Creating Multidimensional Data in SAS for Excel Pivot Tables
Chandy Karnum, Ace Analytics Inc., Madison Wis.

Abstract:

If you are an analyst in data driven industry you are constantly faced with the challenge of digesting large volumes of data into meaningful information in the cheapest way possible. Although there a number of sophisticated/expensive software packages and business intelligence systems specifically designed to provide the solution many of the analysts and organizations have used analytical tools such as Base SAS and Excel to get the job done. This paper will attempt to demonstrate how using only a few steps in Base SAS you can create the multidimensional data for Excel-pivot tables.

Nowadays, Excel has become the common denominator for business analysis in the industry. However, Excel lacks the power to digest huge volumes of data found across any industry. Over the years Base SAS has grown to become the leader in data extraction and summarization regardless of the data source or volume. Here we will discuss the concepts of a dimensional data model, multidimensional dataset and the definition of dimensions and facts. Then we will review data extraction, creating dimensions and summarization techniques using Base SAS. Finally we will see a demonstration of Excel pivot tables that use the multidimensional dataset to reveal patterns and trends.

Introduction:

A simple cross-tab report that you generate using PROC TABULATE (figure 1) may consists of sales totals expressed at two levels of summarization or dimensions. This is also referred as a “two dimensional report” where the product and region variables represent the “dimension” and the sales totals is called as “fact”.

<table>
<thead>
<tr>
<th>Product</th>
<th>Canada</th>
<th>Mexico</th>
<th>U.S.A.</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BED</td>
<td>$460,671</td>
<td>$312,039</td>
<td>$1,823,734</td>
<td>$2,596,444</td>
</tr>
<tr>
<td>CHAIR</td>
<td>$457,739</td>
<td>$300,831</td>
<td>$1,838,075</td>
<td>$2,596,645</td>
</tr>
<tr>
<td>DESK</td>
<td>$469,536</td>
<td>$304,936</td>
<td>$1,830,806</td>
<td>$2,605,278</td>
</tr>
<tr>
<td>SOFA</td>
<td>$477,230</td>
<td>$313,853</td>
<td>$1,927,313</td>
<td>$2,718,396</td>
</tr>
<tr>
<td>Grand Total</td>
<td>$1,865,176</td>
<td>$1,231,659</td>
<td>$7,419,928</td>
<td>$10,516,763</td>
</tr>
</tbody>
</table>

If one were to produce such reports periodically the “time” dimension will need to be considered to track performance trends. To visualize this dimension imagine if the monthly reports were stacked to represent a slice of time.
To extrapolate further as these reporting requirements get complex one can imagine the demands on the cross tab reports getting quite complex as more facts and dimensions are added to the fact-dimension matrix. This matrix is often simply represented as a cube but understanding multidimensional data may require the assistance of OLAP visualization tools.

Business analysis often requires visualizing the results by slicing and dicing in different perspectives or dimensions to discover patterns, clusters or trends. Instead of cross tab reports this visualization can be better rendered by tools such as pivot tables in Excel. Pivot table is the simplest form OLAP (On-Line Analytical Processing) tool and it is fairly easy to implement and understand. Before we look into pivot tables let’s examine a few concepts and definition.

What is a Multi-Dimensional Dataset/Table?

In simple terms, a multidimensional dataset is constructed with one record for each discrete metric/fact accompanied by all the relevant or associated dimensions that describe the fact. These datasets are normally an aggregation of detail transactional or operational data. The lowest level of aggregation is called as the “grain”.

Grain: The lowest level of detail at which the data is recorded.
What is a Dimension?

A dimension is a context, aspect or perspective by which the facts may be accessed, grouped, selected, sequenced, filtered and presented.

Example: Product, Region, Time etc...

What is a Fact?

A fact is a business performance measurement, typically numeric and additive. It is also called as a “metric”. This measurement is true to the “grain”

Example: Quantity Sold, Price, Total Sales, etc...

What is a Grain?

A grain is the level aggregation defined by the business requirement where there is one record containing a single measurement value or “metric” that has a descriptive contexts to related dimensions.

Guideline to acquiring a Multi-Dimensional Dataset:

Even though the larger topic of understanding the concepts around building a dimensional data models is out of scope, the present work assumes the availability of multidimensional datasets in the industry. The source of these datasets may vary depending on the data environment.

For instance:

- **Transactional data**: Sourced from application transaction tables. The data is highly de-normalized and highly granular where facts and all the associated dimensions are available one each record. This is the easiest source where you can get all the facts and dimensions in a single table.

- **Star Schema**: Sourced from a dimensional data mart. The data is structured into fact tables and dimensional tables linked by primary/foreign key relationship. This is an ideal source and designed for improved query performance and capable of complex analysis.

- **Relational data**: Sourced from relational tables where the analyst needs to link source system tables to derive the facts and dimensions. The grain of the data creates the fact table and all the associated dimensions are sourced by linking to tables containing the dimensions. This can prove to be a very daunting approach requiring complex transformations.

Before sourcing your multidimensional dataset you will need to determine the business requirement around data analysis. This will help in defining the scope of “grain” of the data.
Here are a few thumb rules:

- The granularity of the data or the “grain” of the data has to be defined. The grain is based on the “fact” that needs to be measured. The “dimension” is the context within which the facts were generated.
- The facts and dimensions are sourced in such a way that there needs to be one record per numeric measurement described by all the relevant dimensions of interest.
- There needs to be a one to one association between the dimension record and the fact record. (..is true to grain)
- Dimensional hierarchies must be pre-defined by a business process

Creating a multidimensional data involves three steps:

1. **Extraction & Transformation**: Multidimensional reporting requires gathering the facts/metrics and their respective groupings/dimensions at the granular level.
2. **Summarization**: The granular data then needs to be aggregated to produce N-dimensional row level data
3. **Load**: N-dimensional data then has to be loaded as source data into Excel to create the pivot table as a presentation layer for visualization.

1. **Extraction & Transformation**:
   
   The initial extraction of data is the most daunting task of the three steps because you have deal with huge volume of detail transactional / operational data. This detail data may reside in several databases that need to be collated to gather the required facts and dimensions. The DATA step or PROC SQL happens to be robust enough to handle all these challenges. When certain dimensions needed for the analysis do not exist we can programmatically derive them by applying the appropriate business logic that is referred to as a transformation. Here is how a simple data extraction and transformation step may look like:

```
PROC FORMAT; *formats to represent sales metrics;
  value slsrange
    low - <5000 = "$0 to $5,000       "
    5000 - <10000 = "$5,001 to $10,000  
    10000 - <100000 = "$10,001 to $100,000"
    100000 - high = "$100,000 and above 
    OTHER           = "Amount not known   
RUN ;

Data Product_Sales;
  Length Country $10;
  set sashelp.prdsal2; *source transaction table;
  /* Create a Custom Dimension : Region Grouping */
  If Country = "U.S.A." and State = "California"
  then Region = "USA-West";
  else
  If Country = "Mexico" and State = "Campeche"
```
then Region = "MX-West";
else
If Country = "Canada" and State = "British Columbia"
  then Region = "CAN-West";
else
  Region = "Other";
/* Create Range based dimensions for Sales amounts*/
  format ACTUAL PREDICT slsrange.
run;

2. Summarization:

Once the data is successfully extracted creating an N-Dimensional data aggregation is a breeze when you use PROC SUMMARY. This procedure has all the methods you need to create a fully functional n-dimensional dataset. Here is an example:

```
proc summary data = Product_Sales missing nway ;
  class Year Qtr Country Region Prodtype product ;
  var actual predict;
  output out = Product_Sums
    sum(actual)=Actual_Sum
    sum(predict)=Predict_Sum
  ;
run;
```

- CLASS statement defines the dimensions and the dimensional hierarchies are specified from right to left.
- NWAY creates one record per fact with their respective dimensions
- VAR statement captures the facts/metrics and the OUTPUT statement can be used to produce a variety of statistics.
- If you are planning on loading the resulting dataset into excel make sure the record count is less than 65,536 rows

3. Load / Export to Excel:

The summarized data is the multidimensional data needed to implement pivot tables. This dataset can be exported to excel using PROC EXPORT. This requires the “SAS/ACCESS to PC file formats” module or you can use ODS in Base SAS to generate a XML file. XML files can be easily opened with excel and save as an excel file. Here is a PROC EXPORT example:

```
PROC EXPORT DATA= Product_Sum
   OUTFILE=  
"C:\SASDATA\PivotData\Product_Sum_Data.xls"
   DBMS=EXCEL2000
   REPLACE;
RUN;
```
Visualization and Analysis:

Once the data is Excel you can either launch the pivot table wizard to build a pivot table or you can dynamically create a pivot table by calling a VB Script from a SAS session. Here is an example of calling a VB script:

```sas
* call VB script to make the pivot table ; data _null_; x 'C:\userdata\sas\pivot.vbs'; run;
```

Here is the VB script code:

```vb
Set XL = CreateObject("Excel.Application")
XL.Visible=True
XL.Workbooks.Open "C:\userdata\sas\Product_Sum_Data.xls"
Xllastcell= xl.cells.specialcells(11).address
XL.Sheets.Add.name = "PivotTable"
xldata="PRODUCT_SUM"
XL.Sheets(xldata).select
XL.ActiveSheet.PivotTableWizard SourceType=xlDatabase,XL.Range("A1" & ":" & xllastcell),"PivotTable!R1C1",xldata
XL.ActiveSheet.PivotTables(xldata).PivotFields("YEAR").Orientation = 1
XL.ActiveSheet.PivotTables(xldata).PivotFields("QUARTER").Orientation = 1
XL.ActiveSheet.PivotTables(xldata).PivotFields("COUNTRY").Orientation = 1
XL.ActiveSheet.PivotTables(xldata).PivotFields("REGION").Orientation = 1
XL.ActiveSheet.PivotTables(xldata).PivotFields("PRODTYPE").Orientation = 1
XL.ActiveSheet.PivotTables(xldata).PivotFields("PRODUCT").Orientation = 1
XL.ActiveSheet.PivotTables(xldata).PivotFields("Actual_Sum").Orientation = 4
XL.ActiveSheet.PivotTables(xldata).PivotFields("Predict_Sum").Orientation = 4

After the pivot table is loaded I will demonstrate how a pivot table can be used to dynamically create cross-tab reports, 3-D histograms, subtotals, roll-up and do drill-downs.

Cross-tab report:

<table>
<thead>
<tr>
<th>Sales Month</th>
<th>(All)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PRODUCT</strong></td>
<td><strong>Data</strong></td>
</tr>
<tr>
<td>BED</td>
<td>Actual Sales</td>
</tr>
<tr>
<td></td>
<td>Predicted Sales</td>
</tr>
<tr>
<td>CHAIR</td>
<td>Actual Sales</td>
</tr>
<tr>
<td></td>
<td>Predicted Sales</td>
</tr>
<tr>
<td>DESK</td>
<td>Actual Sales</td>
</tr>
<tr>
<td></td>
<td>Predicted Sales</td>
</tr>
<tr>
<td>SOFA</td>
<td>Actual Sales</td>
</tr>
<tr>
<td></td>
<td>Predicted Sales</td>
</tr>
<tr>
<td>Total Actual Sales</td>
<td>$1,865,176</td>
</tr>
<tr>
<td>Total Predicted Sales</td>
<td>$2,002,259</td>
</tr>
</tbody>
</table>
Swapping Dimensions:

Create 3-D histograms that can be rotated using the active-x controls.
Conclusion:

The goal of this paper was to provide the audience with a simple explanation to the underlying principles governing multidimensional datasets. Multidimensional datasets is the foundation for pivot tables or any OLAP technology. My hope is that the audience will leave with a basic understanding of the underlying principles so that they can further digest more complex topics like dimensional models, star schemas, etc.

References:


Nelson, Gregory S.B. “Implementing a Dimensional Data Warehouse with the SAS System” Published in the proceedings of SUGI 22


Ecklund, P. “An Introduction to Excel Pivot Tables” published as white paper in the FUQUA School of Business, 2002


Contact Information:

I would really appreciate any feedback, questions, comments and ideas. I would like to hear from you. Here is my contact information:

Chandy Karnum | Principal
Ace Analytics Inc.
1845 Jenifer St.
Madison WI 53704
☎: 608.206.9063
karnumc@yahoo.com | AceAnalytics.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.