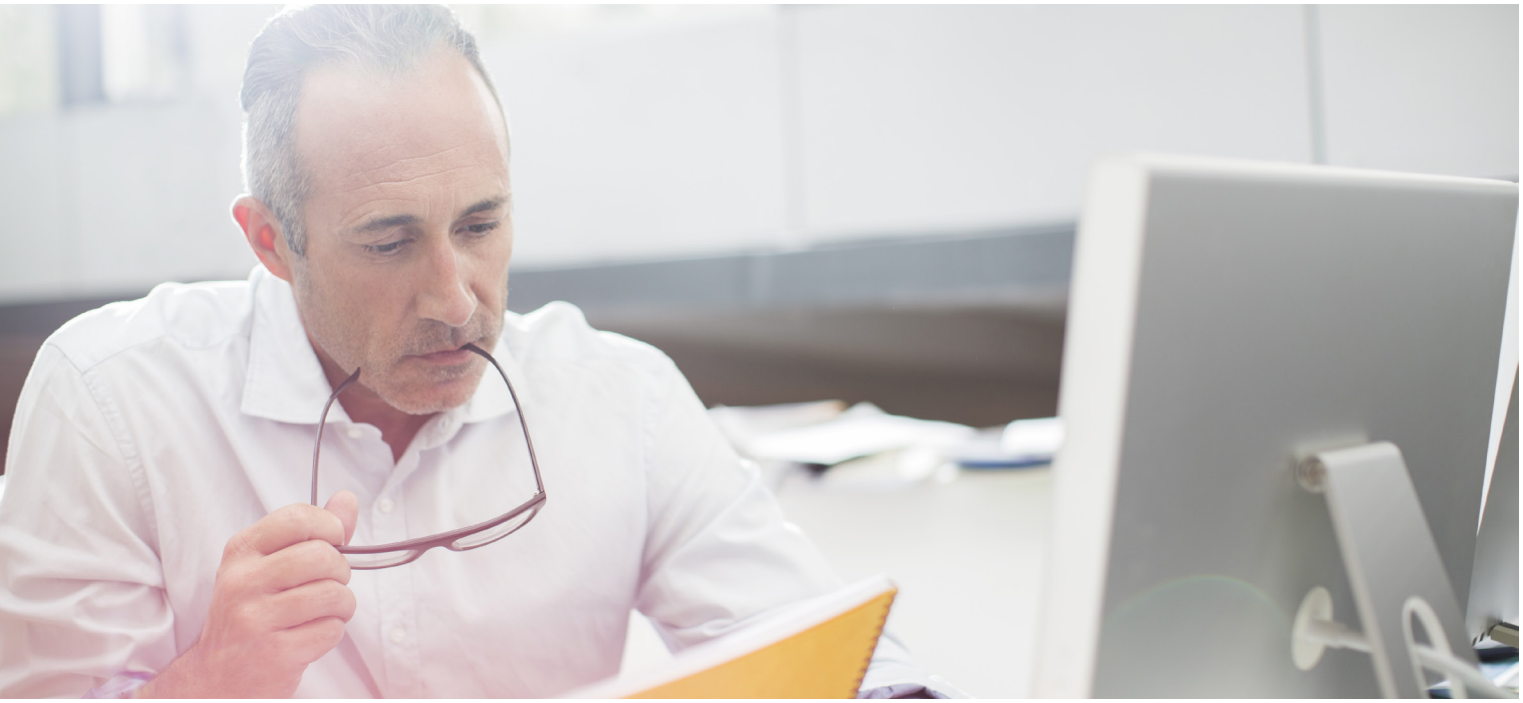
Quest

Deduplication: The hidden truth and what it may be costing you

Not all deduplication technologies are created equal. See why choosing the right one can save storage space by up to a factor of 10.

By Adrian Moir, Senior Consultant, Product Management, Quest Software



Data growth is something we all have to contend with. We have to store more and more data for longer and longer time to satisfy business or regulatory requirements. The global data-sphere will likely grow to 163 zettabytes by 2025, by which time almost half of it will reside in the enterprise.[1]

Protecting your data is a primary business function. It comes with its own set of challenges. Add those to the ever growing and shifting data sets, and it starts to eat into resources quickly.

While protecting data of course is necessary, we do not want it to be a burden. We all look for solutions that can move data from point A to point B as fast as physics will allow! We look towards technology to make the experience better, go faster, store more while forever reducing costs. Quite a challenge indeed.

We have seen many technologies evolve that help smooth the path to secondary data storage optimization, with a view to reducing the resource burdens of backup and recovery. One such technology is deduplication. While deduplication has been available in the market for some time, not all solutions are equal, and it is often useful to understand exactly what the impact of each is likely to have on your available resources and budgets.

NOT ALL THINGS ARE CREATED EQUAL

This is the same for deduplication technologies. Deduplication has become a byword for data reduction; however, using a single word to describe different methodologies can be misleading at best.

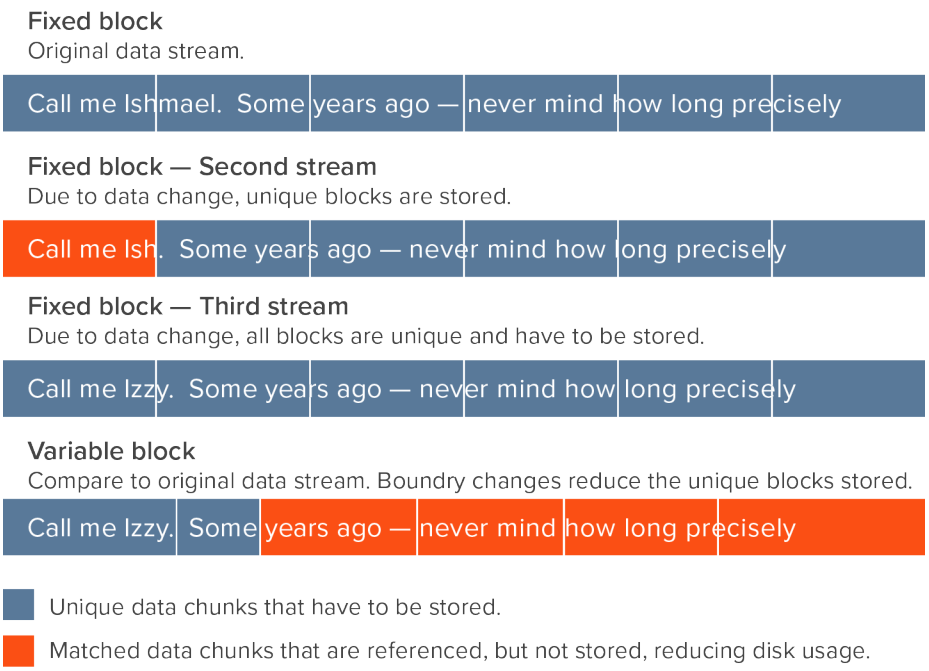


Figure 1: Fixed-block de-duplication limitations.

Fixed-block deduplication is a good start, however it is limited. It only works well on some data types that are stored directly on the file system.

Let us consider some different data reduction technologies:

Compression

A good example is lossless data compression. It relies on exploiting statistical redundancy without losing data, so you can 'undo' the compression and reinstate the data. We have been using these techniques for many years, now found as default in technologies. For example, a GIF image will use LZW (Lempel-Ziv-Welch) compression to reduce the file size of the image without losing information.

Single instancing

Single instancing describes a storage methodology quite well. If I store the same file twice, then I will keep one and reference it for the other. However, this only works if the files are identical. Change a small thing in the file and the whole file is stored again. Now this is great for systems that have their structure populated with the same content, production storage perhaps, but not so well for general data protection.

Fixed-block de-duplication

This is our first foray into what would be described as a proper deduplication methodology. Fixed-block deduplication takes a stream of data and slices it up into blocks of a fixed size. We call these 'chunks' of data.

These chunks are then closely compared using several methods, and if they are deemed the same, then only a single 'chunk' is stored and a reference is kept for each subsequent match. Sound familiar? Think of this as single instancing but at a subfile level, looking at the blocks that make up that file.

This methodology sounds better, however it is limited. It works well on some data types that are stored directly on the file system, since they are byte-aligned and have their file systems written in 4k, 8k, 32k chunks, like virtual machines, for example.

In this case, fixed-block solutions can be very effective. However, this does not work well on a mixture of data where those boundaries are not consistent or if they are backed up through different types of software, which changes the alignment. We know data is anything but consistent, and depending on its type, will have a different make up, block size, byte alignments and content.

ABOUT QUEST

At Quest, our purpose is to solve complex problems with simple solutions. We accomplish this with a philosophy focused on great products, great service and an overall goal of being simple to do business with. Our vision is to deliver technology that eliminates the need to choose between efficiency and effectiveness, which means you and your organization can spend less time on IT administration and more time on business innovation.

If you have any questions regarding your potential use of this material, contact:

Quest Software Inc.
Attn: LEGAL Dept
4 Polaris Way
Aliso Viejo, CA 92656

Refer to our website (www.quest.com) for regional and international office information.