



Craig S. Mullins

[Return to Home Page](#)

October / November 2009



zData Perspectives

by Craig S. Mullins

[Collecting Histogram Statistics With RUNSTATS](#)

Among the many new enhancements that have found their way into DB2 9 for z/OS is the ability to gather histogram statistics with the IBM RUNSTATS utility.

A histogram is a way of summarizing data that's measured on an interval scale. A histogram is particularly helpful when you want to highlight how data is distributed, to determine if data is symmetrical or skewed, and to indicate whether or not outliers exist.

The histogram is appropriate only for variables whose values are numerical and measured on an interval scale. To be complete, let's define interval: a set of real numbers between two numbers either including or excluding one or both of them. Histograms are generally used when dealing with large data sets.

You should learn about histogram statistics because they can be quite useful to the DB2 optimizer for certain types of data and queries. Instead of the frequency statistics, which are collected for only a subset of the data, sometimes DB2 can improve access path selection by estimating predicate selectivity from histogram statistics, which are collected over all values in a table space.

Consider collecting histogram statistics to improve access paths for troublesome queries with RANGE, LIKE, and BETWEEN predicates. They also can help in some cases for equality (=), IS NULL, IN list, and COL op COL predicates.

How to Collect Histogram Statistics

OK, so how can you go about collecting histogram statistics? Well, the IBM RUNSTATS utility has been enhanced in V9 so that it can collect statistics by quantiles. DB2 allows up to 100 quantiles. You can specify how many quantiles DB2 is to use—from one to 100. Of course, specifying one isn't wise because it's the same as collecting for everything and so it isn't helpful.

You can tell RUNSTATS to collect histogram statistics by coding the HISTOGRAM keyword in conjunction with the COLGROUP option. This way you can collect histogram statistics for a group of columns. You also must tell DB2 the number of quantiles to collect by specifying the NUMQUANTILES parameter. NUMQUANTILES also can be specified with the INDEX parameter, in which case it indicates that histogram statistics are to be collected for the columns of the index.

A single value can never be broken into more than one interval. This means the maximum number of intervals is equal to the number of distinct column values. For example, if you have 40 values, you can have no more than 40 quantiles, each consisting of a single value. In other words, be sure that you don't specify a value for NUMQUANTILES that's greater than the total number of distinct values for the column (or column group) specified. Also, keep in mind that any NULLs will occupy a single interval.

So, then, how do you decide on the number of quantiles to collect? If you don't specify NUMQUANTILES, the default value of 100 will be used; then based on the number of records in the table, DB2 will adjust the number of quantiles to an optimal number. Therefore, unless you have a good understanding of the application or a viable reason to deviate, a reasonable rule of thumb is to simply let NUMQUANTILES default and let DB2 work it out.

RUNSTATS will attempt to produce an equal-depth histogram. This means each interval will have about the same number of rows. Please note that this doesn't mean the same number of values—it's the same number of rows. In some cases, a highly frequent single value could potentially occupy an interval all by itself.

Where Are the Histogram Statistics Stored?

The histogram statistics are collected in three new columns: QUANTILENO, LOWVALUE, and HIGHVALUE. These columns can be found in the following six DB2 Catalog tables: SYSCOLDIST, SYSKEYTGTDIST, SYSCOLDIST_HIST, SYSCOLDISTSTATS, SYSKEYTGTDIST_HIST, and SYSKEYTGTDISTSTATS.

Here's a sample RUNSTATS specification that shows histogram statistics for the key columns of the indexes:

```
RUNSTATS TABLESPACE DB.CSMTS02
INDEX ALL
HISTOGRAM NUMCOLS 2 NUMQUANTILES 10
SHRLEVEL (CHANGE)
UPDATE ALL
REPORT YES
```

Summary

When you've moved to DB2 9 for z/OS (and if you haven't, you soon should), consider gathering histogram statistics. Although a RUNSTATS specifying HISTOGRAM will probably consume more CPU than just gathering frequency distribution statistics, the performance gains from improved access paths for your queries could more than offset that cost.

From [zJournal](#), Oct / Nov 2009

© 2009 Craig S. Mullins, All rights reserved.

[Home.](#)