

PIVOT TABLES/CHARTS MAGIC BEANS WITHOUT LIVING IN A FAIRY TALE

Prepared by
John Van Wagenen
Caterpillar Inc, Peoria IL USA
van_wagenen_john_s@cat.com
for the

34th Annual International Conference of the Computer Measurement Group, Inc.
December 7-12, 2008
Las Vegas, Nevada USA

This presentation will show how one analyst uses the pivot table feature of Microsoft EXCEL to arrive at a quick performance analysis for DB2 and BATCH mainframe applications. At our enterprise, we use pivot tables and charts as monitoring tools, performance analysis, and management communication. An explanation of the basic concepts to create and manipulate pivot tables will be combined with a real-time demonstration of the tables and charts. At times in your career you find a tool that is truly worth learning about. This is one of those times and one of those tools!

INTRODUCTION

There is a simple math assignment that we all experienced back in our school days. It involved tables of data and our ability to cross-tabulate some count. We collected data on the type of pet each classmate had, or what breakfast cereal they ate each morning. The teacher asked us how many kids had dogs/cats/fish or whether more Cheerios were eaten than Lucky Charms. These were clever but tedious exercises since we only had 20 rows of data. What we really learned was that we would NEVER wish to solve these same questions with lots of data (like 50,000 rows). If someone told you that there was a tool that could easily manipulate and cross-tabulate data AS WELL AS produce presentation-ready graphs and charts, you would use it right? Well the tool exists. It is the pivot table feature within the Microsoft EXCEL product. This paper illustrates some basic usage and value.

HISTORY of PIVOT TABLES

According to Bill Jelen and Mike Alexander, authors of the book Pivot Table Data Crunching, the concept that led to today's pivot table came from Lotus Development Corporation with a revolutionary spreadsheet program called Lotus Improv. Improv was envisioned in 1986 by Pito Salas, a Cambridge, MA-based software developer. While working with the Lotus Advanced Technology Group in 1986, Salas realized that spreadsheets have patterns of data. By designing a tool that could recognize these patterns, one could quickly build advanced data models. This insight led to the next generation spreadsheet concept that later became the basis for Pivot Tables in Microsoft EXCEL. The product was released by Lotus in 1989 as Lotus Improv.

My personal experience with this tool crosses about six years and would be described as an intermediate level. An upfront disclaimer is that this tool allows many methods to accomplish the same function (for example toolbars versus dropdown menus). The examples shown in this paper are an illustration of how this author thinks about solving problems. The examples are not necessarily the smartest, fastest, or most elegant solutions. They are however, the BEST, because they work at our enterprise and have become highly dependable. This paper is intended for an introductory audience. Some knowledge of EXCEL functionality is required to understand the pivot table/chart features. The examples should encourage confident experimentation by the neophyte user.

When would you use pivot tables?

Quite often the management in our enterprise desires sorted lists identifying who or what are the top 10 or bottom 10 for some characteristic. Perhaps it is storage usage, or CPU hours, or even safety statistics. We collect data fast and furious about almost everything anymore. And inevitably someone notices the data and begins asking questions. For example:

- What are your top 10 applications by CPU utilization?
- “ “ “ “ “ by transaction count?
- “ “ “ “ “ by average transaction time?
- What application has the maximum transaction time?
- What applications run average transaction time above the enterprise standard?
- What are the top 10 applications that have changed this month?
- “ “ “ “ “ “ “ in the last quarter?
- “ “ “ “ “ “ “ in the last year?
- What are the top 10 applications that varied from baseline workload?
- During which hour of the day does the peak application load occur?
- What are the biggest applications?
- Which applications have not been used in six months?

Believe it or not, a single set of data and a pivot table can answer all of these questions. While each question may require a different manipulation of the pivot table grouping and summarizing variables, a skilled analyst could provide tables and charts for the scenarios listed above in under an hour (assuming the data exists already).

Notice that several of the questions were similar in nature but wanted a different duration of time, or a different summarization variable. This is one of the best features of a pivot table and can be invaluable when trying to convey information to an audience that is not totally familiar with the detail data.

Not only can a pivot table be used for these types of lists, but every pivot table can also be represented as a pivot chart. The range of chart types and chart options is as broad as whatever EXCEL offers, so the only limitation will probably be your imagination.

For performance analysis on the mainframe, pivot tables are well-suited to processing mass data that can be generated by mining SMF records. Even summarized data could represent 1000's of rows. A pivot table can summarize all of that data into a single page table or chart. In the past, a pivot table was only able to process 64,000 rows of data in any spreadsheet tab (a limitation this author still lives with). Recent enhancements to EXCEL increase this limit to a million rows. Unless you are a glutton for detail, most data analysis can be scoped down to include a million records or less.

Understanding pivot tables

Individual queries on the web about pivot tables often end in frustration because there is no definitive answer to the question “what is a pivot table?” The anxiety can be reduced by simply accepting that pivot tables and charts, are a tool in an EXCEL spreadsheet. And a spreadsheet is just a tabular set of data. So the first idea to digest is that pivot tables are useful when you have a table of data values. The next idea is to identify that you want to gather some numerical metric. This could be a total, an average, the max or min value, or even a simple count. If you do not have this requirement for analysis, then a pivot table will probably do you no good. The third idea to recognize is that in some way, there is descriptive data in each table row that would be useful or interesting to summarize. By descriptive data, think about the values of a column of data.

Thinking back to the introduction of this paper, if all you wanted was a grand total of pets, you could simply sum a column in the table. But if you want to know the cross-summarization of how many students had dogs/cats/fish, then a pivot table makes this grouping very easy to do and puts the information in an easily understood format.

In terms of performance data, the numeric data we want to summarize is usually CPU seconds. The metric we needed to track, is the total. The descriptive data is at first an application code. In lower detail, we may wish to summarize on jobnames or transaction ids. The same pivot table can easily do both levels of reporting. The next exhibit is the basic format of a pivot table when it is "launched" in a spreadsheet (see figure 1a).

There are five basic concepts to understand with pivot tables: (a) DATA ITEMS; (b) ROW FIELDS; (c) COLUMN FIELDS; (d) PAGE FIELDS; (e) Pivot Table Toolbar. Each of these will be explained from the perspective of the sample data shown in figure 1b.

Figure 1a

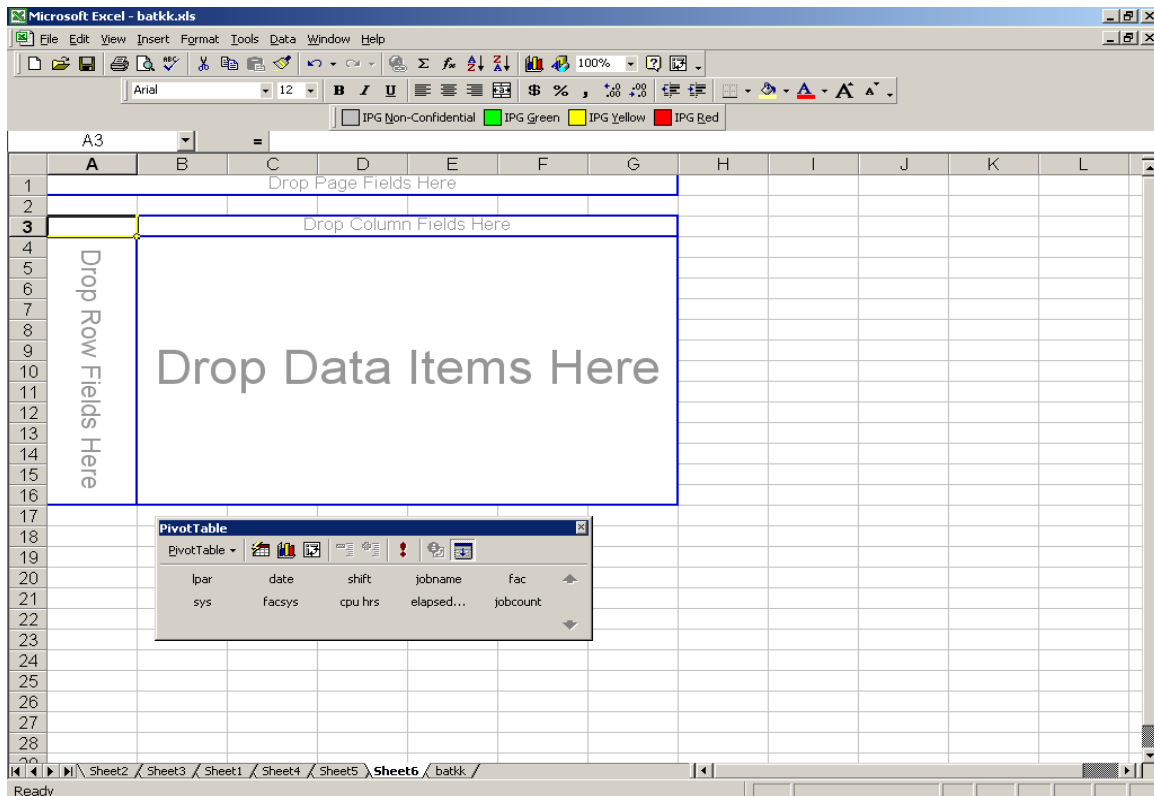


Figure 1b

```

BROWSE      TST.Z1DW.BATCH.DAILY.G1274V00          Line 00046637 Col 001 080
Command ==>                               Scroll ==> PAGE
A2S2  2008-05-27 PRIME      H0EV1830 H0 EV H0EV      0.000161    11      2
A1S1  2008-05-27 PERIOD3    H0EV1805 H0 EV H0EV      0.000003     0      1
A1S1  2008-05-27 PERIOD3    H0EV1850 H0 EV H0EV      0.000028     3      1
A1S1  2008-05-27 PRIME      H0EV1805 H0 EV H0EV      0.000006     0      2
A1S1  2008-05-27 PRIME      H0EV1850 H0 EV H0EV      0.000039     2      2
A1S1  2008-05-27 PRIME      H0EV1881 H0 EV H0EV      0.000033     1      1
BGSG  2008-05-27 PERIOD2    H1FA2010 H1 FA H1FA      0.000108     8      1
BGSG  2008-05-27 PERIOD2    H1FA2020 H1 FA H1FA      0.000003     0      1
BGSG  2008-05-27 PERIOD2    H1FA4110 H1 FA H1FA      0.000014     0      1
BGSG  2008-05-27 PERIOD2    H1FA4230 H1 FA H1FA      0.000017     1      1
BGSG  2008-05-27 PERIOD2    H1CE0096 H1 CE H1CE      0.000875    46      2
BGSG  2008-05-27 PERIOD2    H1CE0807 H1 CE H1CE      0.000692    65      1
  
```

LPAR	DATE	SHIFT	JOBNAME	FAC	SYS	FACSYS	CPU hours	Wait seconds	JOBCOUNT
------	------	-------	---------	-----	-----	--------	-----------	--------------	----------

There are three numeric fields in this data. Any of them could be the **DATA ITEM**. In our experience, we are either summing the “CPU hours” or the “job count”. When using a pivot table, you do not have to decide the type of metric you use on the numeric column, as this can be adjusted dynamically later.

When thinking about **ROW FIELDS** or **COLUMN FIELDS**, the best way is to picture a graph with an X and Y axis. Rows go across the bottom on the X axis and the columns go up following the Y axis. So when you assign a column from your original table to the **ROW FIELD**, it becomes the bottom of the chart. In our case we assigned “date” so that we created a timeline picture. Sometimes you will not need two dimensions to your analysis, but if you do, a **COLUMN FIELD** adds delineation to the values in each row. In our case, we used the table columns of “lpar” or “fac” to show which computers or which customers were using the batch job workload on each day.

Often times when first using a pivot table, there may be no obvious answer as to how the summary of data and categorization should be reported. One of the greatest advantages of the pivot table concept is that they can be altered instantly to produce varying images of the same data. Pivot tables cannot alter the original data that is being summarized. Multiple copies of pivot tables can be created using the same original data, or a single pivot table can be reused over and over to produce various results. When the neophyte user understands this precept, individual creativity will be unleashed.

Sometimes after you begin to see the summarization by different ROW and COLUMN FIELDS, you would like to freeze a particular value and see a more specific picture and not the aggregate. This is when you use a **PAGE FIELD**. Any descriptive column in your base data could be a page field. In our case, we would often use this feature to isolate specific “lpars” in order to understand day patterns.

Another way to think of a **PAGE FIELD** is a chart that showed population totals for the entire country. If you wanted to see the individual population by state, you would drag that column to the top area of the pivot table as shown in figure 1a. Then in a dropdown box, you could select any specific state. To generate groupings of states, there are other tools that could be used.

Each of the pivot table features above works on a “drag & drop” basis between the pivot table structure shown in the upper left-hand corner of figure 1a and the **Pivot Table Toolbar** shown towards the bottom in the same figure 1a. All columns selected for metric analysis or descriptive reporting will be shown in the toolbar. These column names will be the icon that can be “grabbed and dragged”. When you first build or when altering any of the **FIELDS**, the toolbar will be used. At other times, the toolbar can be minimized and retrieved with the wizard. When printing tables and charts, the toolbar will not appear.

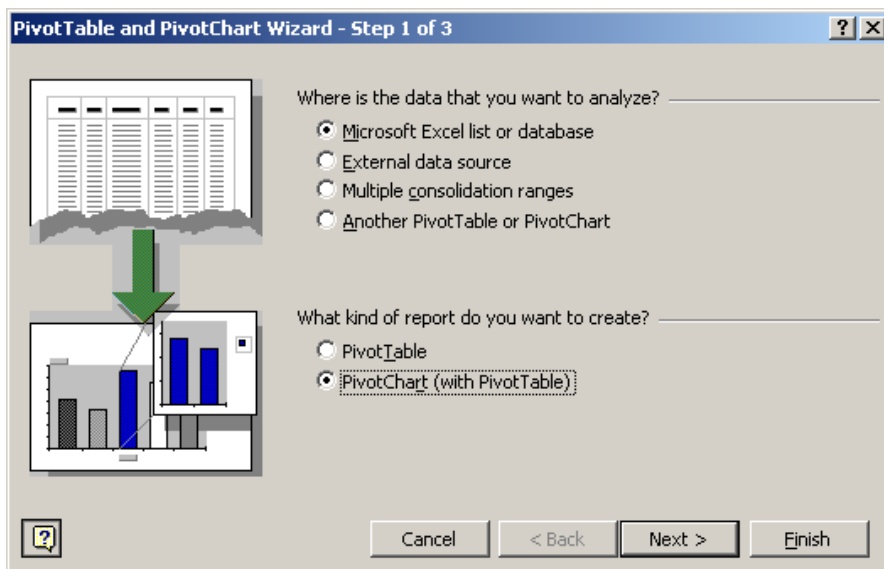
The toolbar also allows you to instantly create a pivot chart, refresh data, and other formatting features for the pivot table or pivot chart. In our case, we have established consistent methods of reporting data, so use of the toolbar is routine and minimal. This is true for many of the features for the pivot table/chart tools. There is much more for the expert user to discover and utilize.

Basic pivot table creation

The first step to creating any pivot table is to be in a spreadsheet or spreadsheet tab. The next step is to select the cells you want analyzed. In most cases, this will be the whole table. In some cases, it is useful to perform a pivot table on a limited set of data, but this can also be easily accomplished from filtering after the pivot table has been built. For the neophyte, it is recommended you always select all of the columns and rows from a table. A degree of proficiency needs to be applied when intentionally placing limitation on data that is analyzed. A pivot table can easily filter out unwanted data. There is generally no need to sub select a set of table data before making a pivot table or chart.

Once you have selected the data to analyze, you simply request EXCEL to produce a pivot table. This can be done from the DATA dropdown menu, or from a toolbar icon (if you know which one). Either way produces a popup dialogue box as shown in figure 2a. Most of the time we just press the “Finish” button. For those who want to become experts, you can follow the “Next” path.

Figure 2a



After you press “Finish”, the initial pivot table will be built (in a new tab) as shown in figure 1a. The toolbar should also be visible. If it is not, we normally scrap this tab and start over. The creation of a pivot table/chart is a “READ-ONLY” event to your original data. Nothing in this activity can alter or destroy the original tab. You can also create multiple pivot tables from any single spreadsheet. The initial pivot table is built as a framework image with the current toolbar. To complete the pivot table, the columns on the toolbar will be dragged onto the framework.

The next step will be to “drag & drop” the appropriate fields. Every pivot table has at least one **DATA FIELD**, and either a **ROW FIELD** or a **COLUMN FIELD**. A pivot table is simply a summarization and rearrangement of the data in the original spreadsheet, so now you can experiment with various combinations of variables and pivot table formats. For some data, the date looks good running down the side of the table. In other cases, a date running across the top of the table is great for comparison (see figures 3a and 3b). With simple dragging and dropping, you can create and alter tables until the presentation of data provides the answers you were looking for.

When the first column from the toolbar is dragged onto the framework, the image of the pivot table will disappear. Initially this may be confusing. Fear not! The framework is always present even when it is not shown. You simply need to drag a column onto the pivot table in the general location of a **ROW FIELD** or a **COLUMN FIELD** or a **PAGE FIELD**. The pivot table will act on the location that the column was dropped. If it isn’t what you intended, just drag the field back to the toolbar and start over.

In the case of figure 3a, the cpu hrs column was dragged to the **DATA FIELD**. The columns for date and fac were dragged to the **ROW FIELD** and **COLUMN FIELD**. In the case of figure 3b, the only difference is that date was switched from a **ROW FIELD** to a **COLUMN FIELD**, and we used the Shift column to show a different picture by time-of-day instead of fac.

Figure 3a

Sum of cpu hrs	fac					
date	Q1	Q2	Q3	Q5	Z1	Grand Total
1/1/2007	0.68	1.06	1.27	0.03	26.28	29.32
2/1/2007	0.61	1.08	1.12	0.02	60.37	63.20
3/1/2007	0.75	1.22	1.24	0.03	55.44	58.69
4/1/2007	0.65	1.11	1.19	0.03	72.83	75.80
Grand Total	2.69	4.46	4.82	0.11	214.92	227.00

Figure 3b

Sum of cpu hrs	date				
shift	1/1/2007	2/1/2007	3/1/2007	4/1/2007	Grand Total
PERIOD3	14.53	41.01	42.17	48.89	146.59
WEEKEND	11.75	17.69	14.69	21.28	65.42
PRIME	1.74	3.23	0.78	4.39	10.13
PERIOD2	1.08	1.27	1.05	1.11	4.51
HOLIDAY	0.23			0.14	0.37
Grand Total	29.32	63.20	58.69	75.80	227.01

Basic pivot table operations

The ability to rapidly arrange the data with a variety of variables from the raw data is the single most powerful feature of pivot tables. Data is used to answer questions. When the data can be presented in a manner that is parallel to the question or the way the audience thinks, then you become more successful in communicating ideas based on the data. It is so true that a GOOD picture is worth a thousand words.

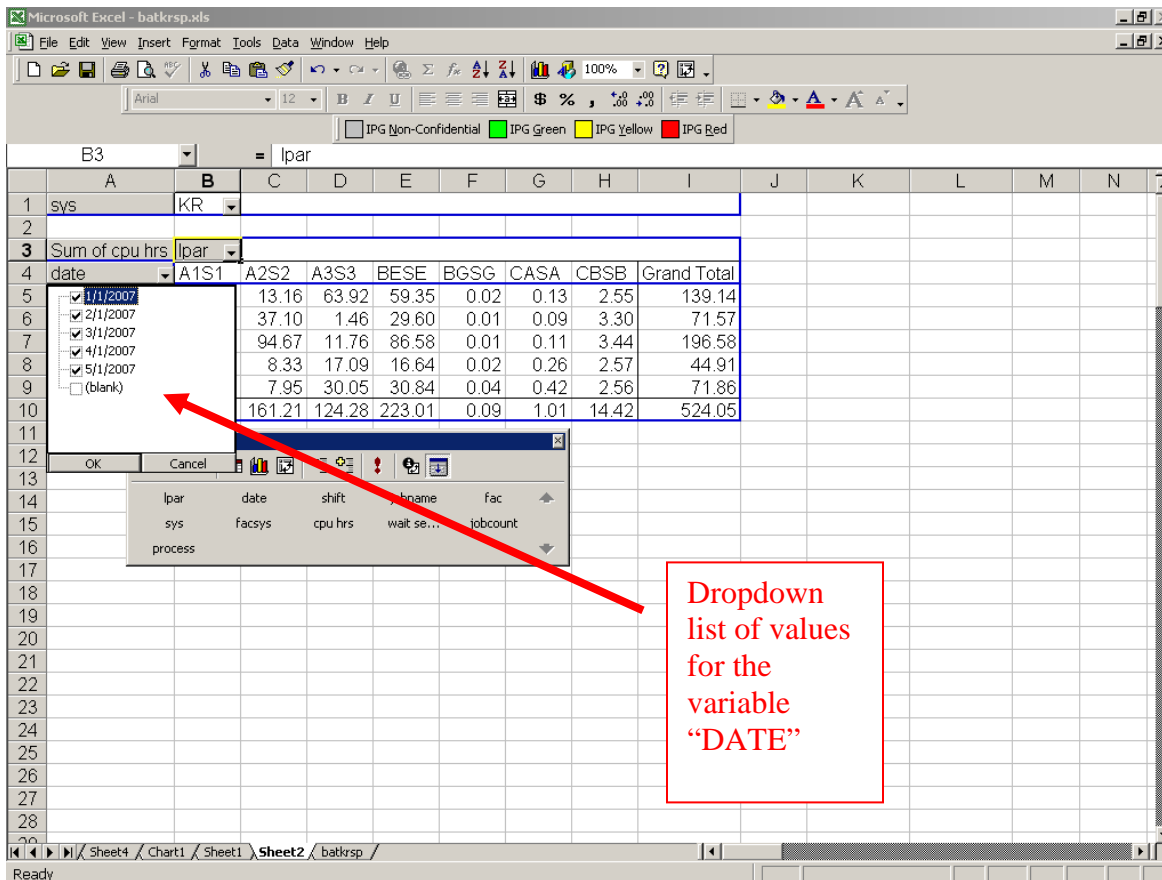
At any time, multiple variables can be used in the **DATA, PAGE, ROW, COLUMN FIELDS**. This can be useful at times to illustrate subcategories. For example, when analyzing performance data, it is often useful to show not only the total of cpu seconds for an application, but also the total of jobs. In that way, analysis quickly indicates whether a particular total was an anomaly or a pattern.

The numeric variables are normally totaled, but a variety of mathematical functions can be applied, such as MIN, MAX, AVG, COUNT. For example, our performance detail data contains the observed job or transaction unit-of-work. By looking at the average or max value of this metric, we can observe how much above baseline an application is running and the probable cause if variance exists.

Any cell in a pivot table can be "exploded" into the detail that represents it. By double-clicking the cell, a new spreadsheet tab will be built with an extraction of the raw data. For example, in Figure 3b to understand where the 14.53 hours comes from, simply doubleclick the cell and all records for PERIOD3 in JAN2007 would be extracted. This is particularly useful when you are looking for data anomalies.

Whenever a variable is used as a **PAGE, ROW or COLUMN FIELD**, the pivot table provides the option to select or hide various data values from the pivot table summarization. There will be a dropdown toggle next to the variable that opens a list of all values with checkboxes. Selecting the values will make them inclusive in the pivot table and all totals. Deselecting the values excludes the data from the presentation in the table and in any summarizations/calculations. This can be useful to exclude data from weekends if you collect data every day but only need workday information. (see Figure 4a)

Figure 4a



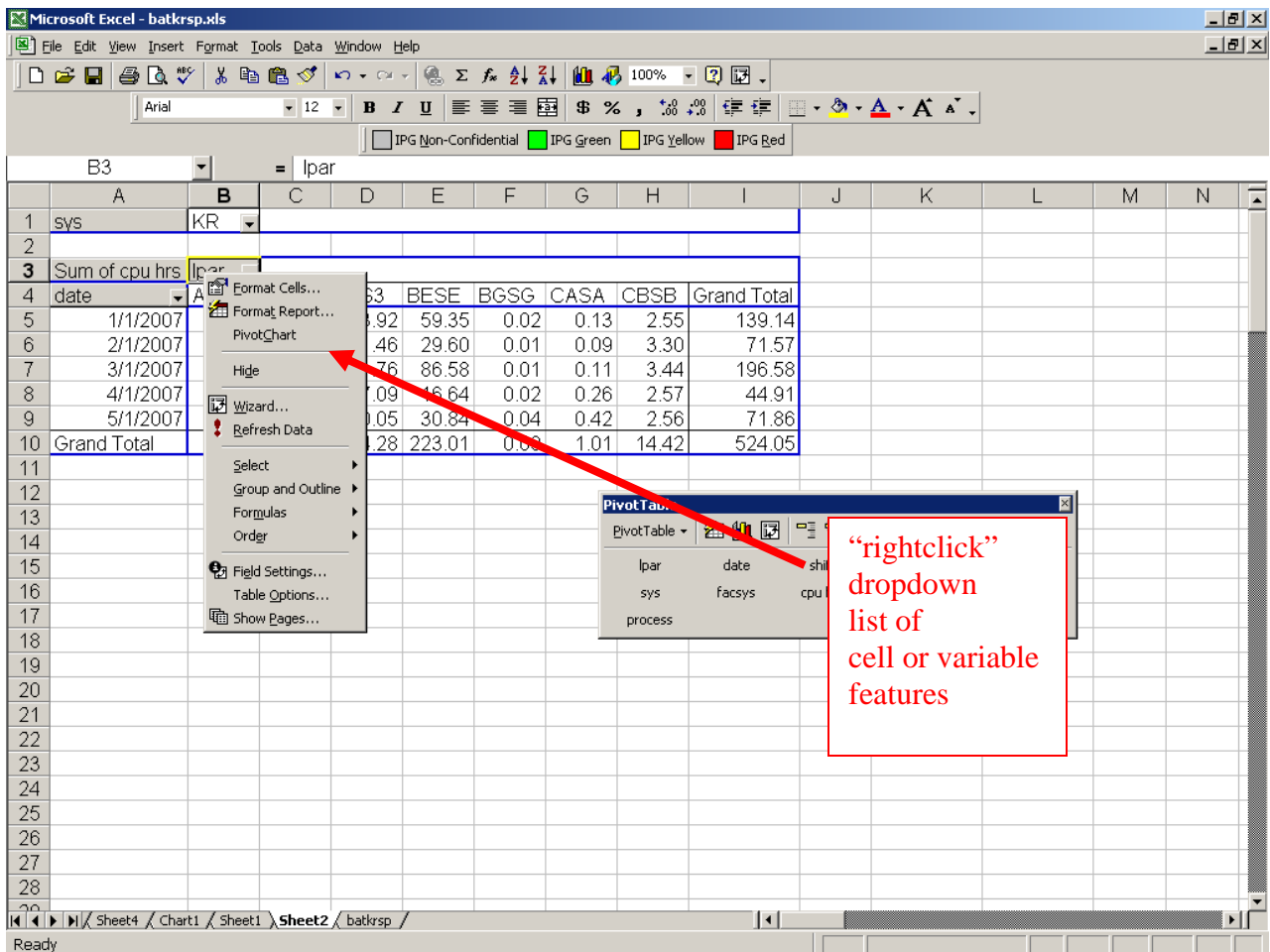
When you right-click on a **PAGE, ROW or COLUMN FIELD**, you will be presented with a variety of options to alter the format or processing of that variable (see Figure 4b). Two features the author uses the most are "Format Cells" and "Field Settings". It is important for the neophyte user to realize that a pivot table is still an EXCEL spreadsheet. Every cell can be selected and processed using a right-click or left-click.

Think of the function of "Format Cells" as being able to adjust the appearance of the values in a cell. This includes things as simple as font sizes and colors, to the format of dates and numeric data. These types of functions generally do not change the value of data in the table. Just the presentation of the cell the data is in.

The function of "Field Settings", changes the way the pivot table summarization and categorization works. This is the area where you can decide to TOTAL or AVERAGE. This is also where you can easily define a pivot table to create top 10 lists. Sorting of data and grouping can also be accomplished within the "Field Settings" popup list of functions.

Other functions provide the expert user the ability to customize reports and charts. If your data isn't quite like you want it, there are features to extend or translate the raw data while it is processed. A skilled EXCEL analyst who works on lots of pivot tables can use this like a programming language.

Figure 4b



Since the data in a pivot table is still an actual spreadsheet, you can operate on the values by referring to the cell location. For example many of our reports show the daily totals of CPU. To determine the daily average, we simply use an EXCEL formula and point to the column (or row) of daily data. After a pivot table is built, custom cell assignments can be made that enhance the value of the data being presented, or to illustrate extended metrics. We use this technique in a history pivot table. The monthly data is imported and updated in the spreadsheet while extra computation cells (on the outside of the pivot table) calculate the historical baseline. In this way, pivot table data can be used to easily report calculated metrics and constant values.

One of the icons on the **Pivot Table Toolbar** allows you to refresh a pivot table if the raw data has changed. This is particularly useful if you are collecting daily data to merge into a monthly summary. As each day (or batch) of data is collected, a simple "refresh" will update the table and all related cells.

Some care must be observed when refreshing the data in a pivot table. New data inevitably introduces new values (like another date). This will change the dimensions of the table. Even if it is by one row or one column extra, this can distort existing data, especially if spreadsheet modifications have been made that are external to the cells of the pivot table. The UNDO tool will become very useful if you happen to trash an existing pivot table due to the after-effect of some adjustment. This probably isn't an error that you will make often, but once can be enough to destroy a pivot table. So make sure that all files are routinely backed up so you could recover to a previous date.

Basic pivot chart functionality

Many times after a pivot table is built, it is useful to represent the data in a chart. One of the icons on the **Pivot Table Toolbar** allows you to immediately build a chart representing the table. By default, a vertical bar chart is built, but dozens of charts are possible by simply clicking the chart icon. See figure 5a and 5b for examples of charts that were made in a matter of seconds.

This feature once again is simply an extension of what EXCEL can already do in charting data. The primary difference is that your pivot chart will have a toolbar and all of the same basic features for the **DATA, PAGE, ROW, COLUMN FIELDS**. This makes it easy to treat your charts the same as the pivot tables, and quickly arrange data and variables to present the best picture.

In many cases when analyzing performance data, a pivot table and chart can be used to plot interval data (such as a daily/weekly/monthly total). The entire purpose of the data collection is to plot a chart that grows by the interval increment each time it is updated and published. During our enterprise peak, we plot an hourly image of utilization versus total capacity (see figure 5c). This data is imported at the end of the day, and with one click of a refresh icon in the toolbar located on the chart, the previous day is plotted by hour. While a pivot table is required to support a pivot chart, we do not even have to look at the table to produce the desired chart. Total time required is approximately 5 seconds.

In a pivot table, a cell is cursor sensitive to what is in it. On a pivot chart, any data point has this same sensitivity and will show you the actual value without requiring extensive chart labeling. However, for the more advanced reporting analysts, there is a full range of features to make charts "ready to publish". At times, we simply cut/paste them into a Powerpoint slide and annotate as we wish. Each bar or part of a bar in a chart can be acted on like the "Format Cell" feature. If your boss has an aversion to the color maroon or requires a specific date format, the charting tools can do almost anything.

Figure 5a

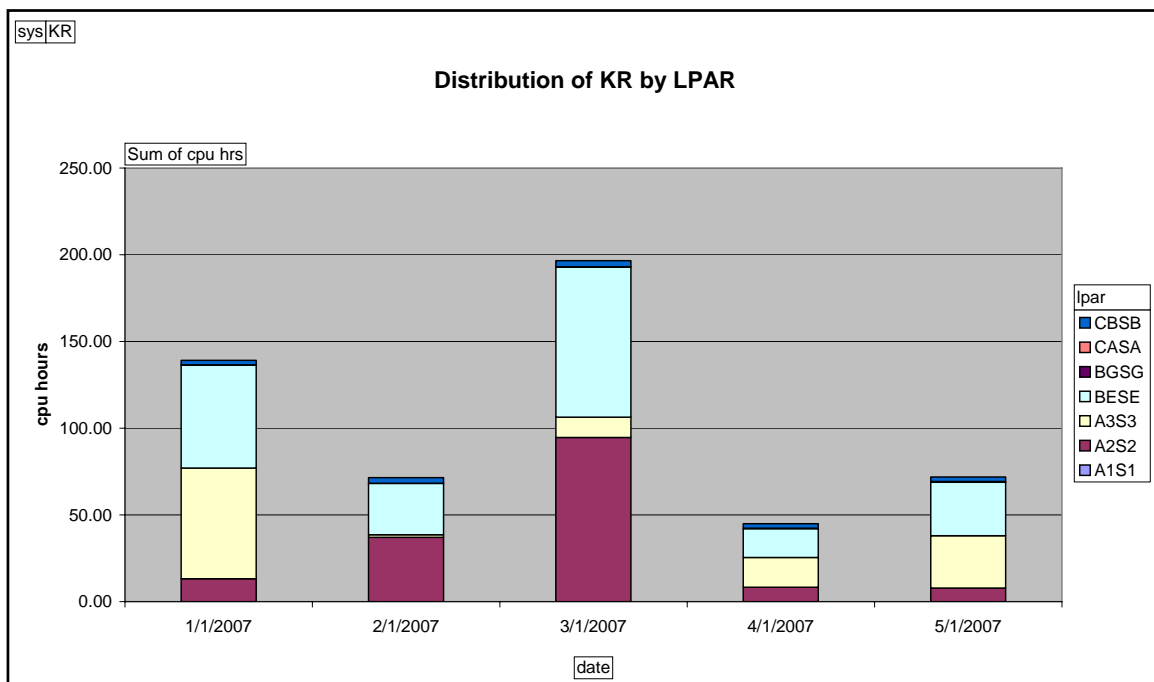


Figure 5b

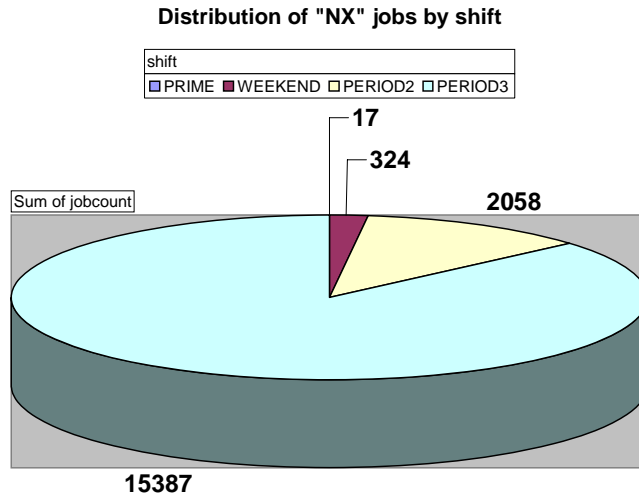
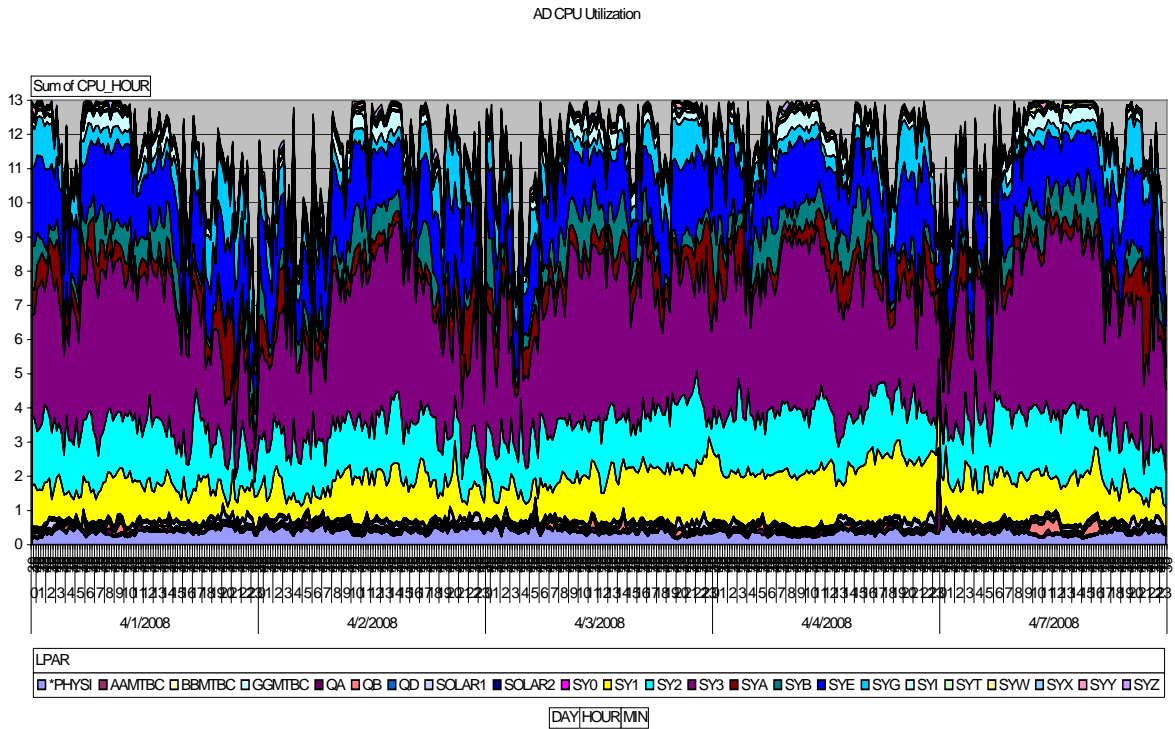


Figure 5c



Pivot tables for intermediate and advanced users

These topics are beyond the scope of this paper. However, only your imagination can limit you. The next steps are data linkage, report and chart formatting, and macro automation. These are the same advanced skills that one would need if they were to become an expert in EXCEL functions. Remember, a pivot table is just a fancy spreadsheet table. For those of you who venture into this territory, This author would ask that you provide a future training session at CMG for advanced users. This author will be one of the first to enroll.

Summary

A pivot table is a tool. Under the right circumstances and appropriate usage, this tool can be quite useful for any enterprise. At our enterprise, the circumstances involve volumes of meaningful data records that are not useful unless they are summarized. In addition, there seems to be ongoing requirements to analyze the data from different audiences. Pivot tables are so easy and quick to manipulate that we have the speed to be indecisive. Referring back to the list of questions at the beginning of this paper, one set of data can be used to answer all of those questions in a matter of minutes. A pivot table may not make you clairvoyant, but to your management you may seem like it.

CMG2008 presentation note

This paper is associated with a live demonstration of an EXCEL spreadsheet with pivot tables/charts. The data is rather simple batch job performance data. The purpose of this paper is to help the first-time user or the neophyte to increase the confidence in using pivot tables and to improve general skills. Since the creation/manipulation of a pivot table can occur in seconds, seeing a live demo of the tool is particularly effective in communicating the value and ease-of-use. Not only is there personal value in this incredible tool, an increase in data sharing becomes more prevalent. Using a common tool like pivot tables/charts to summarize and present data can increase confidence in the data and establish credibility for the author.

Note regarding version of EXCEL for this paper

This paper is based on the pivot table toolset in the EXCEL product within MS OFFICE 2000. This version has been part of the standard software configuration for our enterprise since this author began using the tool. There has been no effort applied to a cross-comparison of versions. It is the author's belief that the introductory nature of this paper will be supported by any version of the tool as offered in MS OFFICE. However, this assumption has not been validated and the author concedes that different versions may change the way some features are accessed or how they function. This is an inherent risk for every piece of software that has versions.