Data Warehousing

at

Avondale College

Presented by: David Heise

Date: 29 November, 1996

DW1 Title Slide

Introduction

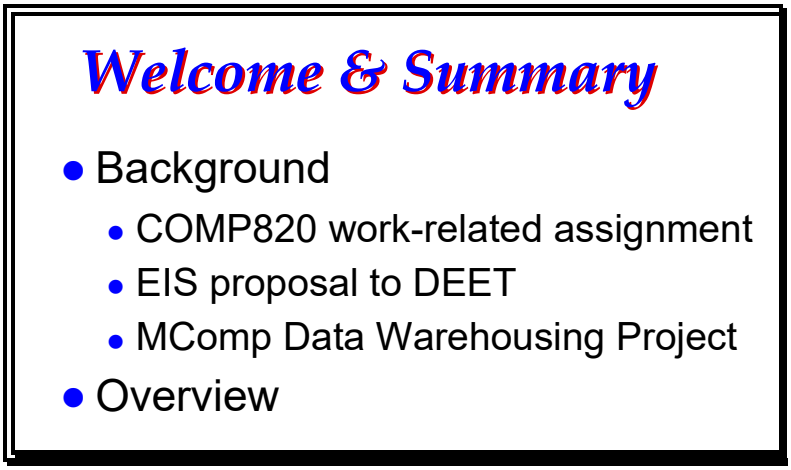
Hello Professor Dampney, fellow M. Comp. students, and others. My name is David Heise, and the topic of my presentation is Data Warehousing at Avondale College. When I began working towards this Master of Computing in 1994, I was employed as Director of Information Systems at Avondale College, a small independent university of about 6-700 students, 1½ hours drive north of Sydney. Since August this year, I have been working in a similar capacity at Andrews University in Michigan, USA. Andrews has an enrollment of about 3,000 students.

I trust you are all having a very pleasant and warm Friday afternoon in Sydney. Here in Michigan, it is quite pleasant indoors, but outside it is certainly not warm nor is it Friday afternoon. Temperatures outside are just above 0°C, and it is still only Thursday night here at present.

N, N

I am pleased to be able to tell a little about the Data Warehousing project being implemented at Avondale College

1:04



DW2 Welcome & Summary

To introduce the presentation, I will give a little background, and an overview of the project.

N **Background**

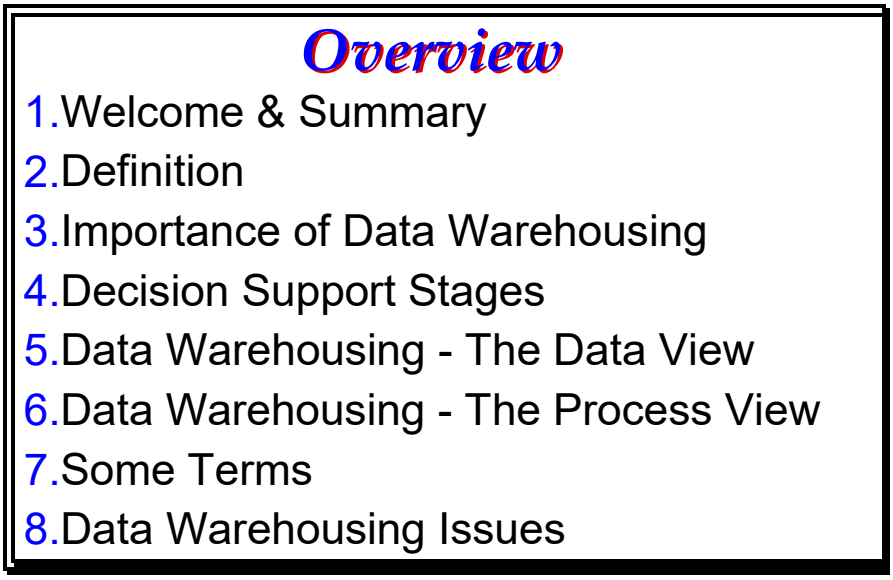
N
* The initial ideas for this Data Warehousing project started with a work-related assignment for COMP820 (Information Systems Management and Analysis), taken in second semester of 1994.

N
* These ideas were further developed in a successful proposal requesting a grant from the National Priorities (Reserve) Fund for a "management improvement" project. At this point, we were calling it an EIS Project.

N
* This management improvement project received further impetus when it was accepted as the project topic for my Master of Computing. By this time, EIS (Executive or Enterprise Information System) and DSS (Decision Support System) were seen as parts of a larger data management picture which has come to be known as Data Warehousing.

N I will complete this summary with an overview of the project.

1:03



DW3 Overview

I will start with a definition of data warehousing and give some of the reasons it is why it is considered to be such a vital part of the information revolution.

By describing briefly the stages of decision support from operational data processing through to intelligent agents and information discovery, I will place data warehousing in the context of data processing and information systems.

Then I will attempt to demonstrate what data warehousing is, from the points of view of both the data and the process. The terms described in point 7 encapsulate some of the fundamentals of data warehousing in terms of data storage and access technologies.

I will conclude with some observations on what we have learned about data warehousing issues at Avondale College.

0:57

N, N

Definition

This definition of a data warehouse is from Bill Inmon, the "father of data warehousing".

N
N
N
N
N, N

Definition

A data warehouse is a:

- Subject oriented
- Integrated
- Time-variant
- Nonvolatile

collection of data in support of management decisions.

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

DW4 Definition

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, beware of vendors who try to sell you a ***data warehouse***. A data warehouse is not a product. It is a model of ***your*** data, put together in such a way that it answers ***your*** business questions.

According to this definition, a data warehouse is different from an operational database in 4 important ways.

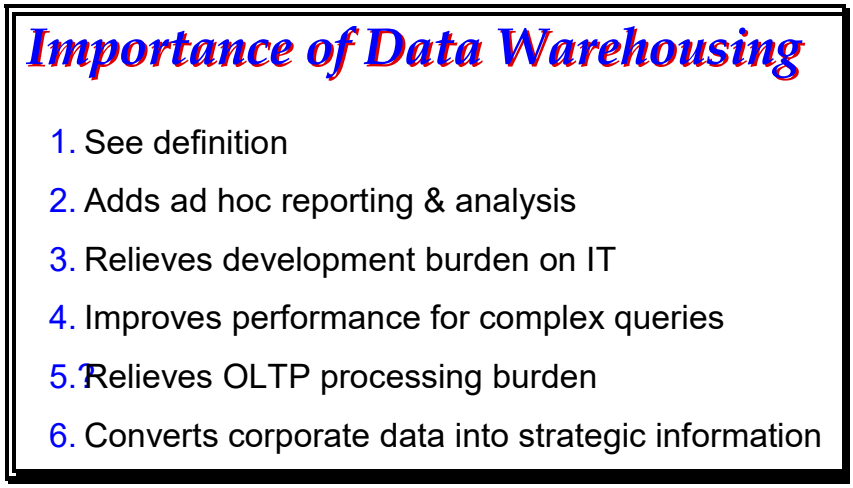
Data Warehouse	Operational Database
subject oriented	application oriented
integrated	multiple diverse sources
time-variant	real-time, current
nonvolatile	updateable

- It is application oriented
- Data comes from multiple diverse sources
- It holds current, real-time data values
- It is updateable

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

N, N

1:50



DW5 Importance of Data Warehousing

So is data warehousing important? Well the obvious answer is yes! But why? Why is data warehousing considered to be so vital? One of the reasons is that it is seen as part of the ***answer to information overload***.

P, P

But the definition itself gives some of the reasons why data warehousing is important.

It is **subject oriented**
Data is arranged by ***subject*** rather than by application, and is more intuitive for users to navigate.

It is **Integrated**
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 imposed by a data warehouse was seen as very appealing.

It is **Time-variant**
Data snapshots taken at times that are significant to the decision making process make it possible to analyze ***trends over time***. While it is true that operational systems maintain some history (such as permanent academic records), it is usually not sufficient for trend analysis except in very specific cases. 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.

N, N, N

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.

N

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.

N, N

However, analytical processing is different from online transaction processing in many ways. Optimization rules for large complex queries are not handled well by today's relational database management systems. Large analytical queries cause ***poor response time*** for the analysts, as well as severely impacting system performance for transaction processing functions. Separating the data warehouse from the operational data overcomes this problem.

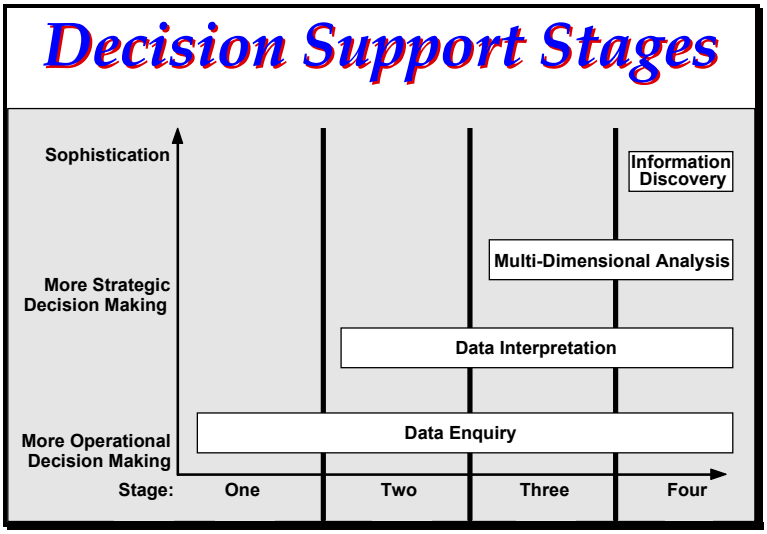
N

The end result is that corporate data becomes strategic information.

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.

N

3:20



DW6 Decision Support Stages

Each of the 4 levels of sophistication depicted in this diagram - Data Enquiry, Data Interpretation, Multidimensional Analysis and Information Discovery - corresponds to a different set of computer-based tools.

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

Going up through the levels, there are:

- traditional data processing applications
- 'point-and-click' reporting tools
- 'slice and dice', 'drill-down' tools
- intelligent agents

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.

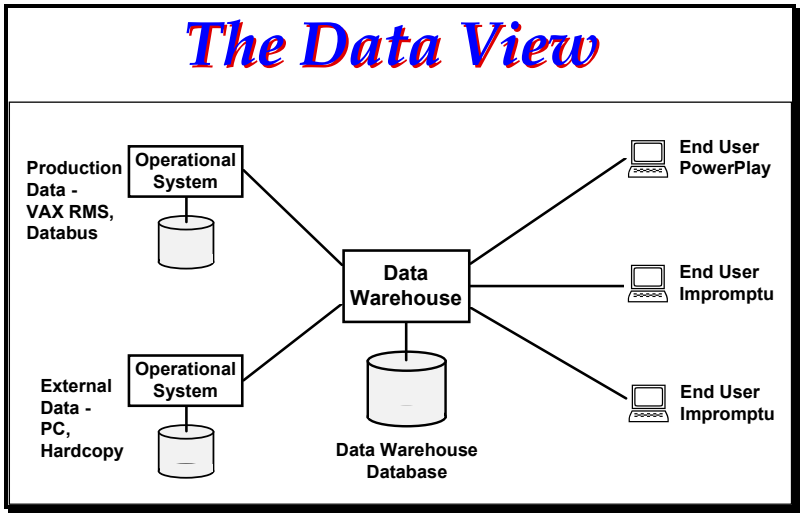
Users will be able to perform **data enquiry** using Impromptu without having to wait for special programs to be written by IT. As further questions are raised in response to this, users will be able to refine their questioning with further enquiries, and will be better able to **interpret** the data.

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

Notice 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.

N

1:51



DW7 Data Warehousing - The 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 in PowerPlay was quite different from one supporting knowledge workers writing reports in Impromptu.

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

1. coded data is replaced with values or meanings,
2. 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)

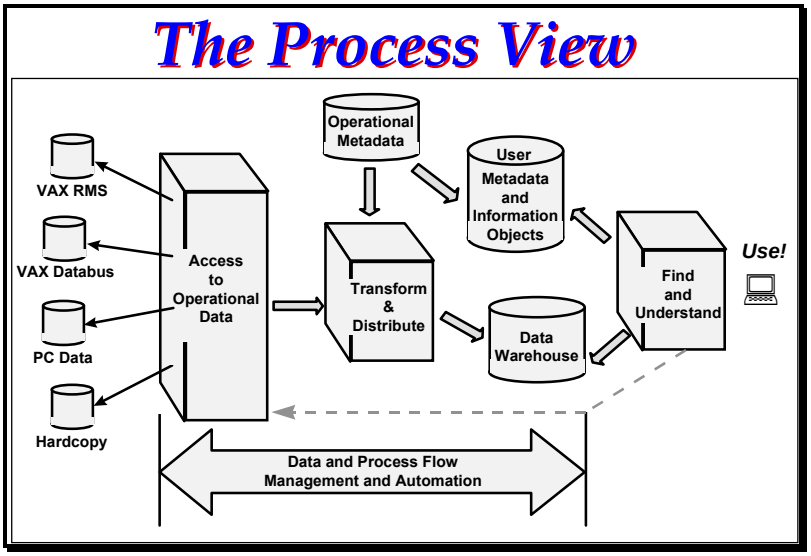
3. and thirdly, significant aggregation and summarization occurs

As a result, the data warehouse we are building consists of two parts:

1. there is an integrated relational view of all operational data, with history - for Impromptu
2. and there are summarized and pre-joined data to support multi-dimensional analysis - for PowerPlay

N

1:15



DW8 Data Warehousing - The Process View

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

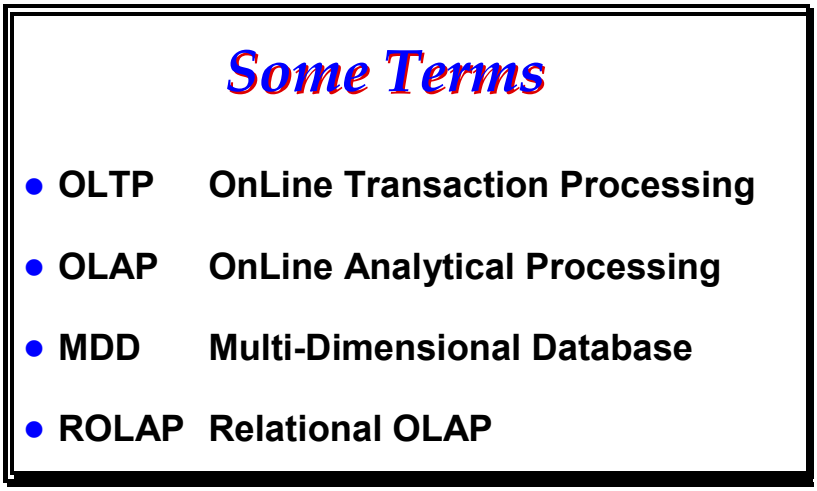
The proprietary multi-dimensional databases work well for smaller data warehouses or departmental 'data marts' up to 4 or 5 giga bytes. Using the more open Relational OLAP database products, data warehouses into the tera bytes in size have been built, running on multi-CPU platforms.

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

You start with:

1. talking to the users
2. then you 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. then you 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

N



DW9 Some Terms

These terms relate to storage and processing technologies.

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**.

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.

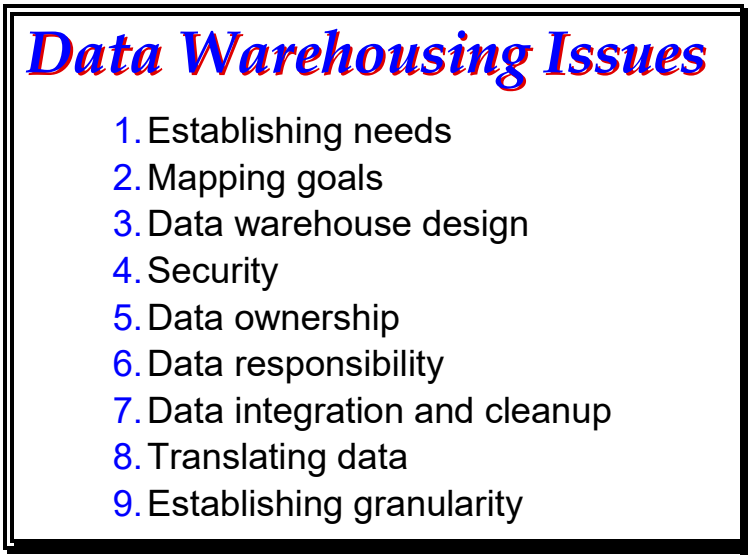
MDD Multi-Dimensional Database
MDD. An analysts view of the enterprises' universe is typically multi-dimensional in nature. For instance, they will want to analyze a given student population with regard to course, department, school, year in course, gender, age, etc. These could be some of the dimensions 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 Hyper cube - 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.

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.

N

2:55



DW10 Data Warehousing Issues

Many of the issues I will refer to here seem to relate pretty closely to steps in the process of building a data warehouse, and that is true. The actual list of issues or problem areas would vary in detail from one installation to another, although there would be some commonality.

N

1. 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.

N

2. Mapping goals to (measurable) performance indicators
The first so-called measures that were proposed were very vague, and in fact were not capable of being measured. As we got the hang of it, 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.

N

3. Data warehouse design
I have mentioned the concepts of denormalization and aggregation in a data warehouse previously, so I won't repeat them here.

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. 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.

N

4. 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.

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

N

5. Data ownership
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.

N

6. Data responsibility
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.

N

7. Data integration and cleanup
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                    PIC 9(10).
STUDENT_NUMBER                PIC 9(5).
NAME_NUMBER                   PIC X(6).
   (NAME_NUMBER replaces STUDENT_NUMBER plus CHECK_DIGIT)
```

Semantic Differences in STUDENT ID

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.

N

8. Translating data

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
```

Sample QTP source code

N

9. Establishing 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. 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. In larger data warehouses, change detection becomes vital, so that only altered records are refreshed. In the case of QTP, the single statement

```
OUTPUT <table name> IN DW UPDATE ADD
```

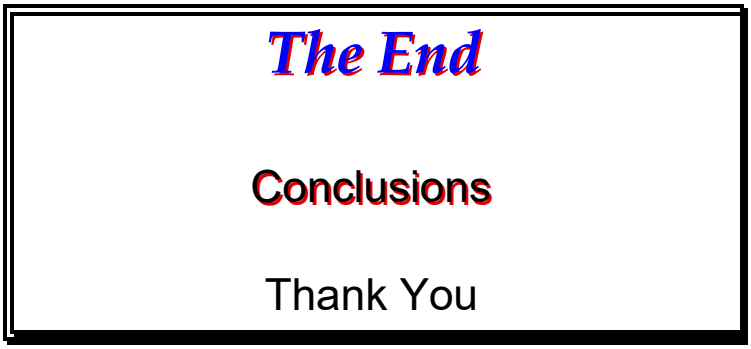
automatically adds new records and updates only records that have changed. No logic needs to be coded to read the old record and determine if a refresh is required. This means that the same program can be used to perform initial loads as well as performing the scheduled replications.

```
$ 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
```

Sample JCL for scheduling `daily.com`

N

6:25



DW11 Conclusions

It is now 4 months since I left Avondale College, but my involvement in the Avondale Data Warehouse Project has continued from a distance. I have moved to a new country and have had to learn a new job, and my successor at Avondale has also had to learn a new job, so the project understandably has lost a bit of momentum. We had hoped to be able to make some observations on the impact of the project on decision making processes at Avondale, by the end of October, but we were not ready in time to do that.

Nevertheless, we believe we can claim some credit for helping to focus attention on planning and decision making at Avondale. It is agreed that the models chosen for the pilot project will provide vital strategic information, and other areas are already being suggested 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.

N

0:57