What is a Hash Value?

Description

A hash value results from a calculation (hash algorithm) that can be performed on a string of text, electronic file, or entire hard drive contents. The result is also called a checksum, hash code, or hashes. Hash values are used to identify and filter duplicate files (i.e., emails, attachments, and loose files) from an ESI collection or verify that a forensic image or clone was captured successfully.

Each hashing algorithm uses a specific number of bytes to store a " thumbprint" of the contents. The following is a list of hash values for the same text file. Regardless of the amount of data fed into a specific hash algorithm or checksum, it will return the same number of characters. For example, an MD5 hash uses 32 characters for the thumbprint, whether a single character in a text file or an entire hard drive.

HASH

MD5: 464668D58274A7840E264E8739884247

SHA-1: 4698215F643BECFF6C6F3D2BF447ACE0C067149E

SHA-256: F2ADD4D612E23C9B18B0166BBDE1DB839BFB8A376ED01E32FADB03A0D1B720C7

SHA-384:

2707F06FE57800134129D8E10BBE08E2FEB622B76537

A7C4295802FBB94755BBEE814B101ED18CC2D0126BD66E5D77B6

SHA-512:

C526BC709E2C771F9EC039C25965C91EAA3451A8CB43651EA4CD813F338235F495

D37891DD25FE456FE2A8CA89457629378BE63FB3A9A5AD54D9E11E4272D60C

RIPEMD-128: A868B98EAEC84891A7B7BA620EDDE621

TIGER: F31A22CEED5848E69316649D4BAFBE8F9274DED53E25C02D

PANAMA: 7E703B1798A26A0AF21ECD661CBADB9C72B419455814CA7B82E29EE0C03FA493

CHECKSUM

CRC16: 117C

CRC32: FA2D47D4

ADLER32: CF7D65FF

As you can see, there are also various length hashes within a family (SHA-1, SHA-256 et.) The most common hash values are MD5, SHA-1, and SHA-256. The longer hash values require more time to calculate and are designed to reduce the probability of a collision.

digital forensics hash verification

Image not found or type unknown

A few other ways that hash values are used:

– Verify a downloaded file was created by the publisher (as opposed to a virus-infected version)

– Identify and filter files on the NSRL/NIST list (“deNISTing”)

– Locate known contraband (illegal images and videos)

Here are a few reasons why hash values are so widely used as a means to validate and compare content:

1) Privileged Data – There would be apparent issues storing and providing multiple copies of the contents of a company’s files or entire hard drive data in a database to perform a byte comparison. Not to mention illegal images and videos (child pornography) would have to be stored and used in each system scan.

These scenarios are unacceptable.

2) Speed – Comparing an indexed hash value versus what could be billions or trillions of bytes or source data is much quicker. Optimized hash engines (Pinpoint Harvester) can compare thousands of hash values in a second.

3) Security – Hashing data is a one-way trip. The original data can't be recreated or reverse-engineered from the hash value. This provides additional security because a person can't determine the source data from the hash.

The argument that data sources could be different and have the same hash value has raised much concern. Countless threads related to this issue on litigation support and computer forensic forums exist. The bottom line is that the only way to compare the original data is to store it everywhere you need to deduplicate or verify the information; however, as mentioned, this isn't a practical alternative.

More complex hashing functions have been introduced (SHA-256, SHA-512, etc.), reducing the likely hood of a collision. It is also worth noting that even in those cases where scientists have created collisions, it resulted from exploiting the weaknesses in a specific hash algorithm. The same alterations would not create a collision in a different hashing algorithm.

So, if you still aren't satisfied with the incredibly remote possibility a collision could happen using a single hash value, then the easiest way to implement an extra precaution is to take the time to have your processes calculate hash values from two separate algorithms (i.e., MD5/SHA256) for each item. Unfortunately, most EED applications and forensic imaging tools don't support this option, especially in a single pass.

What to Remember

Hash values are a reliable, fast, and secure way to compare individual files and media contents. Whether it's a single text file containing a phone number or five terabytes of data on a server, calculating hash values are an invaluable process for Deduplication and evidence verification in electronic discovery and computer forensics.

Date
04/18/2025
Date Created
12/10/2010