MULTIDIMENSIONAL DATA MODELLING: DATA SETS AND TOOLS FOR BUSINESS INTELLIGENCE

Clare Stanier
Staffordshire University
Beaconside
Stafford
c.stanier@staffs.ac.uk

ABSTRACT

Business Intelligence (BI) is an expanding data management area. BI analysis requires students to understand key concepts such as multidimensional data modelling, data cleansing and data mining algorithms, tools and techniques. Practical work for BI is often based around Excel which has a number of limitations as far as the teaching of BI is concerned. This paper discusses the experience of using the SSAS (SQL Server Analysis Services) tool to introduce students to multidimensional data modelling. We describe the way in which the tool was used, the issues encountered, the strengths and limitations of working with SSAS and possible future extensions. We also discuss issues with identifying and using data sets and outline a proposal to extend the data sets available to our students.

Keywords
Business Intelligence; Multidimensional Data Modelling; Data Sets; Data Mining.

1. INTRODUCTION

Business Intelligence (BI), which is understood here as the technologies, systems and processes used to help an enterprise to understand its business operations and to inform enterprise decisions [1], is an expanding data management area and one which is forecast to continue to grow [2]. Changes in technologies and increasing customer demand mean that BI is no longer restricted to large organisations [3] and this has implications in terms of student skills and employability [4]. The 2007 undergraduate Computing Benchmark [5] and the 2011 Masters Computing Benchmark [6] both include topics which are part of the field of BI and business intelligence concepts are now widely taught in UK universities. The teaching of Business Intelligence, however, presents a number of challenges. In order to promote deep understanding of concepts, we need to offer students the opportunity to apply theory through practical work. This requires a balance between tools which have a steeper learning curve than can be accommodated within the available teaching time and tools which do not sufficiently represent the complexity of the concepts. There are acknowledged problems with the resources available to teach BI [7,8]. One possible approach is to develop a full scale data centre [9] but this is a time consuming as well as resource intensive solution and was not an option for us in the short term. Resource restrictions and the limitations imposed by the curriculum, meant that we needed a tool, and supporting data set, which had a short learning curve for students but which gave them experience of working with an industrial standard business intelligence tool. This paper presents our experience of using the SSAS (SQL Server Analysis tool) to support the teaching of multidimensional modelling and data mining concepts to undergraduate and masters level students, discusses the strengths and limitations of working with this tool and the lessons learnt. We also experienced some issues when identifying suitable data sets for student use; the paper outlines these issues and suggests a possible solution.
2. ISSUES IN THE TEACHING OF BUSINESS INTELLIGENCE

2.1 The Business Intelligence Curriculum

Business intelligence, as defined in the introduction, is a portmanteau concept which could be stretched to cover almost any technology, system or process which produces information used to support decision making. Some elements however are generally accepted as core. Decision support, data mining, data warehousing and data modelling are listed as elements of the computing curriculum in both the Undergraduate and the Masters Computing Benchmark statements [5,6]. The teaching literature provides more detail suggesting data modelling based around the multidimensional cube and related structures such as star and snowflake schema [10], data warehouse and data mining concepts, tools and techniques [4] and an understanding of the role and contribution of BI [7]. Of the two modules discussed here, in one module we wanted to provide a general introduction to BI and OLAP (Online Analytical Processing), an understanding of multidimensional modelling (the multidimensional cube, slice and dice, star schema, fact and dimension tables) and the contribution these can make to business analytics; in the other module, we wanted to cover these topics and also to go further, to discuss issues such as data warehousing and ETL (Extract, Transform and Load) and to provide an introduction to data mining concepts and hands on experience of using data mining tools and techniques.

2.2 Resources for the Teaching of Business Intelligence

We had covered similar topics with other cohorts in previous years. Practical work with these cohorts had been based around data cleansing using PL/SQL, querying using OLAP SQL and online demos and tutorials – we found Alspace [11], for example, to be a useful resource. The limitations of the tools we were working with meant that students had little exposure to the process of creating and visualizing a multidimensional cube and little experience of working with a real life BI tool that was not spreadsheet based. This appears to be a sector wide issue since much of the reported work on teaching Business intelligence focusses on the use of MS Excel [12,13]. We also used Excel, working with built in features and particularly with pivot tables. It has been our experience that although Excel is an appropriate tool for students who are starting to explore data analysis, it does not allow more advanced students to develop a sufficient understanding of multidimensional concepts and of data mining algorithms. We found that students using the pivot table tool manipulated the data without fully understanding the implications of the way in which the data was visualized and that it was difficult for them to make the connection between the data sets in the databases they had been working with and the data set used in Excel. We also found that students were already familiar with Excel and did not feel sufficiently ‘stretched’ by the tool. Excel 2010 includes an advanced cube function; this was closer to our requirements but still did not provide students with the experience of creating and manipulating the cube in the same way as a dedicated BI tool. Dedicated BI tools, however, have a significant learning curve and we were limited in time and other resources. We identified our requirements for the teaching tool as:

- Accessible for students;
- A low initial learning curve but supports additional, more complex, functionality;
- Recognizably a real world tool;
- Allows students to visualize the multi-dimensional cube;
- Allows students to build an artefact and carry out BI analysis;
- Support for data mining in a BI context;
- Follows on from previous work on database concepts
- Support for the teaching of the key concepts identified in 2.1;
- Minimal resource (financial/software/hardware) implications.

We investigated WEKA [14]; this is a well-known data mining teaching tool available free from the University of Waikato. The tool is relatively easy to download and to use and provides a very good introduction to data mining techniques, with scope for more advanced work. A number of students used WEKA independently in the course of one of the modules. However, WEKA focuses on data mining algorithms and for teaching purposes we wanted a tool which could introduce students to additional BI elements. We reviewed a number of dedicated commercial and open source business intelligence tools. We felt that for teaching purposes, these tools provided more functionality than we needed and that the associated learning curve/resource requirements were too great. We then looked at the BI functionality included with the enterprise databases used in the university (Oracle and SQL Server) and revisited Excel 2010. A DBMS based approach had the advantage that the students would be able to make the connection between transaction processing data and analytical operations. Oracle offers a data warehouse builder and a range of Business Intelligence applications but these were not installed on our system and resource constraints meant that this was not a
The tool finally selected was SQL Server Analysis Services accessed via the SQL Server Business Intelligence Development Studio which ships with SQL Server Enterprise 2008 and Visual Studio. SSAS is an OLAP tool rather than a full data warehouse but had sufficiently functionality, and could handle sufficient data volumes, for our purposes. There were minimal resource implications as the software was already licensed and (we thought) was correctly deployed. Students had previously worked with SQL Server or were introduced to SQL Server in the course of the modules. All the students had a sound background in relational concepts and relational design – this was important when using the SSAS wizards which asked students, for example, to approve foreign keys and confirm how relationships should be set up. The tool connects to an underlying relational database, illustrating the way in which data gathered through OLTP (Online Transaction Processing) systems can be used to support analysis and allowing students to work with familiar data structures. SSAS supports multidimensional data modelling and a number of different data mining algorithms and there is a large amount of online tutorial material available to support private study. The interface is accessible but has the look and feel of a BI tool rather than a spreadsheet and we received positive comments from a number of students who were enthusiastic about working with a tool used by real world companies.

2.3 Data Sets

The decision to use SQL Server Business Intelligence was relatively straightforward given our requirements and the constraints within which we were working: deciding on the data set was more problematic. A number of data sets used in existing database modules were available but this data had originally been developed to support the teaching of design and transaction processing. The data volumes were small and some of the data structures were not a good fit for the teaching of business intelligence. We identified 4 possible sources for data sets, as shown in Figure 1:

<table>
<thead>
<tr>
<th>Large/Research data sets</th>
<th>Publicly available data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNAP data such as Gowalla</td>
<td>UK Government data such as crime and smoking data; National Centre for Health Statistics. EU data</td>
</tr>
<tr>
<td>British Oceanographic Survey Data</td>
<td></td>
</tr>
<tr>
<td>Iris Data set</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample data sets</th>
<th>Tutor &amp; student generated data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data provided with data mining tools such as Weka; DBMS provided datasets such as MS AdventureWorks</td>
<td>Purpose build data sets created through scripting or by use of a data generator</td>
</tr>
</tbody>
</table>

Figure 1: Data Set Sources

Large data sets such as those available through the open source SNAP (Stanford Large Network Data Set Collection) project and the British Oceanographic Survey would allow students to work with realistic data volumes. We looked at SNAP and in particular at the Gowalla data set\(^1\). Gowalla is a data set from a now closed social networking website \([15]\); the data includes identifiers, names, checkin times, latitudes and longitudes. Gowalla data is simple to download, easy to import into a data storage medium and the data structures are intuitive. From a BI perspective, the data did not map easily to the business scenarios with which students were familiar although it would support work in other areas such as NoSQL. The data available through the British Oceanographic Data Centre \([16]\) was more complex to access and the data semantics were less accessible for our students but the data was interesting from the point of view of advanced work as

\(^1\) With thanks to Mrs Zakia Idress Ali of Staffordshire University for introducing us to this data set
for some data sets, anomalies are flagged but not removed\(^2\). We felt, however, that this would require a higher skill level and more time to explore than we had available. We looked at the Iris data set \([17]\) but felt that undergraduates would find this data set difficult to relate to and that it did not have the necessary BI focus. Publicly available data was a better match in terms of BI. We looked a number of different datasets available through the data.gov.uk site \([18]\), some of which were available as Excel spreadsheets, some as csv files. Some of this data had already been pre-processed, making it unsuitable for our purposes and some of the data would require students to have a more in depth understanding of the context than we had time to support in the modules being taught in the current academic year. In the next academic year, 2013/14, we will be delivering a masters level module on analytical data and the intention is to use publicly available data sets in that module.

We next reviewed sample data sets, disregarded the sets used in WEKA, as we were not intending to work with this tool. We looked in detail at the Microsoft AventureWorks database \([19]\). There are different versions of AdventureWorks depending on analysis context but the standard version provides transaction data organised into relational data structures with sufficient data volumes to support the teaching of multidimensional data modelling. AdventureWorks is available with the SQL Server suite but depending on the installation, may need to be installed separately. For two reasons, however, we decided against using AdventureWorks. There are a large number of online tutorials written explicitly for the AdventureWorks database. Using a different data set for classroom teaching meant that the AdventureWorks tutorials would be available for private study without the risk that exercises/results will be duplicated. The second factor was that if AdventureWorks was used for teaching or assessment, answers to exercises would be readily available online; students would be able to copy and paste answers without necessarily having worked through the exercises or understanding how the results were arrived at.

The final option was to work with tutor and student generated data. For an optimization assignment that we use with final year students, students use data generators to generate tables with million plus rows but for the BI element we provided the students with tutor generated scripts to populate tables. The advantage of providing scripts is that data sets can be dropped and recreated as often as the student wishes, allowing him/her to work across university machines and their own machines. Data is frequently changed or corrupted during classroom exercises so the facility to easily drop and recreate was useful, particularly as our installation means that some students experience difficulty saving on the university configuration. Developing our own scripts meant that results were predictable and we could tune the data as we wish, designing in anomalies (one exercise required students to cleanse the data before they could use it). We were able to build on a transaction processing data script called Samples with which the students were already familiar, reducing the amount of time taken to understand the data. The Samples script was extended to create a script called Analysis which had a larger, although in BI terms still very small, data set. The restriction on data volumes was intentional; it was felt that a smaller data set would allow students to cross check their results and would support the data cleansing work. As discussed in section 4.4, we experienced some issues when we came to work with the Analysis data set.

3. **Teaching with SSAS**

SSAS was used as a tutorial tool to support the teaching of BI concepts. The aim was not to make students proficient in the use of SSAS but to illustrate concepts and familiarize students with the issues. This meant that we took some shortcuts with the tool, focusing on allowing students to create and deploy data structures quickly rather than providing a structured introduction to SSAS. We identified a number of topics where we felt that practical exercises would strengthen student understanding of BI concepts but some of these topics, for example the introduction to data cleansing exercise, were supported by use of SQL Server and Excel rather than SSAS. We used SSAS principally to support understanding of multidimensional modelling concepts and to introduce data mining concepts.

3.1 **Creating a MultiDimensional Cube**

SSAS enables students to create a star schema from a relational source, in our case the tables in an SQL Server 2008 database. Students first select the relational tables they will be using, for this exercise the tables created and populated through the Analysis script. They then create a data source and define a view based on the data source. The data source view is then used to create a cube. When students see the cube, they are required to build the cube and once the build is successful, to deploy the built cube for processing. Students are guided through the stages by SSAS wizards and students found the process straightforward. Three different groups were able, in the space of a one tutorial, to create and populate the underlying

\(^2\) Personal communication
relational data source, create a cube with a fact table and dimensions and use this to complete a (simple) multidimensional analysis task. The SSAS stages are shown in Figure 2 overleaf:

When the process is completed, the student sees a representation of the cube which, for our data looked like the structure shown in Figure 3. Yellow indicates the fact table which holds the measures and blue indicated the dimension tables. SSAS provides a large amount of functionality to work with the cube ranging from very simple tasks, such as aggregation to more sophisticated calculations and use of KPIs.
3.2 Using the Cube for Teaching

Students experienced few difficulties completing the exercise. The connection to the underlying database was not complex although some students initially encountered connection problems because they had selected the wrong connection option. As already noted, all the students working with SSAS had previously studied relational databases. This background in relational design enabled them to respond to requests for instructions from the SSAS wizards – for example, they could select the correct options to identify primary and foreign keys. The cube wizard offers suggestions as to the fact table which led to some students creating two fact tables and only one dimension. The students recognised the problem themselves and retraced their steps to produce a more appropriate cube structure for the data. This was useful as it illustrated the human versus system input into creating the cube. Students had no difficulty manipulating the cube to aggregate measures and carry out a simple analysis against the product dimension.

The exercise was designed to give students a more hands on understanding of OLAP concepts and to help them to visualise the multidimensional cube and in particular to reinforce understanding of the fact table/dimension element. Students had previously carried out a paper based star schema creation exercise using the Analysis data. The cube created through SSAS was the same as the paper based schema and this helped to make the design element more concrete. In previous work on multidimensional modelling, it had been our perception that students understood how to create, for example, a star schema, but did not fully appreciate that they were working with a different data model. The visualisation provided through SSAS reinforced the multidimensional concept. This was the first time we had worked with SSAS and with hindsight, we would have designed the data set differently. We had taken the decision to work with an extension of the Samples data set as we felt that students would find it helpful to work with data and data structures which were familiar. However, the data we used did not lend itself to a sensible time dimension. This was not a major issue as the cube wizard highlighted that a time dimension was missing and offered to create one. More significantly, the limited number of dimensions supported by the data led some students to ask if only three dimensions could be supported. We were able to clarify this but for future exercises we plan to use a more realistic data set. On our installation, all students work with SQL Server as administrators and were thought to have all the necessary privileges to work with SSAS. One unexpected problem when working with SSAS was that when students came to deploy the cube, they received a deployment failed message although the cube deployed on the staff account. On investigation, additional privileges were required to allow students to deploy the cube.

3.3 Data Mining with SSAS

This section discusses some initial work on using SSAS to support the teaching of data mining concepts and describes the way in which it is planned to use SSAS in the next academic year.

Previously we have introduced data mining concepts in lectures and then built on this material through paper based exercises, use of online tools, limited use of WEKA in some modules and through use of Excel. For example, in one exercise which has been used with a number of different groups, students are presented with a spreadsheet and asked to filter the data in order to carry out a market basket analysis. This allows students to identify outliers and relevant domains, to determine co-associations, conditional probability, association rules and support for the rule. A related exercise allows them to create a decision tree. This supports understanding of concepts but the Excel format understates the complexity of BI analysis. We found the paper based/demo/Excel approach very useful for reinforcing introductory concepts but it did not support more advanced teaching or give students the hands on experience of working with data mining algorithms in a BI context which was one of our teaching requirements. Following on from the work on the cube, we investigated data mining functionality in SSAS. Using SSAS, students can create a mining structure which allows the selection of an appropriate data mining technique, the creation of a model and visualises the results. To create a mining structure, the student selects the data source, which can be relational tables or a cube and selects a data mining technique from an option list which includes clustering, naïve Bayesian and Neural Networks. The structure itself is created via a wizard but students are asked to specify elements such as the case for analysis, the input tables and columns and the predictable column, as well as the composition of the testing and training sets. Some of these elements can in fact be left unspecified to be handled by the wizard but selecting elements requires students to understand the significance of, for example, testing sets. When the structure has been created, it is processed and deployed and the user can then see and interpret the results. The process is shown in Figure 4 overleaf:
In the current (2012/3) academic session we worked only with the clustering feature. The tool enabled students to create a cluster diagram similar to the partial diagram shown in Figure 5:

**Figure 4: Data Mining Structure in SSAS**

SSAS provides a number of features to refine the diagram and visualise the results. It was possible, for example, to drill into the cluster diagram to examine the profiles of the different clusters, to review cluster characteristics and to examine the distinctiveness of clusters using the cluster discrimination feature. The user is also able to select the degree (strength/weakness) of links shown. The initial cluster diagram was very complex in relation to the data on which it was based. Selecting only the stronger links produced a diagram which was easier to interpret as shown in Figure 6:

**Figure 5: Section of a cluster diagram**
3.4 Using the Mining Structure for Teaching

In the current academic session, SSAS for data mining was used only in one module and was an optional topic. Only the more advanced students attempted the exercise and restrictions on teaching time meant that it was only possible to explore one data mining technique with SSAS. We chose to use the clustering algorithm as we felt this would be more intuitive for the students than, for example, a neural network algorithm. We introduced the concept of clustering by looking at the \( k \)-means algorithm and the strengths and weaknesses of this approach. Excel provides a \( k \)-means feature and one group looked at the implementation of \( k \)-means in Excel. For the practical work with SSAS, we used the default SSAS clustering algorithm. This is the EM (expectation maximization) algorithm which is a soft clustering algorithm as opposed to the hard \( k \)-means algorithm. For the practical exercise, the analysis data set was used to create the cube. Working through the exercise, it was found that although the data supported the clustering algorithm, there was insufficient distinctiveness between clusters to produce informative results. The data set was tweaked to provide visible cluster discrimination as shown in Figure 7:

![Figure 6: Section of a cluster diagram showing strongest links](image)

![Figure 7: Cluster Discrimination with the revised data set](image)

We found that the learning curve with the data mining element was significantly steeper than with the cube exercise and that more theoretical underpinning was needed. The cube wizard was self-explanatory for students with a grounding in relational theory; to make sense of the data mining element, students needed an equally strong underpinning in data mining. The data mining wizard required students to understand, and make decisions about, the structure of the analysis case. As with the cube exercise, it was relatively easy to create a diagram but interpreting the results required more knowledge and greater understanding. From a teaching point of view, the wizard that creates the mining structure has two particularly useful features. The first is that the user is required to specify, for numeric data, whether the data is to be treated as discrete or continuous. One of the issues with using Excel was that data issues tended to be smoothed out by the tool; in SSAS, users were required to make explicit decisions. The second feature is that wizard asks users to specify the amount of data to be included in the training set and the testing set and provides a good explanation as to the reason for requiring this data. Experimenting with this, it proved possible to ‘crash’ the
model, by specifying the wrong percentages or by including so many attributes in the case to be tested that
results could not be processed. This reflected the fact that the data set in use was very small but from a
教學 standpoint, provides a good introduction to the use of training data sets. With the cube exercise, the
goal was to create a cube to illustrate concepts such as fact table, measures and dimensions. With the data
mining exercise, it was important that the students recognised that the goal was to interpret the results of the
data visualisation (in this case, the cluster diagram) rather than to create the diagram itself. SSAS provides a
tabbed menu, as shown in Figure 7, which provides additional information about the cluster diagram to
support students in drilling down into the composition of the diagram.

4. CONCLUSIONS AND FURTHER WORK

4.1 Conclusions – Working with the SSAS Cube

SSAS has proved to be a versatile tool for teaching purposes, allowing students to take data from relational
tables and model it in a multidimensional environment. We felt that students were comfortable working with
this tool and that they were able to visualise the data. Creating the cube proved straightforward and students
appeared to recognise the implications of creating the fact and dimension tables from the relational database.
The cube exercise, for example, made an Excel based data cleansing exercise more relevant as students had
to cleanse the data before creating the tables on which the cube was based. The difficulties experienced were
mainly due to the data set we used and to the limitations on time which restricted the amount of analysis
which could be carried out against the cube. As discussed in section 3.2, some students originally interpreted
the schema to mean that a cube could never support more than 2 dimensions. A further limitation was that the
data did not support a sensible time dimension. It is planned to use SSAS in the next academic year to
support a module which has a greater BI component. This will allow us to complete more work with the cube,
using a number of different data sets and carrying out a wider range of analysis against the cube.

4.2 Conclusions – Working with SSAS Mining Structures

SSAS introduces students to a real world OLAP tool and provides a richer learning experience than is
available through Excel. The more technical nature of the mining structure element means that students will
require more preparation and support than when working with the cube and that the practical work must be
based on a good understanding of underpinning concepts. We would suggest introducing data mining
concepts through lectures and paper based exercises, employing Excel to allow students to explore concepts
using a tool with which they are familiar, and then moving to SSAS. In the more advanced module planned for
next year, students will be introduced to more advanced functionality and will work with realistic data sets and
data volumes.

4.3 Conclusions – Data Sets

Working with SSAS for the first time, our main concern had been to provide a data set with which the students
were familiar so that they would be able to focus more on the analysis than on understanding the underlying
data structures. We found that while familiarity with the data was an advantage, the data set we used, which
had been adapted from a design and development teaching set, was not best suited to exploring BI concepts.
We had to tweak the data set to support some of the exercises and the small number of dimensions used
risked confusing students. On the other hand, we noted some clear advantages to a small scale, well
understood data set. Students were able to visualise the results expected, for example when data cleansing
and working with the fact table, and could manually cross check results. Small scale data sets were quick to
load – an advantage with our system where students dropped and recreated the data set at the start of each
tutorial. SSAS supports the building of an analysis set over a relational data source; the students understood
the underlying tables and this appeared to smooth the progression to BI. It also emphasised the
interconnection between transaction data and analysis. The intention for next year is to work initially with a
small, scripted data set to introduce concepts and then to move to larger data sets, based on publicly
available data, to support more advanced work.

4.4 Overall Conclusions and Future Work

BI tools are evolving rapidly and it is likely that in the longer term we will introduce additional tools. For the
immediate future, however, as we have the necessary hardware and software licenses and have found that
the tool meets our requirements, we plan to continue to use SSAS and to extend the range of tasks that
students carry out, particularly with regard to data mining and the evaluation of the accuracy of models. In the
course of researching a suitable BI data set to support work with SSAS, we came across a number of data
sets which could be used to support other work such as NoSQL and advanced data analytics. Identifying data
sets was time consuming and there were also a number of other issues such as mapping to data structures,
exporting data and the amount of pre-processing required. Our students regularly create large data sets for testing and optimisation; this involves a certain amount of duplication of effort and some students have difficulty generating the necessary data. For the next academic year it is proposed to set up a data set portal for our students, to provide a one stop access point for use by any Staffordshire University student who wishes to access a data set.

5. REFERENCES