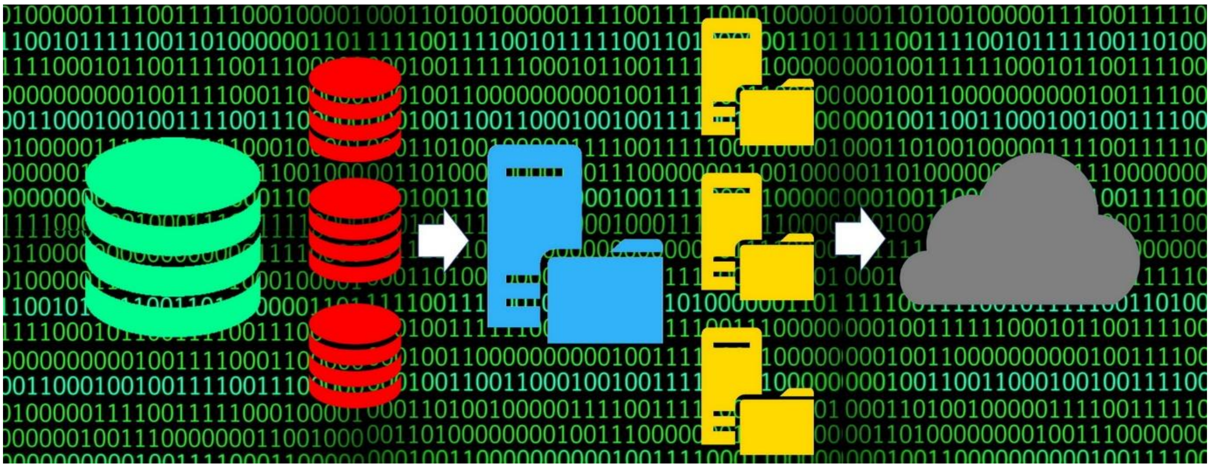
Data is Never Deleted!

Shôn Ellerton, May 20, 2021

I can 99.99 percent guarantee that your data is never deleted on a commercial system!



A friend of mine recently recommended me to watch the British TV police drama series, *Line of Duty*, and, as a result got me helplessly hooked on it. It is thoroughly engaging no doubt; however, there was a major flaw which I picked up on almost immediately. There was a sequence showing one of the police officers deliberately deleting incriminating evidence from the police computer. Now, had this data was still available, the murder case in the first series could have been solved in one fell swoop. But no. According to the story, the data was lost which meant a prolonged investigation requiring the anti-corruption arm of the police to be involved.

Now in the real world, I can guarantee 99.99 percent that this data would ***not*** have been deleted. I will also say this. 99.99 percent of the time, ***your*** data stored on third-party commercial systems (social media platforms, banks, police, etc) is ***not*** deleted either, even if you have your data requested to be deleted or after a time where it is stated in terms of legalities that the data must be deleted. Not only does this apply to all those social media sites but this also applies to passwords in which many databases are set up in an amateur or even nefarious way in which passwords are held as plain text or held as easily decipherable.

Data architects, like me, are often one of those peculiar individuals which seek to debunk IT practices on fictional police dramas on TV or have the tendency to archive everything as humanly possible so we can access something that we may, *possibly*, need in the future with ease. The reality is that, most of what we archive, will probably never ever be looked at again or used, *BUT*, there is that 'gem' moment when, as if, by magic, we either need something that most mortals would have long forgotten about or to re-discover something from the past as if it's something brand new. Many in the profession of data management tend to be archivists, an activity very much shared with keepers and custodians of knowledge like librarians. Archivists tend to have a predilection to be organised; however, they often transgress unintentionally into the world of hoarders often on the basis that they do not want to lose data. For example, one of the main afflictions of the archivist is the accumulation of backed up hard drives (usually piled in various locations around the house), the love of creating lists for everything, and of course the overzealousness to ensure everything that can be alphabetised *is* alphabetised. However, there is a cost factor. Storage, or the lack of it. My garage is crammed with things I will probably never use again. And then I need to create a list or spreadsheet to show me *where* the items are. The stuff in the garage is the *data* and the list to show me where it is, the *metadata*.

This mindset suggests to me one crucial thing.

Data is NEVER deleted!

Why?

There are two main reasons and a third reason which I will come to at the end.

1. The fear of losing the data.

2. And entropy (I'll explain in a bit).

The first point, the fear of losing data, is reasonably obvious. If one's job is to ensure that data is safe, one ensures that a sufficient number of backups are made on a frequent basis. What this invariably means is that numerous backups of databases, file systems, configuration files, or anything in the digital domain will be stored, often in many different locations. There is probably no data manager who will issue the SQL command, 'DROP', to delete an object without first creating a backup of it. Much like hoarders don't like throwing things away, archivists (and librarians) don't like deleting or removing anything.

And the second point, being entropy, is the unintentional side-effect of creating numerous backups and then forgetting about them or needlessly duplicating them, because it is often the case, that managing those backups becomes another exercise in organising and managing the data, or metadata, *about* the backups of the data which are frequently being backed up! If you're confused, I fully understand. In simple terms, the orderly way in which data is organised can, and often turns out, chaotic, if not governed under strict control. And, unfortunately, in most cases in the real world, this does not happen. It is likely that, *somewhere* is a copy of the data which, in some cases, maybe *should* have been deleted if it was meant to be deleted.

And this is before the advent of the popularity of BLOB containers and datalakes in the cloud making it worse which, to make a very crude analogy, are giant rubbish skips in which you chuck anything data-related you *might* find useful, the difference being is that you need the right data mining tools and the expertise of data science at one's disposal to hunt for them later. When data storage is cheap, many are not that exacting in what not to throw in the datalake or not.

It is not difficult to visualise how easy it is for chaos to ensue.

I work in an environment in which we take daily backups of all our operational databases and then store into a data warehouse. We ***extract*** the data. Then we ***load*** and ***transform*** it into a data warehouse. This means that all the data, or certainly all the updated, deleted and added data is added as a snapshot into a data warehouse, so ***at any point in time***, we can go back and view the data as it was then. Think of it as a database with an additional ***time*** dimension.

This all sounds fine especially when there are third-party tools which are available to manage your database backups on the file server. But what of the *file server*? Backups of the file server are normally carried out by the *system administration* teams who are not concerned with what is stored on the databases specifically, but of *all* the data stored on the servers. And when all the storage is running out? Compress it and put the oldest of it in cold storage or even magnetic tape. This basically means that, if, for some reason that a database backup was not performed on a specific day, it is often always possible to get the system administrators to retrieve and restore a portion of the file system in which the database was on.

Now one may be wondering what happens if data is deleted in a database *before* it gets updated to the data warehouse? Is the data truly deleted? Probably not. Certainly, in business-critical systems or in which the data is of prime importance, most well-designed operational databases will *flag* a piece of data as being deleted and hide it from user interaction. Sometimes this data is *never* deleted even after being updated into the data warehouse, but even if it is, a snapshot of the data can be retrieved at any moment in time.

So back to the police drama, *Line of Duty*. The simple thing to do would have been for the police officer to send an email to the IT guys and get them to retrieve the lost data!

There is a third reason that data is never deleted, which I promised to cover at the end of this piece, and this alludes to the possible intention of being nefarious. Data is power. Data can be bought. Data can be used for sinister reasons like blackmail or subversiveness. Taking a gloomy non data-related analogy, how can we be certain smallpox had been eradicated?

Taking these three reasons into account. Fear of losing data, entropy and retention for nefarious purposes. Data is never deleted. Or, at least, almost never.