this address
address1	C	30		CHAR	30	1st line of address, eg number and street
address2	C	30		CHAR	30	Line 2 of address (optional)
town	C	30		CHAR	30	Town or suburb
state	C	10		CHAR	10	State part of address
postcode	C	7		CHAR	7	Post code
country	C	20		CHAR	20	Country part of address
phone	C	15		CHAR	15	Telephone number part of person address

PRIMARY KEY: address_key : name_number
: address_number

SR_CODE_COURSE_MASTER

Course code, description and academic details

Item	Type	Size	Dec			Description
year_yyyy	C	4		CHAR	4	Year (YYYY) for this record
course_code	C	5		CHAR	5	Course code
course_name	C	50		CHAR	50	Course name in full for Avondale
credit_lev1_sem1	N	3	1	ZONED-U	3	Standard course credit points for level 1, semester 1
credit_lev1_sem2	N	3	1	ZONED-U	3	Standard course credit points for level 1, semester 2
credit_lev2_sem1	N	3	1	ZONED-U	3	Standard course credit points for level 2, semester 1
credit_lev2_sem2	N	3	1	ZONED-U	3	Standard course credit points for level 2, semester 2
credit_lev3_sem1	N	3	1	ZONED-U	3	Standard course credit points for level 3, semester 1
credit_lev3_sem2	N	3	1	ZONED-U	3	Standard course credit points for level 3, semester 2
credit_lev4_sem1	N	3	1	ZONED-U	3	Standard course credit points for level 4, semester 1
credit_lev4_sem2	N	3	1	ZONED-U	3	Standard course credit points for level 4, semester 2
year_first_offered	N	4		ZONED-U	4	Year (YYYY) this cours was first offered
year_last_offered	N	4		ZONED-U	4	Year (YYYY) this course last offered
currently_offered	C	1		CHAR	1	Y=course is currently offered
accredited	C	1		CHAR	1	Is the course accredited?
govt_funded	C	1		CHAR	1	Indicates if this is a Government funded course
misc_course	C	1		CHAR	1	This is a miscellaneous course
edu_type	C	1		CHAR	1	P=Primary, S=Secondary, N=Non-teaching
course_name_full	C	72		CHAR	72	E308: Course name - full
course_name_abbrev	C	30		CHAR	30	E309: Course name - abbreviated
aou	C	3		CHAR	3	E333: Academic organisational unit code
fos	C	6		CHAR	6	E311: Field of study code
major_code	C	10		CHAR	10	Standard code for major
course_type	N	2		ZONED-U	2	E310: Course type code
course_load	N	2		ZONED-U	2	E350: Course load

PRIMARY KEY: course_key : year_yyyy
: course_code

SR_CODE_SUBJECT

Subject codes, descriptions and academic details

Item	Type	Size	Dec			Description
year_yyyy	C	4		CHAR	4	Year (YYYY) for this record
semester	C	1		CHAR	1	E353: Semester in which the load occurs code
subject_code	C	5		CHAR	5	Alphanumeric code assigned to this subject
subject_name	C	50		CHAR	50	Full description of this subject
subject_name_abbrev	C	15		CHAR	15	Abbreviated name of this subject
subject_credit_points	N	2	1	ZONED-U	2	Number of credit points, for academic credit and fees
subject_type	C	1		CHAR	1	Degree/Certificate type of subject
discipline_group	C	4		CHAR	4	E336: Discipline group code
campus_taught	C	1		CHAR	1	C=Cooranbong, W=Wahroonga
department	N	3		INTEGER-S	2	Avondale department code
equivalent_subject	C	5		CHAR	5	Other name for this subject (Same as subject code if none)
offered	C	1		CHAR	1	This course is currently offered
number_attended	N	8		INTEGER-S	4	Total number of students attending this subject
number_grades	N	8		INTEGER-S	4	Total number of numeric grades for this subject
sum_grades	N	9		INTEGER-S	4	Sum of all numeric grades for this subject
sum_grades_squared	N	12		INTEGER-S	8	Sub of the squares of all numeric grades

PRIMARY KEY: subject_key : year_yyyy
: semester
: subject_code

SR_STUDENT_COURSE

Course related records for students

Item	Type	Size	Dec			Description
name_number	C	6		CHAR	6	Computerised record number assigned to this person
course_record_number	C	3		CHAR	3	Pointer for course record
year_yyyy	C	4		CHAR	4	Year (YYYY) for this record
semester	C	1		CHAR	1	Semester for this record
course_code	C	5		CHAR	5	Course code
current_level	N	1		ZONED-U	1	Current level in course
current_semester	N	1		ZONED-U	1	Current semester in course
attendance_type	N	1		ZONED-U	1	Full-time or Part-time
date_commenced	C	6		CHAR	6	Date this student first commenced this course
date_recommenced	C	6		CHAR	6	Date this student recommenced this course
date_applic_rcvd	C	6		CHAR	6	Date application was received
date_prov_accept	C	6		CHAR	6	Date provisional acceptance was given
date_acad_prob	C	6		CHAR	6	Date that academic probation status was given
acad_prob_code	N	1		ZONED-U	1	Reason for granting Academic Probation Status
date_full_accept	C	6		CHAR	6	Date full acceptance was given
date_suppl_rcvd	C	6		CHAR	6	Date supplimentary form was received
date_registered	C	6		CHAR	6	Date this student registered
date_terminated	C	6		CHAR	6	Date this student ceased being enrolled in this course
termination_code	N	1		ZONED-U	1	Code giving reason why student is no longer doing course
graduating	C	1		CHAR	1	Y=Student graduates in this year
acad_award	C	1		CHAR	1	Academic Excellence Award given on graduating
in_absentia	C	1		CHAR	1	Graduated in absentia
major	C	10		CHAR	10	Standard code for major
minor	C	10		CHAR	10	Standard code for minor
other_field	C	10		CHAR	10	Study field in addition to 1st major and minor
handbook_year	C	4		CHAR	4	This student course is based on the Handbook for this year
name_record_number	N	3		ZONED-U	3	Name history record number of person's name for this crse

PRIMARY KEY: course_key : name_number
: course_record_number

SR_STUDENT_PERSONAL

Personal details, etc about registered students

Item	Type	Size	Dec		Description
ref_year	C	4	CHAR	4	Reference year for this record
name_number	C	6	CHAR	6	Computerised record number assigned to this person
name_code	C	6	CHAR	6	First 4 letters of surname + first 2 letters of given name
student_account	C	8	CHAR	8	Student ledger account number
contact_address_code	N	1	ZONED-U	1	Address code holding this person's contact address
fees_address_code	N	1	ZONED-U	1	Address code of address record containing fees address
kin_address_code	N	1	ZONED-U	1	Address number of next of kin address
vac_address_code	N	1	ZONED-U	1	Address number of vacation address
baptised	C	1	CHAR	1	B=Baptised SDA, U=Unbaptised SDA, O=Other
sda_church_mship	C	20	CHAR	20	SDA church where membership is held
conference	C	5	CHAR	5	SDA conference of this person's home church
student_religion	C	15	CHAR	15	Religion of this student
religion_father	C	15	CHAR	15	Religion of student's father
religion_mother	C	15	CHAR	15	Religion of student's mother
dorm_room_code	C	4	CHAR	4	Room code of residence hall room allocated to this student
prior_qualification	N	1	ZONED-U	1	Highest qualification prior to application for enrolment
year_achieved	C	4	CHAR	4	Year (YYYY) in which prior qualification was achieved
secondary_institution	C	30	CHAR	30	Institution where prior qualification was received
advanced_standing	C	1	CHAR	1	Is advanced standing being sought?
instructed_english	C	1	CHAR	1	Prior education was given in English
starting_semester	N	1	ZONED-U	1	Indicates in which semester (0/1/2) student will commence
first_aid_cert	C	1	CHAR	1	Indicates if student has a first aid certificate
typing_competency	C	1	CHAR	1	Indicates whether typing competency has been achieved
tert_ent_score	N	1	ZONED-U	1	E369: Tertiary entrance score
te_score_nsw	N	1	ZONED-U	1	NSW equivalent Tertiary Entrance Ranking
te_score_source	C	1	CHAR	1	State where TE score was obtained
application_type	N	1	ZONED-U	1	=New student, 2=Returning, 3=Returning after absence
nz_non_res_fee	C	1	CHAR	1	This student is a NZ non-residential for fee payments
fees_title	C	4	CHAR	4	Title of fee payer
fees_surname	C	20	CHAR	20	Surname of fee payer
fees_initials	C	5	CHAR	5	Fee payer's initials
fees_business_phone	C	15	CHAR	15	Fee payer's business phone number
fees_relationship	C	1	CHAR	1	Relationship of fee payer to student
kin_title	C	4	CHAR	4	Title of next of kin
kin_surname	C	20	CHAR	20	Surname of next of kin
kin_given_names	C	30	CHAR	30	Full given names of next of kin
kin_initials	C	5	CHAR	5	Initials of next of kin
kin_business_phone	C	15	CHAR	15	Telephone number for next of kin
kin_occupation	C	30	CHAR	30	Occupation of next of kin
aborig_torres	C	1	CHAR	1	E316: Aboriginal/Torres Strait Islander code
admission	C	2	CHAR	2	E327: Basis for admission to current course
citizen_resident	C	1	CHAR	1	E358: Citizen/resident indicator
permanent_resident	C	1	CHAR	1	E390: Permanent resident status
country_birth	C	4	CHAR	4	E346: Country of birth code
year_arrival	C	4	CHAR	4	E347: Year of arrival in Australia
fee_student	C	1	CHAR	1	E349: Fee paying student indicator
home_location	C	4	CHAR	4	E320: Location code of permanent home residence
language_home	C	2	CHAR	2	E348: Language spoken at home indicator
term_location	C	4	CHAR	4	E319: Location code of semester/term residence
disability	C	8	CHAR	8	E386: Indicates responses to 3 questions about disabilities

PRIMARY KEY: year_name_number : ref_year
: name_number

SR_STUDENT_PERSONAL_SEM

Semester-specific personal details, etc about registered students

Item	Type	Size	Dec		Description
ref_year	C	4	CHAR	4	Reference year for this record
ref_semester	C	1	CHAR	1	Reference semester for this record
name_number	C	6	CHAR	6	Computerised record number assigned to this person
marital_status	C	1	CHAR	1	Marital status (single, married, etc)
campus	C	1	CHAR	1	C=Cooranbong, W=Wahroonga
residence_type	C	1	CHAR	1	Day or residential
single_room_request	C	1	CHAR	1	Student is requesting a single room
course_record_number	C	3	CHAR	3	Pointer to semester 1 course record
cert_credit	N	1	ZONED-U	1	Total certificate credit points taken
dip_coor_credit	N	1	ZONED-U	1	Credit points for Cooranbong degree subjects
dip_wah_credit	N	1	ZONED-U	1	Credit points for Wahroonga degree subjects
ma_credit	N	1	ZONED-U	1	Total Masters credit points taken
eftsu	N	1	ZONED-U	1	Equivalent full-time student unit
late_application	C	1	CHAR	1	Application was late
late_registration	C	1	CHAR	1	Student registered late
application_status	N	1	ZONED-U	1	Application and supplementary form received, etc
acceptance_status	N	1	ZONED-U	1	Stage of acceptance or otherwise
student_status	N	1	ZONED-U	1	Registration status (registered, on leave, withdrawn, etc)
fin_status	C	1	CHAR	1	Financial status
dependent_disc_flag	C	1	CHAR	1	Override flag for denominational employee discount
name_number_sibling	C	6	CHAR	6	NAME_NUMBER of a sibling (for family discount)
bridging_course	C	1	CHAR	1	Y=taking chemistry bridging course
acad_bridging_course	C	1	CHAR	1	Bridging course required for academic reasons
college_sship	C	1	CHAR	1	Y=college scholarship available
bursary_provider	C	20	CHAR	20	Name of provider of this student bursary
bursary_amt	N	1	ZONED-U	1	Amount of bursary granted to this student
overseas_sship_code	N	1	ZONED-U	1	Override code for overseas scholarship
hecs_exmt_status	C	2	CHAR	2	E380: HECS exemption status
hecs_amt_paid	N	1	ZONED-U	1	E381: HECS amount paid - semester
hecs_prexmt_tot	N	1	ZONED-U	1	E382: HECS amount prior to exemption - total

PRIMARY KEY: year_sem_name_number : ref_year
: ref_semester
: name_number

SR_STUDENT_SUBJECT

Permanent record of student subjects taken, with results

Item	Type	Size	Dec		Description
name_number	C	6	CHAR	6	Computerised record number assigned to this person
year_yyyy	C	4	CHAR	4	Year (YYYY) for this record
semester	C	1	CHAR	1	E353: Semester in which the load occurs code
subject_code	C	5	CHAR	5	Alphanumeric code assigned to this subject
attendance_status	C	1	CHAR	1	Type of attendance in this class - eg, Auditing, Normal
date_started	C	4	CHAR	4	Date (DD/MM) student commenced this subject
completion_status	C	1	CHAR	1	Final status of this subject for this student
date_sub_terminated	C	4	CHAR	4	Date (DD/MM) student quit doing this subject
grade	C	3	CHAR	3	Grade assigned to a subject
unit_status	C	1	CHAR	1	E355: Unit of study completion status
checksheet_category	C	2	CHAR	2	Checksheet category to assign this advanced standing to
checksheet_subject	C	5	CHAR	5	Subject code being granted advanced standing status

PRIMARY KEY: subject_key : name_number
: year_yyyy
: semester
: subject_code

Appendix F: Data Warehouse Schema Definition

```
create domain YEAR_DOM char(4);
comment on domain YEAR_DOM is 'standard definition for year with century';

 CREATE TABLE DW_STUDENT
 (
    ref_year            YEAR_DOM,
    course_code         CHAR(5),
    student_id          CHAR(10),
    semester            CHAR(1),
    aou                 CHAR(3),
    course_name_abbrev  CHAR(30),
    graduating          CHAR(1),
    fos                 CHAR(6),
    date_of_birth       DATE,
    age_in_ref_year     INTEGER,
    sex                 CHAR(1),
    aborig_torres       CHAR(1),
    citizen_resident    CHAR(1),
    term_location       CHAR(4),
    home_location       CHAR(4),
    admission           CHAR(2),
    commencement_date   DATE,
    attendance_type     CHAR(1),
    country_birth       CHAR(4),
    language_home       CHAR(2),
    tert_ent_score      CHAR(3),
    total_eftsu_half    INTEGER,
    hecs_prexmt_tot     INTEGER,
    hecs_exmt_status    CHAR(2),
    hecs_amt_paid       INTEGER,

    PRIMARY KEY ( ref_year, student_id, course_code, semester ) deferrable);
 CREATE UNIQUE INDEX PKDW_STUDENT ON DW_STUDENT
    ( ref_year ASC, student_id ASC, course_code ASC, semester ASC );

 CREATE TABLE DW_SUBJECT
 (
    ref_year            YEAR_DOM,
    semester            CHAR(1),
    student_id          CHAR(10),
    course_code         CHAR(5),
    unit_study          CHAR(10),
    aou                 CHAR(3),
    discipline_group    CHAR(4),
    eftsu               INTEGER,
    unit_status         CHAR(1),

    PRIMARY KEY ( ref_year, semester, student_id, course_code, unit_study ) deferrable);
 CREATE UNIQUE INDEX PKDW_SUBJECT ON DW_SUBJECT
    ( ref_year ASC, semester ASC, student_id ASC, course_code ASC, unit_study ASC );
 CREATE TABLE DW_PP_COHORT
 (
    student_id         CHAR(10),
    course_code        CHAR(5),
    analysis_year      YEAR_DOM,
    semester           CHAR(1),
    course_name_abbrev CHAR(30),
    cohort_year        YEAR_DOM,
    age_in_anal_year   INTEGER,
    age_group          CHAR(7),
    sex                CHAR(1),
    aou                CHAR(3),
    course_type_name   CHAR(10),
    course_status      CHAR(10),
    spu_annual         INTEGER,
    spu_cumulative     INTEGER,
    headcount          INTEGER,
    eftsu              INTEGER,

    PRIMARY KEY (course_code, student_id, analysis_year, semester ) deferrable);
CREATE UNIQUE INDEX PKDW_PP_ANALYSIS ON DW_PP_COHORT
  (course_code, student_id, analysis_year, semester );
```

Appendix G: Data Dictionary - Elements Alphabetically (selected entries)

Item	Type	Size	Dec		Description	Location
aborig_torres	C	1	CHAR	1	E316: Aboriginal/Torres Strait Islander code	SR_STUDENT_PERSONAL
acad_award	C	1	CHAR	1	Academic Excellence Award given on graduating	SR_STUDENT_COURSE
acad_bridging_course	C	1	CHAR	1	Bridging course required for academic reasons	SR_STUDENT_PERSONAL_SEM
acad_prob_code	N	1	ZONED-U	1	Reason for granting Academic Probation Status	SR_STUDENT_COURSE
acceptance_status	N	1	ZONED-U	1	Stage of acceptance or otherwise	SR_STUDENT_PERSONAL_SEM
accredited	C	1	CHAR	1	Is the course accredited?	SR_CODE_COURSE_MASTER
address_number	C	1	CHAR	1	Address number of this address record	CI_ADDRESS
address1	C	30	CHAR	30	1st line of address, eg number and street	CI_ADDRESS
address2	C	30	CHAR	30	Line 2 of address (optional)	CI_ADDRESS
admission	C	2	CHAR	2	E327: Basis for admission to current course	SR_STUDENT_PERSONAL
advanced_standing	C	1	CHAR	1	Is advanced standing being sought?	SR_STUDENT_PERSONAL
aou	C	3	CHAR	3	E333: Academic organisational unit code	SR_CODE_COURSE_MASTER
application_status	N	1	ZONED-U	1	Application and supplementary form received, etc	SR_STUDENT_PERSONAL_SEM
application_type	N	1	ZONED-U	1	=New student, 2=Returning, 3=Returning after absence	SR_STUDENT_PERSONAL
attendance_status	C	1	CHAR	1	Type of attendance in this class - eg, Auditing, Normal	SR_STUDENT_SUBJECT
attendance_type	N	1	ZONED-U	1	Full-time or Part-time	SR_STUDENT_COURSE
baptised	C	1	CHAR	1	B=Baptised SDA, U=Unbaptised SDA, O=Other	SR_STUDENT_PERSONAL
bridging_course	C	1	CHAR	1	Y=taking chemistry bridging course	SR_STUDENT_PERSONAL_SEM
bursary_amt	N	1	ZONED-U	1	Amount of bursary granted to this student	SR_STUDENT_PERSONAL_SEM
bursary_provider	C	20	CHAR	20	Name of provider of this student bursary	SR_STUDENT_PERSONAL_SEM
campus	C	1	CHAR	1	C=Cooranbong, W=Wahroonga	SR_STUDENT_PERSONAL_SEM
campus_taught	C	1	CHAR	1	C=Cooranbong, W=Wahroonga	SR_CODE_SUBJECT
cert_credit	N	1	ZONED-U	1	Total certificate credit points taken	SR_STUDENT_PERSONAL_SEM
checksheet_category	C	2	CHAR	2	Checksheet category to assign this advanced standing to	SR_STUDENT_SUBJECT
checksheet_subject	C	5	CHAR	5	Subject code being granted advanced standing status	SR_STUDENT_SUBJECT
citizen_resident	C	1	CHAR	1	E358: Citizen/resident indicator	SR_STUDENT_PERSONAL
college_sship	C	1	CHAR	1	Y=college scholarship available	SR_STUDENT_PERSONAL_SEM
completion_status	C	1	CHAR	1	Final status of this subject for this student	SR_STUDENT_SUBJECT
conference	C	5	CHAR	5	SDA conference of this person's home church	SR_STUDENT_PERSONAL
contact_address_code	N	1	ZONED-U	1	Address code holding this person's contact address	SR_STUDENT_PERSONAL
country	C	20	CHAR	20	Country part of address	CI_ADDRESS
country_birth	C	4	CHAR	4	E346: Country of birth code	SR_STUDENT_PERSONAL
course_code	C	5	CHAR	5	Course code	SR_CODE_COURSE_MASTER
course_code	C	5	CHAR	5	Course code	SR_STUDENT_COURSE
course_load	N	2	ZONED-U	2	E350: Course load	SR_CODE_COURSE_MASTER

Item	Type	Size	Dec			Description	Location
course_name	C	50		CHAR	50	Course name in full for Avondale	SR_CODE_COURSE_MASTER
course_name_abbrev	C	30		CHAR	30	E309: Course name - abbreviated	SR_CODE_COURSE_MASTER
course_name_full	C	72		CHAR	72	E308: Course name - full	SR_CODE_COURSE_MASTER
course_record_number	C	3		CHAR	3	Pointer for course record	SR_STUDENT_COURSE
course_record_number	C	3		CHAR	3	Pointer to semester 1 course record	SR_STUDENT_PERSONAL_SEM
course_type	N	2		ZONED-U	2	E310: Course type code	SR_CODE_COURSE_MASTER
credit_lev1_sem1	N	3	1	ZONED-U	3	Standard course credit points for level 1, semester 1	SR_CODE_COURSE_MASTER
credit_lev1_sem2	N	3	1	ZONED-U	3	Standard course credit points for level 1, semester 2	SR_CODE_COURSE_MASTER
credit_lev2_sem1	N	3	1	ZONED-U	3	Standard course credit points for level 2, semester 1	SR_CODE_COURSE_MASTER
credit_lev2_sem2	N	3	1	ZONED-U	3	Standard course credit points for level 2, semester 2	SR_CODE_COURSE_MASTER
credit_lev3_sem1	N	3	1	ZONED-U	3	Standard course credit points for level 3, semester 1	SR_CODE_COURSE_MASTER
credit_lev3_sem2	N	3	1	ZONED-U	3	Standard course credit points for level 3, semester 2	SR_CODE_COURSE_MASTER
credit_lev4_sem1	N	3	1	ZONED-U	3	Standard course credit points for level 4, semester 1	SR_CODE_COURSE_MASTER
credit_lev4_sem2	N	3	1	ZONED-U	3	Standard course credit points for level 4, semester 2	SR_CODE_COURSE_MASTER
current_level	N	1		ZONED-U	1	Current level in course	SR_STUDENT_COURSE
current_semester	N	1		ZONED-U	1	Current semester in course	SR_STUDENT_COURSE
currently_offered	C	1		CHAR	1	Y=course is currently offered	SR_CODE_COURSE_MASTER
date_acad_prob	C	6		CHAR	6	Date that academic probation status was given	SR_STUDENT_COURSE
date_address_modified	D	8		ZONED-U	8	Date this address was modified	CI_ADDRESS
date_applic_rcvd	C	6		CHAR	6	Date application was received	SR_STUDENT_COURSE
date_commenced	C	6		CHAR	6	Date this student first commenced this course	SR_STUDENT_COURSE
date_full_accept	C	6		CHAR	6	Date full acceptance was given	SR_STUDENT_COURSE
date_prov_accept	C	6		CHAR	6	Date provisional acceptance was given	SR_STUDENT_COURSE
date_recommenced	C	6		CHAR	6	Date this student recommenced this course	SR_STUDENT_COURSE
date_registered	C	6		CHAR	6	Date this student registered	SR_STUDENT_COURSE
date_started	C	4		CHAR	4	Date (DD/MM) student commenced this subject	SR_STUDENT_SUBJECT
date_sub_terminated	C	4		CHAR	4	Date (DD/MM) student quit doing this subject	SR_STUDENT_SUBJECT
date_suppl_rcvd	C	6		CHAR	6	Date supplimentary form was received	SR_STUDENT_COURSE
date_terminated	C	6		CHAR	6	Date this student ceased being enrolled in this course	SR_STUDENT_COURSE
department	N	3		INTEGER-S	2	Avondale department code	SR_CODE_SUBJECT
dependent_disc_flag	C	1		CHAR	1	Override flag for denominational employee discount	SR_STUDENT_PERSONAL_SEM
dip_coor_credit	N	1		ZONED-U	1	Credit points for Cooranbong degree subjects	SR_STUDENT_PERSONAL_SEM
dip_wah_credit	N	1		ZONED-U	1	Credit points for Wahroonga degree subjects	SR_STUDENT_PERSONAL_SEM
disability	C	8		CHAR	8	E386: Indicates responses to 3 questions about disabilities	SR_STUDENT_PERSONAL
discipline_group	C	4		CHAR	4	E336: Discipline group code	SR_CODE_SUBJECT
dorm_room_code	C	4		CHAR	4	Room code of residence hall room allocated to this student	SR_STUDENT_PERSONAL
edu_type	C	1		CHAR	1	P=Primary, S=Secondary, N=Non-teaching	SR_CODE_COURSE_MASTER
eftsu	N	1		ZONED-U	1	Equivalent full-time student unit	SR_STUDENT_PERSONAL_SEM

Item	Type	Size	Dec			Description	Location
equivalent_subject	C	5		CHAR	5	Other name for this subject (Same as subject code if none)	SR_CODE_SUBJECT
fee_student	C	1		CHAR	1	E349: Fee paying student indicator	SR_STUDENT_PERSONAL
fees_address_code	N	1		ZONED-U	1	Address code of address record containing fees address	SR_STUDENT_PERSONAL
fees_business_phone	C	15		CHAR	15	Fee payer's business phone number	SR_STUDENT_PERSONAL
fees_initials	C	5		CHAR	5	Fee payer's initials	SR_STUDENT_PERSONAL
fees_relationship	C	1		CHAR	1	Relationship of fee payer to student	SR_STUDENT_PERSONAL
fees_surname	C	20		CHAR	20	Surname of fee payer	SR_STUDENT_PERSONAL
fees_title	C	4		CHAR	4	Title of fee payer	SR_STUDENT_PERSONAL
fin_status	C	1		CHAR	1	Financial status	SR_STUDENT_PERSONAL_SEM
first_aid_cert	C	1		CHAR	1	Indicates if student has a first aid certificate	SR_STUDENT_PERSONAL
fos	C	6		CHAR	6	E311: Field of study code	SR_CODE_COURSE_MASTER
good_address	C	1		CHAR	1	Indicates the validity of this address	CI_ADDRESS
govt_funded	C	1		CHAR	1	Indicates if this is a Government funded course	SR_CODE_COURSE_MASTER
grade	C	3		CHAR	3	Grade assigned to a subject	SR_STUDENT_SUBJECT
graduating	C	1		CHAR	1	Y=Student graduates in this year	SR_STUDENT_COURSE
handbook_year	C	4		CHAR	4	This student course is based on the Handbook for this year	SR_STUDENT_COURSE
hecs_amt_paid	N	1		ZONED-U	1	E381: HECS amount paid - semester	SR_STUDENT_PERSONAL_SEM
hecs_exmt_status	C	2		CHAR	2	E380: HECS exemption status	SR_STUDENT_PERSONAL_SEM
hecs_prexmt_tot	N	1		ZONED-U	1	E382: HECS amount prior to exemption - total	SR_STUDENT_PERSONAL_SEM
home_location	C	4		CHAR	4	E320: Location code of permanent home residence	SR_STUDENT_PERSONAL
in_absentia	C	1		CHAR	1	Graduated in absentia	SR_STUDENT_COURSE
instructed_english	C	1		CHAR	1	Prior education was given in English	SR_STUDENT_PERSONAL
kin_address_code	N	1		ZONED-U	1	Address number of next of kin address	SR_STUDENT_PERSONAL
kin_business_phone	C	15		CHAR	15	Telephone number for next of kin	SR_STUDENT_PERSONAL
kin_given_names	C	30		CHAR	30	Full given names of next of kin	SR_STUDENT_PERSONAL
kin_initials	C	5		CHAR	5	Initials of next of kin	SR_STUDENT_PERSONAL
kin_occupation	C	30		CHAR	30	Occupation of next of kin	SR_STUDENT_PERSONAL
kin_surname	C	20		CHAR	20	Surname of next of kin	SR_STUDENT_PERSONAL
kin_title	C	4		CHAR	4	Title of next of kin	SR_STUDENT_PERSONAL
language_home	C	2		CHAR	2	E348: Language spoken at home indicator	SR_STUDENT_PERSONAL
late_application	C	1		CHAR	1	Application was late	SR_STUDENT_PERSONAL_SEM
late_registration	C	1		CHAR	1	Student registered late	SR_STUDENT_PERSONAL_SEM
ma_credit	N	1		ZONED-U	1	Total Masters credit points taken	SR_STUDENT_PERSONAL_SEM
major	C	10		CHAR	10	Standard code for major	SR_STUDENT_COURSE
major_code	C	10		CHAR	10	Standard code for major	SR_CODE_COURSE_MASTER
marital_status	C	1		CHAR	1	Marital status (single, married, etc)	SR_STUDENT_PERSONAL_SEM
minor	C	10		CHAR	10	Standard code for minor	SR_STUDENT_COURSE
misc_course	C	1		CHAR	1	This is a miscellaneous course	SR_CODE_COURSE_MASTER

Item	Type	Size	Dec			Description	Location
name_code	C	6		CHAR	6	First 4 letters of surname + first 2 letters of given name	SR_STUDENT_PERSONAL
name_number	C	6		CHAR	6	Computerised record number assigned to this person	CI_ADDRESS
name_number	C	6		CHAR	6	Computerised record number assigned to this person	SR_STUDENT_COURSE
name_number	C	6		CHAR	6	Computerised record number assigned to this person	SR_STUDENT_PERSONAL
name_number	C	6		CHAR	6	Computerised record number assigned to this person	SR_STUDENT_PERSONAL_SEM
name_number	C	6		CHAR	6	Computerised record number assigned to this person	SR_STUDENT_SUBJECT
name_number_sibling	C	6		CHAR	6	NAME_NUMBER of a sibling (for family discount)	SR_STUDENT_PERSONAL_SEM
name_record_number	N	3		ZONED-U	3	Name history record number of person's name for this course	SR_STUDENT_COURSE
number_attended	N	8		INTEGER-S	4	Total number of students attending this subject	SR_CODE_SUBJECT
number_grades	N	8		INTEGER-S	4	Total number of numeric grades for this subject	SR_CODE_SUBJECT
nz_non_res_fee	C	1		CHAR	1	This student is a NZ non-residential for fee payments	SR_STUDENT_PERSONAL
offered	C	1		CHAR	1	This course is currently offered	SR_CODE_SUBJECT
other_field	C	10		CHAR	10	Study field in addition to 1st major and minor	SR_STUDENT_COURSE
overseas_sship_code	N	1		ZONED-U	1	Override code for overseas scholarship	SR_STUDENT_PERSONAL_SEM
permanent_resident	C	1		CHAR	1	E390: Permanent resident status	SR_STUDENT_PERSONAL
phone	C	15		CHAR	15	Telephone number part of person address	CI_ADDRESS
postcode	C	7		CHAR	7	Post code	CI_ADDRESS
prior_qualification	N	1		ZONED-U	1	Highest qualification prior to application for enrollment	SR_STUDENT_PERSONAL
ref_semester	C	1		CHAR	1	Reference semester for this record	SR_STUDENT_PERSONAL_SEM
ref_year	C	4		CHAR	4	Reference year for this record	SR_STUDENT_PERSONAL
ref_year	C	4		CHAR	4	Reference year for this record	SR_STUDENT_PERSONAL_SEM
religion_father	C	15		CHAR	15	Religion of student's father	SR_STUDENT_PERSONAL
religion_mother	C	15		CHAR	15	Religion of student's mother	SR_STUDENT_PERSONAL
residence_type	C	1		CHAR	1	Day or residential	SR_STUDENT_PERSONAL_SEM
sda_church_mship	C	20		CHAR	20	SDA church where membership is held	SR_STUDENT_PERSONAL
secondary_institution	C	30		CHAR	30	Institution where prior qualification was received	SR_STUDENT_PERSONAL
semester	C	1		CHAR	1	E353: Semester in which the load occurs code	SR_CODE_SUBJECT
semester	C	1		CHAR	1	E353: Semester in which the load occurs code	SR_STUDENT_SUBJECT
semester	C	1		CHAR	1	Semester for this record	SR_STUDENT_COURSE
single_room_request	C	1		CHAR	1	Student is requesting a single room	SR_STUDENT_PERSONAL_SEM
starting_semester	N	1		ZONED-U	1	Indicates in which semester (0/1/2) student will commence	SR_STUDENT_PERSONAL
state	C	10		CHAR	10	State part of address	CI_ADDRESS
student_account	C	8		CHAR	8	Student ledger account number	SR_STUDENT_PERSONAL
student_religion	C	15		CHAR	15	Religion of this student	SR_STUDENT_PERSONAL
student_status	N	1		ZONED-U	1	Registration status (registered, on leave, withdrawn, etc)	SR_STUDENT_PERSONAL_SEM
subject_code	C	5		CHAR	5	Alphanumeric code assigned to this subject	SR_CODE_SUBJECT
subject_code	C	5		CHAR	5	Alphanumeric code assigned to this subject	SR_STUDENT_SUBJECT
subject_credit_points	N	2	1	ZONED-U	2	Number of credit points, for academic credit and fees	SR_CODE_SUBJECT

Item	Type	Size	Dec			Description	Location
subject_name	C	50		CHAR	50	Full description of this subject	SR_CODE_SUBJECT
subject_name_abbrev	C	15		CHAR	15	Abbreviated name of this subject	SR_CODE_SUBJECT
subject_type	C	1		CHAR	1	Degree/Certificate type of subject	SR_CODE_SUBJECT
sum_grades	N	9		INTEGER-S	4	Sum of all numeric grades for this subject	SR_CODE_SUBJECT
sum_grades_squared	N	12		INTEGER-S	8	Sub of the squares of all numeric grades	SR_CODE_SUBJECT
te_score_nsw	N	1		ZONED-U	1	NSW equivalent Tertiary Entrance Ranking	SR_STUDENT_PERSONAL
te_score_source	C	1		CHAR	1	State where TE score was obtained	SR_STUDENT_PERSONAL
term_location	C	4		CHAR	4	E319: Location code of semester/term residence	SR_STUDENT_PERSONAL
termination_code	N	1		ZONED-U	1	Code giving reason why student is no longer doing course	SR_STUDENT_COURSE
tert_ent_score	N	1		ZONED-U	1	E369: Tertiary entrance score	SR_STUDENT_PERSONAL
town	C	30		CHAR	30	Town or suburb	CI_ADDRESS
typing_competency	C	1		CHAR	1	Indicates whether typing competency has been achieved	SR_STUDENT_PERSONAL
unit_status	C	1		CHAR	1	E355: Unit of study completion status	SR_STUDENT_SUBJECT
vac_address_code	N	1		ZONED-U	1	Address number of vacation address	SR_STUDENT_PERSONAL
year_achieved	C	4		CHAR	4	Year (YYYY) in which prior qualification was achieved	SR_STUDENT_PERSONAL
year_arrival	C	4		CHAR	4	E347: Year of arrival in Australia	SR_STUDENT_PERSONAL
year_first_offered	N	4		ZONED-U	4	Year (YYYY) this course was first offered	SR_CODE_COURSE_MASTER
year_last_offered	N	4		ZONED-U	4	Year (YYYY) this course last offered	SR_CODE_COURSE_MASTER
year_yyyy	C	4		CHAR	4	Year (YYYY) for this record	SR_CODE_COURSE_MASTER
year_yyyy	C	4		CHAR	4	Year (YYYY) for this record	SR_CODE_SUBJECT
year_yyyy	C	4		CHAR	4	Year (YYYY) for this record	SR_STUDENT_COURSE
year_yyyy	C	4		CHAR	4	Year (YYYY) for this record	SR_STUDENT_SUBJECT

Appendix H: Review Questionnaire & Responses

Note: Responses to questions are given in italics.

1. The EIS/Data Warehousing concept

1. Were the concepts sufficiently well presented?

"I thought concepts were reasonably well presented."

2. How did management regard the importance of the project?

"It was probably not regarded with sufficient importance by Senior Management. The middle levels would probably be more involved in the actual use of the system - at least if we had a broader middle level. We are too small."

3. What level of commitment did management give to the project?

2. Needs Analysis

1. Was the process used to determine needs effective?

"Analysis of needs was OK but probably idealistic. The pressures are such just coping with other things that taking time to work with the EIS and utilize the benefits at least for me was not and I doubt will be time-efficient."

2. Was it thorough, detailed enough?
3. Comments on what was done well, or what could have been done better.

3. Design & construction

Was there adequate:

1. user involvement,
2. feedback, communication, progress reports
3. consulting over issues,
4. demonstration of prototypes, etc?

4. Training

1. Was training relevant, appropriate, timely, useful?

"Training was unfortunately irrelevant and at the wrong time. I think the day we spent trying to learn the software was wasted. This part of the exercise should have come after the data warehouse was complete and in place so that we could get an introduction to its use and then move straight into its use."

5. Consulting

1. Please comment on your impressions value of the consultants to the project, from your own experiences with them.

"I thought the consulting process was adequate but we probably did not have a clear strategic plan in place to help us zero in on the critical performance indicators. We are still wrestling with the problem of trying to develop a strategic plan."

6. Desktop tools

1. Impromptu - will you regularly find a use for this product to write your own ad hoc queries and reports?
2. PowerPlay - will you regularly use this tool to answer questions that support decision making?
3. Please comment on these tools in terms of ease of use, ease of learning to use, power and functionality.

"I have used neither Impromptu nor PowerPlay since the training day."

7. Impact

1. Please describe your understanding of the impact that the pilot project is having and will have on decision making at Avondale that it affects.
2. Estimate the impact the data warehouse will have when fully implemented.
3. Describe the kinds of changes you think these 'management improvement' technologies will suggest and facilitate.

(ie, to do something better, you will do it differently)

"No impact yet."

8. The Future

1. What subject areas do you think should be the focus of the next level of development after the pilot project is complete?

"If I am to benefit from the data warehouse I will need training again."

Index