Computing Deltas - Check Sums

Computing Deltas - Check Sums

Methods for Computing Deltas in Backup Applications

While there are hundreds of different backup applications all of them use one of several known methods for computing Deltas.

Deltas are simply defined as the data that has changed since the last backup run. Defining it any further than that depends on how the backup application computes deltas. A delta could be a raw disk block, a variables length portion of a file or even a complete file depending on the method.

Check Sums

Checksums can be used to compute Deltas. The main advantage of the checksum method is that granularity beyond complete files is provided. Checksum methods all have some concept of a block where a block is either defined as fixed length or variable length. In the fixed length example a file is broken into fixed length byte ranges or blocks for example 4 KB. A unique signature called check sum is then computed based on the 4 KB. The most popular checksums for this purpose are MD4 and MD5. For more information on checksums see: http://en.wikipedia.org/wiki/Cryptographic_hash_function

The most popular algorithm for computing deltas in backup applications is the rsync algorithm. Many commerical backup applications use the rsync algorithm to compute deltas for backup purposes. One of the best known is Evault. Evault actually has a patent on the process of using variable length block deltas for incremental backup purposes. Some backup application vendors like Vembu actually brag about being based on rsync algorithm.

The are two major challenges with the process of using check sums to compute deltas.

1. The entire file system tree must be indexed and walked
2. The backup application must read all the contents of all files to compute deltas. This can take many hours or days on a large data set.

Requires Time Consuming Walk of Entire File System Tree to Compute Deltas	Yes
Delta Granularity	Block

Accuracy in Identifying Changes	Perfect unless optimized by checking file attributes
Disk I/O Impact for Small Files	Must Read All Files to Compute Deltas
Disk I/O Impact for Large Files	Must Read All Files to Compute Deltas