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".

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

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.

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

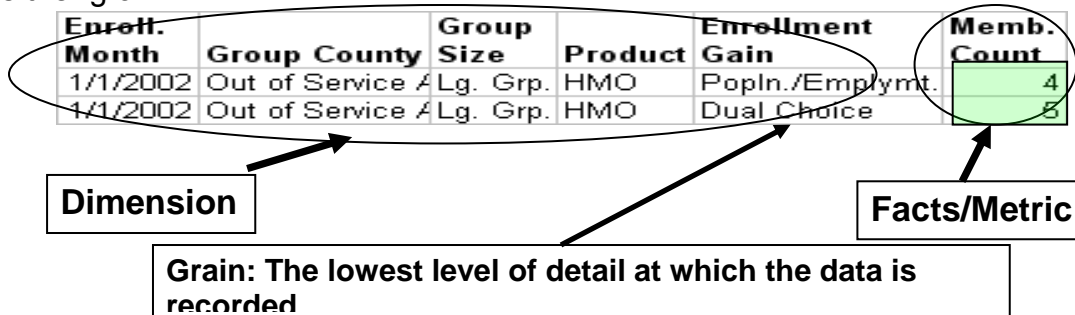| PRODUCT	Canada	Mexico	U.S.A.	Grand Total
BED	$ 460,671	$ 312,039	$ 1,823,734	$ 2,596,444
CHAIR	$ 457,739	$ 300,831	$ 1,838,075	$ 2,596,645
DESK	$ 469,536	$ 304,936	$ 1,830,806	$ 2,605,278
SOFA	$ 477,230	$ 313,853	$ 1,927,313	$ 2,718,396
Grand Total	$ 1,865,176	$ 1,231,659	$ 7,419,928	$10,516,763

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.
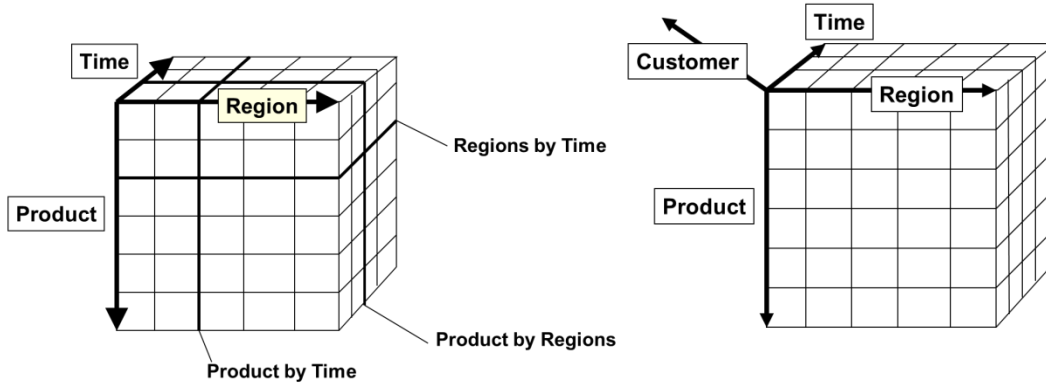
Business analysis often requires visualizing the results by slicing and dicing in different perspectives or dimensions to discover patterns, clusters or trends. Instead of cross tab reports this visualization can be better rendered by tools such as pivot tables in Excel. Pivot table is the simplest form 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".

Enroll. Month	Group County	Group Size	Product	Enrollment Gain	Memb. Count
1/1/2002	Out of Service A	Lg. Grp.	HMO	Popln./Emplymt.	4
1/1/2002	Out of Service A	Lg. Grp.	HMO	Dual Choice	5

Dimension

Facts/Metric

Grain: The lowest level of detail at which the data is recorded

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

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.

<u>Guideline to acquiring a Multi-Dimensional Dataset:</u>

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.

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

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:

```
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:

Sales Month	(All)				

		COUNTRY			
PRODUCT	Data	Canada	Mexico	U.S.A.	Grand Total
BED	Actual Sales	$460,671	$312,039	$1,823,734	$2,596,444
	Predicted Sales	$496,267	$353,178	$1,894,823	$2,744,268
CHAIR	Actual Sales	$457,739	$300,831	$1,838,075	$2,596,645
	Predicted Sales	$494,753	$333,332	$1,953,701	$2,781,785
DESK	Actual Sales	$469,536	$304,936	$1,830,806	$2,605,278
	Predicted Sales	$492,749	$337,728	$1,969,058	$2,799,535
SOFA	Actual Sales	$477,230	$313,853	$1,927,313	$2,718,396
	Predicted Sales	$518,489	$350,386	$1,937,251	$2,806,126
Total Actual Sales		$1,865,176	$1,231,659	$7,419,928	$10,516,763
Total Predicted Sales		$2,002,259	$1,374,623	$7,754,832	$11,131,715

Swapping Dimensions:

COUNTRY	(All)				

		Sales			
PRODUCT	Data	Jan-04	Feb-04	Mar-04	Apr-04
BED	Actual Sales	$342,960	$345,128	$339,934	$965,589
	Predicted Sales	381958	374997.8	364688.5	1006447
CHAIR	Actual Sales				
	Predicted Sales				
DESK	Actual Sales	$156,234	$148,972	$147,690	$772,774
	Predicted Sales	162146	161554.2	159080.2	816854.6
SOFA	Actual Sales				
	Predicted Sales				
Total Actual Sales		$499,194	$494,101	$487,624	$1,738,363
Total Predicted Sales		$544,104	$536,552	$523,769	$1,823,302

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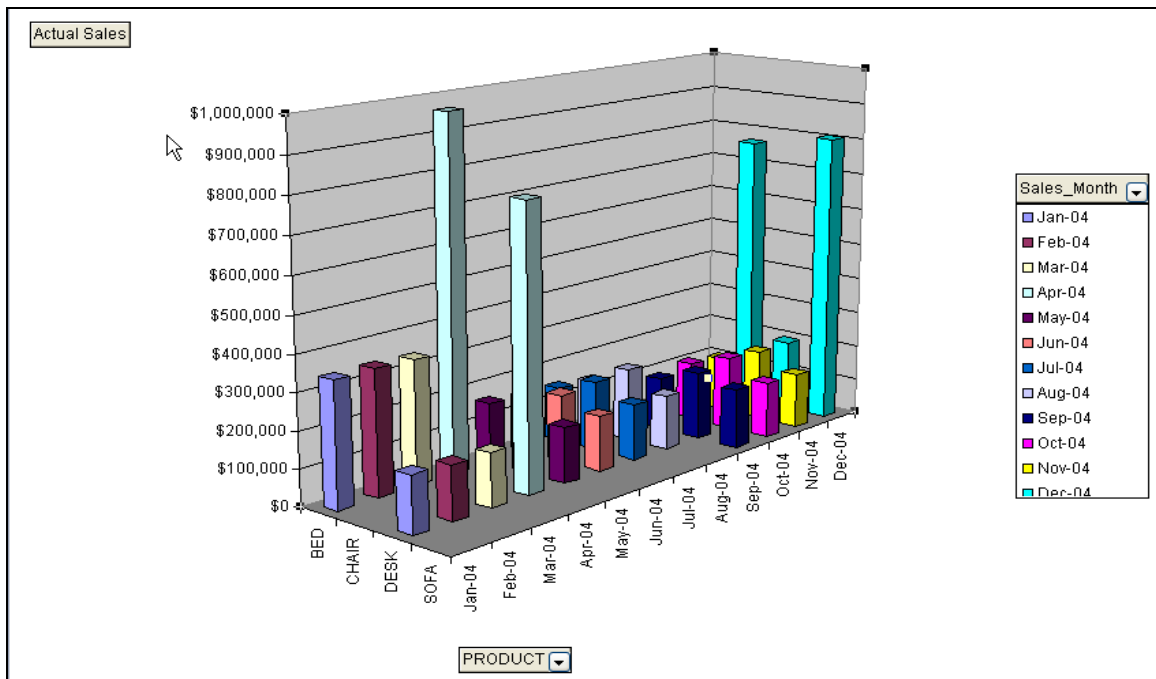
Conclusion:

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:

Schacherer , Christopher W. "Base SAS® Methods for Building Dimensional Data Models". Paper 58-2010. Published in the MWSUG 2010 Conference Proceedings

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

Gray, J. et. al. "Data Cube: A relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-totals" Microsoft research technical report MSR-TR-96-XX, 1996

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

Kimball, R., Ross, M. "The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling". John Wiley and Sons 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