

Calculating Trends and Forecasting with Captell

This white paper uses the sample data supplied with CaptellV6.

Prepared by: Steven M. Dunn
MPP Pty Ltd
Email: steven@mppgroup.com.au
Phone: +61 (0)3 9343 2265
Mobile: +61 (0)4 1735 0318

Table of Contents

Introduction	3
Basic Math	3
Worked examples	4
<i>Step 1</i>	4
<i>Step 2</i>	4
<i>Step 3</i>	4
<i>Step 4</i>	4
Forecasting	6

Introduction

There is often a need to be able to calculate a trend for a specific data field to determine how its value is changing with respect to another variable. For this process to be valid (and of any use at all) the two variables must have some relationship or correlation and the data values should ideally have a normal distribution.

The usefulness of this process can be further enhanced by including forecasting. The calculated trend values can be extrapolated into the future by the addition of extra data points with a suitable period. The ability to calculate when a resource is likely to be 100 percent utilised is also needed when planning new equipment or upgrade purchases.

We will cover off in this document the necessary mathematics, the sample queries to perform this process on example data and some simple charts to demo the results. The queries charts and this document can be found in the folder named "trend_example" within the supplied Captell Samples objects.

Basic Math

The general recommendation is that in calculating the slope of a line, you should try to use a LEAST SQUARES method.

Taking it in the context of the Captell V6 reporting environment: - The generalised overall skeleton SQL for the entire query would look something like:

```
Select ((count(*) * sum(X*Y)) - (sum(X)*sum(Y))) / ((count(*) *  
sum(X*X)) - (sum(X)*sum(X))) as slope  
from [a table]
```

Worked examples

Using the Captell sample data table [Captell\dasd\DASDGIGS] we can use a sequence of simple queries to demonstrate the process and produce the trend data we require. This table contains daily disk utilisation data and we will be trending the usedgigs field to show the growth experienced within the disk storage subsystem for the measurement period.

NOTE: The row number technique used in step 1 will give the correct results for any interval duration you may require.

Step 1

Step 1 creates our X and Y variables

```
SELECT [RUNDATE],
       [NUMVOLS],[usedgigs],
       [USEDGIGS] as y,
       [FREEGIGS],
       row_number() over(order by rundate) as [x],
       [ALLOC]
FROM [Captell\dasd\DASDGIGS]
```

Step 2

Step 2 calculates the slope of the trend line using the X and Y values from Step 1

```
Select ((count(*) * sum(X*Y)) -
        (sum(X)*sum(Y))) / ((count(*) * sum(X*X)) - (sum(X)*sum(X))) as slope
from [STEP1]
```

Step 3

Step 3 calculates the intercept using the data from Steps 1 and 2

```
Select avg(Y) - (max(slope) * avg(x)) as intercept
From [STEP1],[STEP2]
```

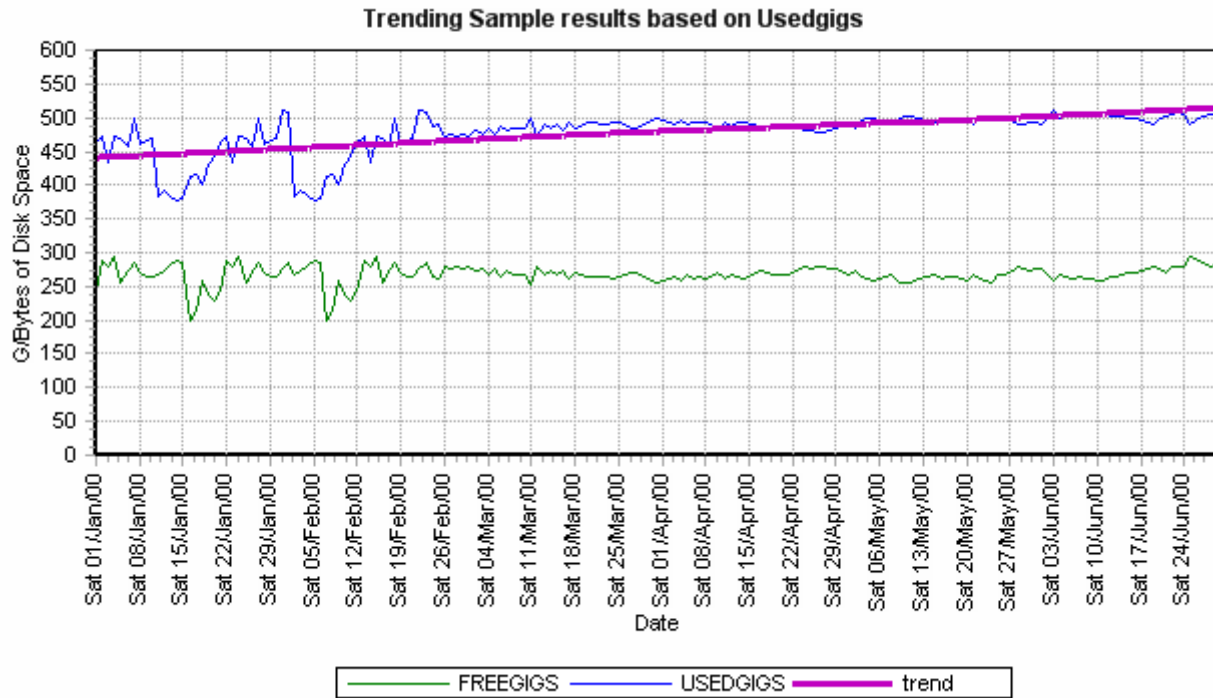
Step 4

Step 4 calculates the projections for the trend line using the results from the previous 3 steps and combines it with the other data points.

```
Select X,
       Y,
       INTERCEPT + (slope * X) as trend,
       [RUNDATE],
       [NUMVOLS],
       [USEDGIGS],
       [FREEGIGS],
       [ALLOC]
from [STEP1],[STEP2],[STEP3]
```

The end result of Step 4 is a table with the data values to plot plus the calculated projection through the data values.

The chart for the unchanged results from Step 4 looks like :-



The above chart can be found in the Captell samples as test1_line, and all of the associated calculations can be seen in the queries named test1*. They are presented as distinct processes to match the steps in the description part of this paper. The same can be achieved in less steps if required but the complexity increases, so this is a support consideration.

We have now shown how to calculate the trend for the range covered by the available data, maybe you would want to use this process as a predictive tool? In the samples we have included an example that extends this process in this way.

Forecasting

If you have a requirement to project past the end of the measured data you need to UNION on some additional X values (in this case dates etc....) to the existing table [STEP1] prior to performing the calculations in Step 4.

To allow us to use this technique as a predictive tool, we have created a table that contains a datetime field with values ranging from 1 July 2000 and 8 April 2003. The same data used in the previous example is combined with this extended date information in the query union1. All of the same calculations and processes are performed using this expanded data source. This can be seen in the queries named test10* in the Captell samples. This results in the following chart :-

