

Alternate Means of “Data Compression”

Part V: The Core Compression Mechanism

By Thomas O’Hare
Thomas@RedTile.Com

Copyright© 2004, Thomas O’Hare – All Rights Reserved.

Overview

In part IV of this series of white papers we outlined a broad general scope of using an external mechanism for extracting only subsets of data to mimic “data compression”. In this paper we will now deal with the specifics of the external mechanism to artificially create “data compression” that we have been leading up to until this point.

A Short Review

We cannot stress enough the importance of Part I of this series of white papers. If you are not totally familiar with Part I, we strongly advise that you now take the time and go back to Part I and read it again. Part I is extremely critical to the understanding of the concepts in this series of White Papers -- especially this part; “The Core Compression Mechanism”.

As a matter of fact, all the white papers in this series need to be fully understood before you proceed. Please take the time to make sure you are up to speed on the first four parts of this series before you read any further.

Defining Your Requirements

After reading this paper and before you proceed any further, sit down with a pen and paper and fully outline your requirements on what you expect to end up with as a result of using this “Alternate Method of Data Compression”. If you realize your end point now, we can better help you plan a path on achieving your desired results.

Remember, the main idea here is that by using this approach *we are in effect eliminating all the data we do not need*. Basically this means duplicate values or values that have changed so little they are what we would classify as insignificant. Bottom line is what we are doing is taking raw data and making the data more manageable by getting rid of redundant data and creating a finely tuned data subset.

The Three Main Criteria

As outlined in Part I of this series of white papers, there are a total of three entities you can use to reduce vast amounts of raw data into much smaller subsets that we have termed “compressed” data. These three entities are *Time*, *Change (Delta)*, and *Time and Delta*. With the latter of the three leaning more towards a failsafe way to make sure all

Copyright© 2004, Thomas O’Hare – All Rights Reserved.

systems were functioning correctly. So realistically we only have two very distinct entities; *Time* and *Delta*.

Your Needs

It is up to you to define which criteria you wish to use to create your “compressed data set”. We highly recommend the latter of the three mentioned above; both *Time and Delta*. As mentioned before, the reason for this is we have a dual type of failsafe mechanism built into the gathering of the data subset. We can rely on both a change in value past preset limits and a preset time of capturing a data sample. But of course you can even make your data subset even smaller by using only one of the mechanisms described at a time; Time *or* Delta. If you are comfortable with your deployed mechanisms, you can then be better able to choose single sampling criteria.

Fail Safes

If we wish to eliminate undue liability and achieve more dependable working criteria, we should concentrate on using the *Time and Delta* mechanism. It will be very easy for you to remove one or the other from this equation on your own if you deem it wise. But in the interest in clarification and stability we will concentrate on explaining both concepts working together with greater detail.

Time

Time, will of course depend on a built in timer in the external data gathering mechanism. The period of time used to gather samples or data subsets is dependent upon your individual needs. This is usually influenced by the critical nature (or lack of) of your data. *Times* should of course have the ability to be changed depending on user preferences. For example, if monitoring temperature on a day when the temperature is expected to remain in a safe high and low range, then sampling of one to four hours may be all that is required. If it is a day when temperatures are expected to be outside of safe zones, then fifteen to thirty minute intervals or less may be necessary.

Again, users should be able to manually change this timing sequences dependent upon their own requirements *and* the state of their present environment at any particular time.

The down side of only using time is great changes may have taken place but we do not know when. In other words, was there a linear progression of change or was change much more like a “spike” effect.

Delta

Using the temperature example above, we can more readily set our Delta limits.

Important: You must remember that when you set Delta limits, they span a “plus” *and* “minus” range (High & Low) from a preset *current* data point. Any variance from the

chosen central point at or past the High or Low artificial limits triggers the Delta mechanism into action.

Deltas do not have to be linear but can be set to any adjusted value. In other words, the higher limit can be greater in value from the central data point than the lower limit, or visa versa.

Delta is for the most part the better of the two criteria to use for achieving a low data storage rate, or as we say in this paper, “better data compression”. This is because we only do a record operation when a preset upper or lower limit is reached. Time has no affecting factor if using only Delta. By eliminating the Time data sampling values we in effect achieve greater “compression”.

Time and Delta

Of course if values do not reach a preset Delta for what may seem like a long period of time, we may begin to wonder if we have some sort of equipment failure in any of the mechanisms used to gather and transport data. We can try to eliminate this guess work by going ahead and also using a Time interval as an extra data point. Obviously there is a penalty to pay for this fail safe measure which is of course extra data stored in our “compressed” storage area. This achieves less “compression” but also insures we have a more accurate and especially a more reliable representation of what is actually occurring.

Conclusion

Time and Delta as a trigger to gather distinct points of data from a much larger data set in effect “compresses” our data. We can use this rationalization because the amount of data we need to do analysis with is greatly reduced in size, or as we say “compressed”.

The use of Time and Delta gives us the best of both worlds. Delta shows us if data change is occurring. Time should let us know if we are still gathering new data and our systems are functioning correctly plus we can determine if the change is linear or non-linear in nature. By combining the two criteria we have made an extremely effective failsafe method of “compressing” data to a format that is much easier and more accurate for later analysis. Again, the key to this “data compression” is by eliminating as many duplicate or near duplicate values as possible. By doing so, we create a smaller or more “compressed” set of working data.

Change Log Part V:

Original: April 9, 2004.