

Data Warehousing

Aims

You should be able to:

- Describe different database models
- Describe the differences between OLAP and OLTP
- Describe the main characteristics of a data warehouse
- Critically evaluate an organisational situation and make appropriate recommendations

A brief overview of database models

Flat Files

In the early days of electronic database technology a simple Flat File was all that was available. For reasons that should be clear by now, such a model does not enable us to successfully model the complexities of “the real world”

Hierarchical Databases

These are organised in a tree structure where one parent can have many children and each child can only have one parent. This type of database is highly inflexible but can be very efficient in the processing of data. The links between records are physical links in the sense that references are made to the actual physical record numbers. Only one to many relations are possible with a hierarchical database.

One of the other main characteristics of a hierarchical database is that there is usually only a relatively small set of pre defined queries that can be applied to the data. This characteristic means that hierarchical databases are inflexible and wholly unsuited to ad hoc querying. In the example below it would be highly efficient to answer the question “What modules are Nigel Hartland’s tutees taking?”, but highly inefficient at answering “Which students are taking the Programming module?”



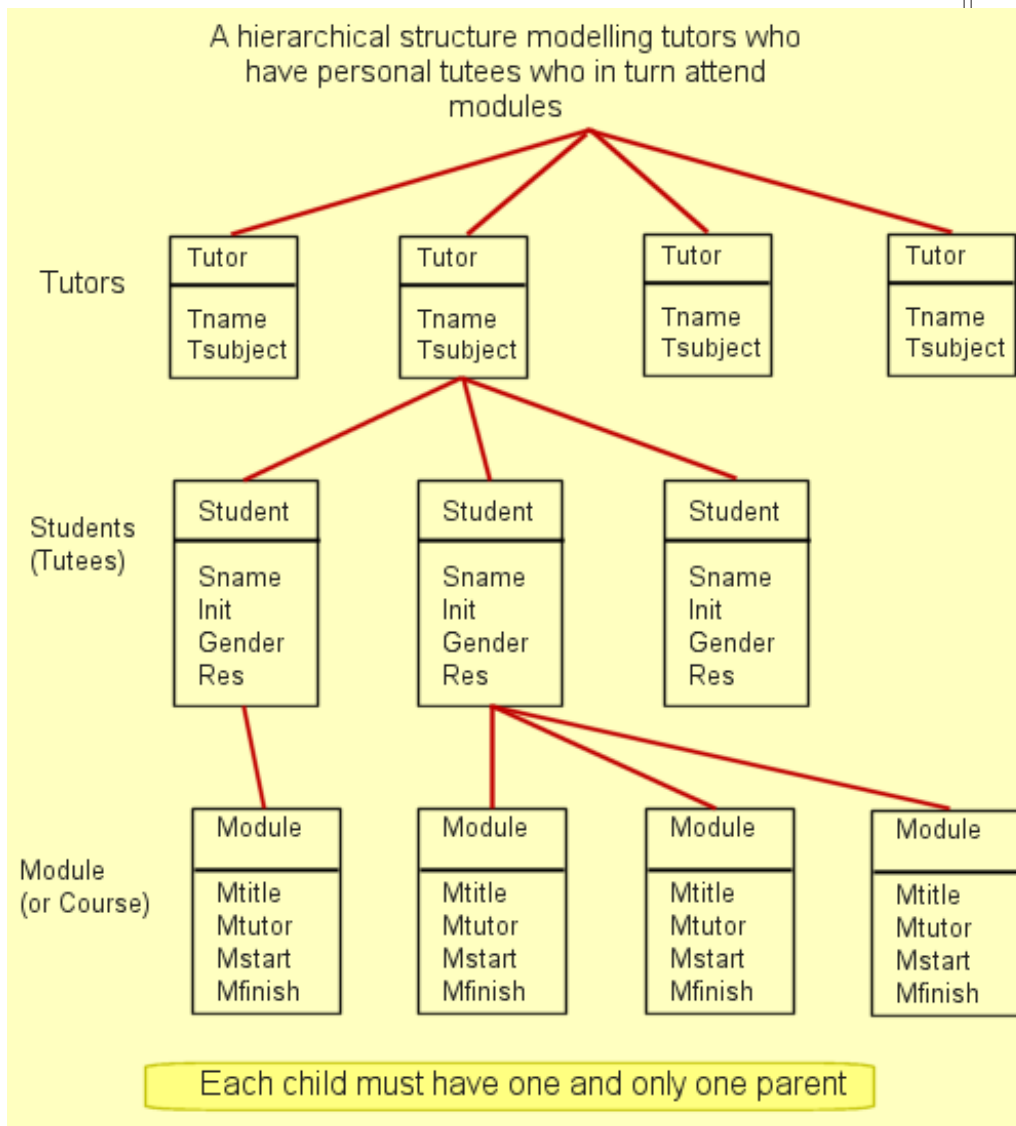


Figure 9.1: A Hierarchical Schema



Network Databases

Network databases were designed to address some of the issues of inflexibility of the earlier hierarchical databases. Each parent could have many children and each child could have many parents. Many to many relationships are possible with a network database but the links are still physical links. This model gives more flexibility of data access than the hierarchical model but at the cost of processing efficiency.

As can be seen from the figure 9.2, adding extra links from the tutor records to the module records will allow greater efficiency in answering the question “Which students are taking the Programming module?” We can identify which tutor teaches that module and then follow the links through to the module record. However, this flexibility does come at the price of great complexity and duplication of data.

A network structure modelling tutors who have students who in turn attend modules. This model also shows who teaches which modules

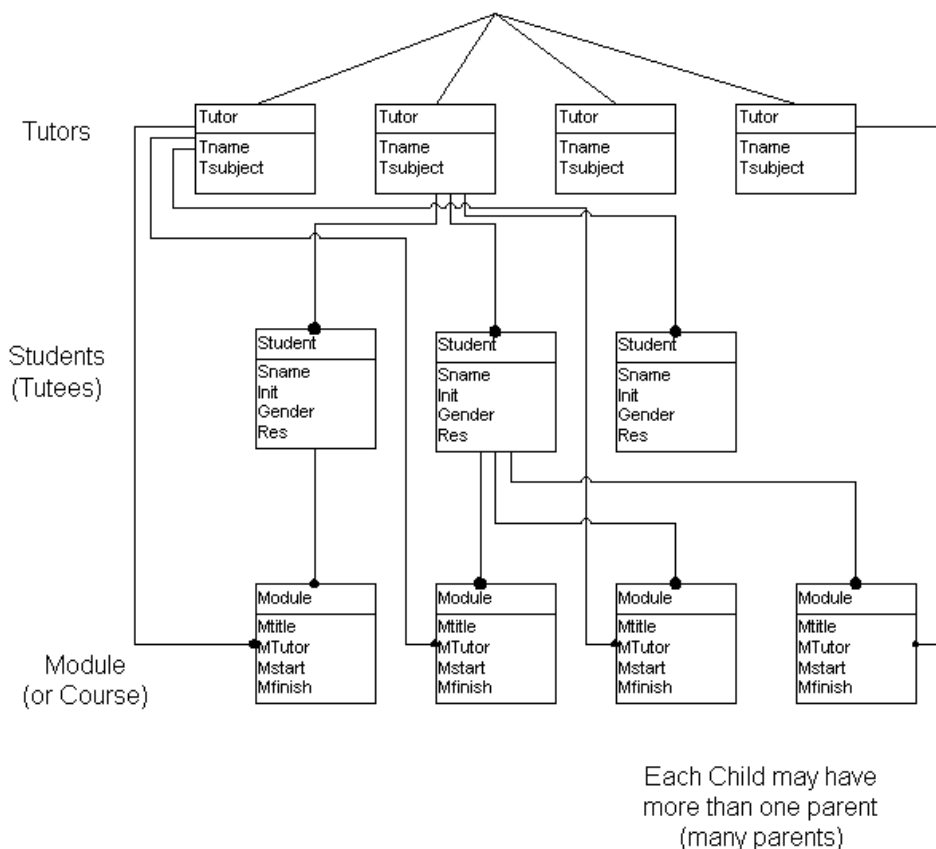


Figure 9.2: A Network Schema



Relational Databases

Relational databases are based on tables and on relations between those tables. Data is highly normalised. The tables are connected (related) via data NOT through physical links. Relational databases are highly flexible and support a wide range of ad hoc querying, however the efficiency of processing is low.

It is important to realise that Hierarchical, Network and Relational models are all concerned with Transaction Processing (TP). This means that they are optimised for collecting and processing data.

Data Warehouse

A data warehouse is NOT a TP system, it is optimised for analysis rather than transactions. Querying a Data Warehouse should be simple and the queries should operate at a very high level, e.g. how does the pass rate for residential students compare with the pass rate for non-residential students over the last academic year? A simple query at this level may involve extremely high amounts of data.

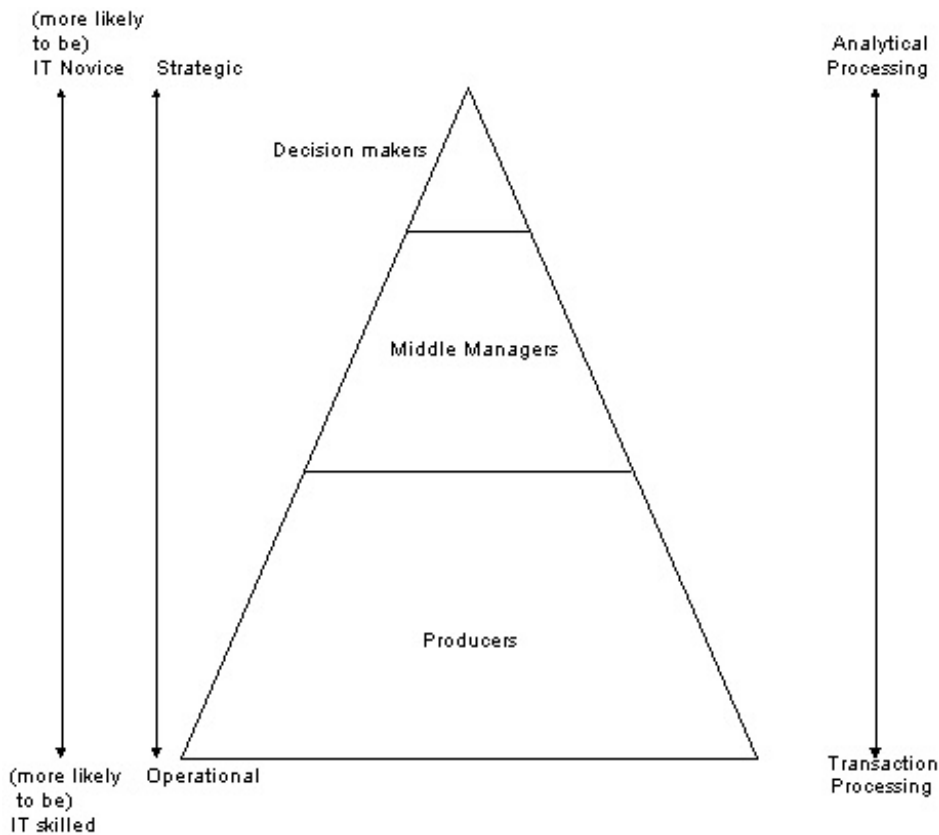
Transaction Processing vs. Analytical Processing

OLTP (On Line Transactional Processing) and OLAP (On Line Analytical Processing) systems are both database management systems, one is optimised for transaction processing, the other for analytical processing.

Consider the structure of many organisations. Generally they are pyramid shaped and hierarchical. At the top of the hierarchy are (relatively few) decision-makers. It is at this level that strategic decisions are made. Senior managers and Chief Executives are rarely if ever concerned with the minutia of how things are done within the organisation, they are more concerned with positioning the organisation "with regard to the market".

At the bottom of the pyramid are the producers who usually form the largest section of the organisation. In the middle are middle managers. Middle managers are often concerned with transmitting decisions from the top to the producers at the bottom and also with transmitting communication from the "shop floor" to the boardroom (couldn't ICT enable direct communication between the decision makers and the producers?).





Notes

Figure 9.3: A Hierarchical Organisation

By the beginning of the 1990's the efficiency of relational databases as Decision Support tools was beginning to be questioned.

1. Senior Management were (still are?) often IT novices and couldn't use the relatively complex tools themselves.
2. The type of question likely to be asked by senior management required highly summarised answers and the complexity of relational queries designed to achieve this was high.
3. It was possible to get different reports on the same data that gave different results (because, despite the claims made by the proponents of relational databases, it does matter how tables are related (order)).
4. Advances in the speed of processing and in the capacity of storage space (combined with increasing affordability) mean that it is feasible to store huge amounts of historical data in a way that is more suited to high level analysis.



5. Organisations knew that they had masses of data, but how could they get at it?

Once it is realised that we require electronic database systems to perform two distinct functions, i.e. the efficient and accurate gathering and processing of operational data (e.g. accounts system, student record system) and the overall organisation wide analysis of data, then it is a natural development to question whether one database model could ever do both jobs satisfactorily. Certainly the proponents of Data Warehousing recommend that relational databases are used to gather data at operational level, but that the analysis is carried out by a system optimised for analysis, i.e. a Data Warehouse.

Defining a Data Warehouse

A data warehouse is “a copy of transaction data specifically structured for query and analysis” Kimball R. (1996): *The Data Warehouse Toolkit*. John Wiley & Sons Inc. A data warehouse is a copy of sets of transactional data, these data can come from a range of transactional systems (e.g. finance systems, student records, stock, personnel etc.). Data that is stored in a data warehouse can be normalised but doesn't have to be, and often isn't.

A data warehouse needs to display the following characteristics:

1. Must be Subject orientated

A College Wide Information System would have Students, Courses, Curriculum and others amongst its subjects.

2. Must be capable of integrating data from a range of sources

Data that is published in a Data Warehouse can be imported from a variety of sources. It is possible (even likely) that this data is stored in incompatible forms, e.g. Date of Birth may be stored as “mm/dd/yy” in a staff record system whilst being stored as “dd/mm/yyyy” in a Student record system. These formats must be vetted and translated into common format before being published in the Data Warehouse.



3. **Must be time variant**

One of the crucial aspects of a data warehouse is that the data in it varies over time. This allows historical analysis of the data. Such historical analysis is useful to organisational managers because it allows a high level (highly aggregated) check to be made on trends. In this sense a data warehouse is not so much a decision support tool but a decision-monitoring tool. Data is imported on a regular basis (daily, weekly, monthly) and added to the existing data. Indeed “the Best and most useful facts are numeric, continuously valued and additive” [Kimball 1996].

Kimball defines continuously valued as “*a numeric measurement that is usually different every time it is measured.*” And defines additive as “*measurements in a fact table that are able to be added across all of the dimensions*”

Kimball also refers to “... [*moving*] snapshots of the OLTP systems over to the data warehouse as a series of layers, much like geologic layers” This image gives a good impression of the ability of a data warehouse to store the history of the organisation. Indeed the term “data mining” seems particularly appropriate for the activity of sifting back in time through the layers (of time) for analytical and comparative purposes.

4. **Must be non volatile**

All data added to the Data Warehouse is kept. A Data Warehouse is a slow moving, “quiet” place as contrasted to the fast rapidly changing world of the OLTP system which may be thought of as a “*twinkling database*” [Kimball, 1996]. Continually adding data has storage and processing implications. Certainly the growth in networking efficiency, speed of processing and the availability of relatively cheap on line or near on line mass storage (Terabytes or even Petabytes) has meant that the Data Warehouse model is capable of being implemented.



Notes

Bit	1 or 0
Byte	a computer "word" of 8 bits
KiloByte	1024 Bytes (2^{10})
MegaByte	1024 KiloBytes
GigaByte	1024 MegaBytes
TeraByte	1024 Gigabytes
PetaByte	1024 TeraBytes
ExaByte	1024 PetaBytes

An Exabyte is 9,223,372,036,854,780,000 bits.

An interesting paper [Lesk, M : *How Much Information Is There In The World?*]:

<http://www.lesk.com/mlesk/ksg97/ksg.html> : accessed 22nd October 2000] sets out to estimate the total amount of information in the world today and to compare it with the amount of on-line storage available. Lesk concludes that by 2000 there will be sufficient on line storage to store all of the information in the world (including text, speech, video, pictures and music). If we acknowledge the rapid growth in computing hardware technology, it is obvious that all information currently being produced will also be capable of being stored on-line.

Incidentally Lesk estimated roughly 12,000 PetaBytes of human information currently exist.

We will need to develop sets of tools to help us to identify important information and to analyse it. Data Warehousing promises to address at least some of the issues.

5. **Must be capable of informing decisions**

Leaving aside the question of how decisions are made, we can at least speculate that an analysis of past trends based on large amounts of historical data can at least form part of the decision making process. Data Warehouses are sold as "Decision Making" tools or as "Business Intelligence" tools, perhaps they don't play such a full part in the initial decisions, but they certainly play a role in monitoring the effects of the decisions.



Some Commercial OLAP Systems

Over the past few years, vendors of database management systems have been gradually introducing two distinct systems namely OLTP and OLAP. The differences between these systems and the way in which they are marketed is widening and it is instructive to survey the current marketing literature.

At the time of writing (October 2000), a quick search on the WWW revealed many vendors claiming to sell and support "Data Warehouses" or "Business Intelligence" systems. A small sample of some of the bigger names is listed below.

Computer Associates	Platinum Erwin
IBM	DB2 V. 7
Informix	Red Brick Decision Server
Microsoft	SQL Server 7.0
Oracle	Oracle Warehouse Builder
Sybase	Industry Warehouse Studio



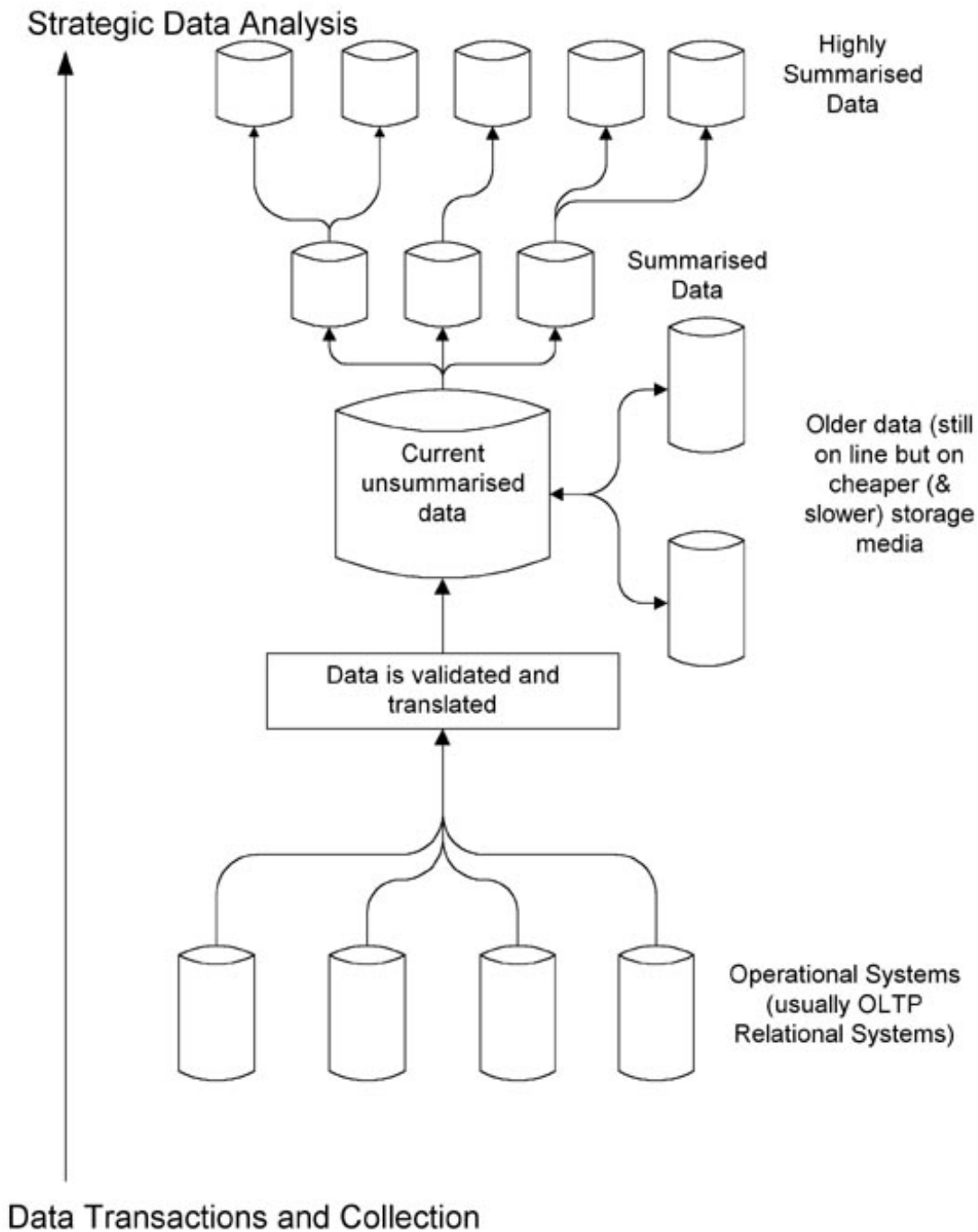


Figure 9.4: The Basic Structure of a Data Warehouse

Dimensional Models

The world of Data Warehousing is populated by terms such as Dimensionality, Star Schema, and Granularity. The structure of a data Warehouse is different from the normalised schema used in a relational OLTP system.



The Basic Structure of a Data Warehouse

At the heart of the design of a Data Warehouse is the visualisation of data as a cube or even as a hypercube (more than three dimensions). Time will of course be a vital dimension, as most analysis will be concerned with changes over time. If we imagine that we will find the measurements of the “performance indicators” of the organisation at the intersection of the axis of the cube, then we’ve got a pretty good visualisation of the “slicing and dicing” of data that strategic decision makers need. In our example in Figure 9.5 we can imagine finding the pass rates or the gender balance or the ratio of residential to non-residential students for any set of courses over a period of time.

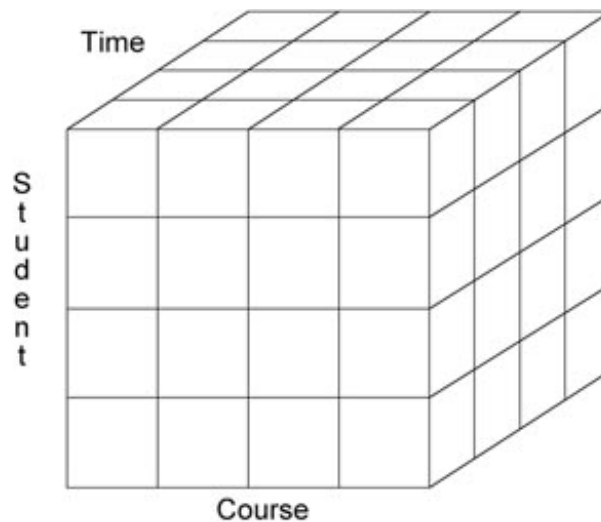


Figure 9.5: A Dimensional “visualisation” of data

Implementing the Dimensional Model

Implementation of a cube or hypercube can be done by using a “Star Schema” as shown in figure 9.6.



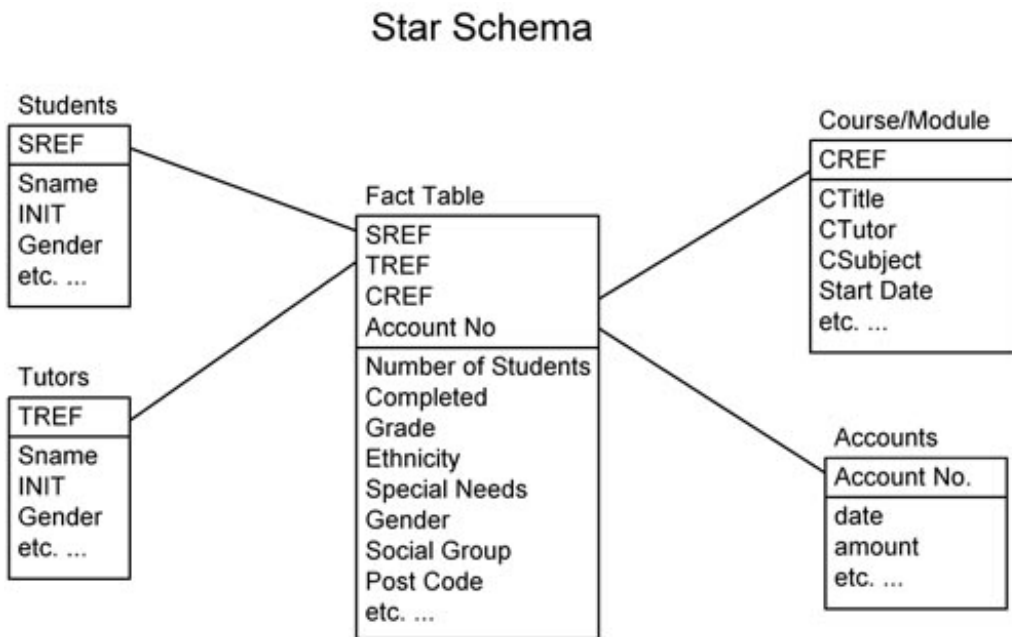


Figure 9.6: A Star Schema

The essential points are:

1. A central “Fact Table” that stores fact about the organisation
2. Any number of “Dimension Tables” that give descriptive details about the facts being measured
3. Only the Fact Table has multiple joins
4. It is the decision of the data warehouse designer to specify exactly what is shown in the Fact Table. In our example we may well decide that the fact table will store the total details for each course/module. This means that we will have a record for each course in our fact table, this degree of summarisation is known as “the grain of the fact table” [Kimball, 1996]. It is good design technique to keep this “grain” as small as possible as this will give a “higher resolution” to any subsequent analysis.



Ralph Kimball defines four essential steps in developing a star schema [Kimball, 1996, p. 27].

1. Choose the Business Process (e.g. a college student record system)
2. Choose the grain for the fact table
3. Choose the Dimensions
4. Choose the measured facts that will be stored in the fact table.

Why we wouldn't recommend a Data warehouse in all situations.

The following section is a summary of the article "The Case against Data Warehousing"

[<http://www.dwinfocenter.org/against.html> : accessed October 2000].

We wouldn't recommend the creation of a data warehouse in all circumstances because implementing a Data Warehouse is costly and is a major organisational commitment, we need to be sure that there will be an organisational benefit from doing so.

1. A data warehouse is valuable only if the organisation has an interest in analysing historical data
2. Data warehousing can complicate existing business systems. The use of a data warehouse needs to be constantly critically evaluated, there's no point in having reports for the sake of it.
3. If existing OLTP systems "do the job" there's no point in fixing what "ain't broke".
4. Data warehouses take time to learn to use effectively and the business practices may need to be re-evaluated.
5. Adding data for the sake of it may devalue the effectiveness of the Data warehousing technology
6. There may be cultural barriers in many organisations against enabling people to ask their own questions.



7. It's difficult to get the staff who are competent with data warehousing
8. It can cost a great deal in terms of IT maintenance
9. It may be a costly exercise to capture, clean up and publish data

Bringing it all together.

As you will have realised our discussion of Data Warehousing has been conducted with the processes of an organisation very much uppermost in our minds. This contrasts with highly technical approach taken earlier in the course with our treatment of relational databases. Data Warehousing techniques are relatively new and only during the late 1990's were they translated into fully marketable products. It seems inevitable that they will gain a larger share of the database market and that they will bring a major increase in the efficiency of on line data analysis. All of this will come at a price and many smaller organisations may well be unable or unwilling to pay for two database management systems.

Any student wishing to specialise in database design cannot be unaware of the developments highlighted in this section and it is hoped that the introduction given here will be built on at some future stage.



Exercise 9

You are advising a college Principal on how to proceed with the provision of IT services. You know that the college already has working finance and student record systems and that both of these systems are relational transaction processing systems. You also know that the Principal is concerned about issues such as consistency of reports and figures, suspect data, uncertainty about how to analyse data. You also know that the Principal is wholly unconcerned with the general mass of data collected and *just wants to have the important things pointed out.*

Produce a brief (guide length - 800 words) report outlining your recommendation and justify your choices (what would you take into consideration). Fully reference any resources you make use of.

