



Grid-Tools
The power of test data

Why data encryption is not data masking

© Grid-Tools Ltd

Why Data Encryption is Not Data Masking

A common misconception within the data community is that encryption is considered a form of data masking – even worse is that there are some that erroneously identify both as one and the same. This white paper seeks to clarify these issues – we'll cover the fundamental differences between data encryption and data masking and discuss the ramifications from a data security standpoint.

They are solutions to different problems

The most immediate difference comes to mind when one thinks about what problems each method was developed to solve. Encryption is a method for securing communications from unauthorised eavesdropping – indeed one of encryption's enduring uses has been to protect secret military information (one of the most famous instances of this being the Enigma code used by the German military in the Second World War). This gives rise to a very important property of encryption algorithms which will be very important to us: namely, that every encryption algorithm is reversible (given the proper information). Their strength lies in the difficulty of brute-forcing the reversal without the decryption key – however the key point that needs to be stressed is that every algorithm *can* be brute forced (though for the strongest algorithms it requires more time than the universe has already existed!). Most algorithms derive their strength from the fact that factorisation of very large numbers is a time-consuming task and there are no shortcuts available (these problems are known as *NP-complete* in terms of computational complexity theory).

Data masking, on the other hand, is a methodology intended to protect the content of data in non-production environments while ensuring it maintains the referential integrity of the original production data. Since the only purpose is to protect the data with no aim to re-construct the original data, we would prefer a data masking method to be irreversible - this brings us to the fundamental difference between encryption and masking:

Fundamental Difference : For encryption, reversibility is required; for masking, reversibility is a weakness.

If a masking algorithm is reversible then it is potentially weak as it still contains the original information but in a different form. The best methods for data masking are those that are not based on the original data at all – random generation is an ideal candidate here, as there is no possible way of working out the original data given only the masked data.



Mathematical Differences

Next we shall discuss the main mathematical differences between encryption and masking in light of what was discussed in the previous section. We have already discussed the most important difference between the two: namely that of reversibility. We shall briefly discuss some corollaries to this fundamental difference.

Given a unique input, an encryption algorithm only has one possible output

To obtain reversibility, one property required is that the transformation is 1-1 (or in mathematical terms, an *injective* map). For most applications, however, the output step is usually the same as the input step (since a byte only has 256 possible unique values), and so we must have that each member of the set can be mapped onto. This is known as a *surjective* map, and a map that is both injective and surjective is said to be *bijjective*. Thus encryption maps must be bijective. To illustrate this, we'll use a small example that should suffice. Suppose we have a rudimentary encryption algorithm that acts on a small set of letters:

`{ a, b, c }`

One possible bijective map would be as follows:

`a ↦ b`

`b ↦ c`

`c ↦ a`

Notice that each letter is represented once and only once on the right hand side. This means it is a bijection. We then have the decryption step defined as:

`a ↦ c`

`b ↦ a`

`c ↦ b`

Notice that, as there is only one possible output given each input, it makes the algorithm *deterministic*. Because data masking algorithms needn't be reversible, it follows that data masking algorithms needn't be deterministic either, and in fact if a map is neither 1-1 nor deterministic then it can be considered to be a very effective map. An easier way of thinking about a non-deterministic map is to consider it as a map that, even if you know all possible inputs (and even if you know how the algorithm works) it is not possible to work out what the output will be (in relation to the input). There are methods of achieving this which we



will not go into here, but it suffices to say that any masking map that does not use the original data can be described as non-deterministic (with respect to the input) as it is not a function of the input.

The Consequences of Reversibility

In addition to the concept of non-determinism, there is a more immediate consequence to reversibility – that is, if it can be reversed then it can be cracked. Strong encryption methods are often advertised as uncrackable but there is one very important caveat that is often omitted: *within a realistic time*. It has been shown that the strongest encryption algorithms require more time than the universe has existed to crack, but this is based on current bounds of computational power (which we know will not stay constant over time). The truth is that every encryption method is crackable given their reversibility.

Irreversible data masking, particularly random generation, has none of these problems, due to their inherent nature. *There is nothing to crack*. The original data does not exist in any form hence it cannot be reproduced. From a security standpoint, this is a very important point that is rarely hammered home.

Security Perspective

Now that we have discussed all the differences from a mathematical point of view, it is important for us now to discuss the differences in a more practical context – data security. To reiterate, because of the reversibility of encryption algorithms it makes them unsuitable as masking functions because the original data is still in there; and we have already seen that almost all properties of encryption algorithms become weaknesses when used as masking functions.

From a data security point of view the best masking solution is random generation as it is completely independent of the underlying data. Encryption does not constitute good masking – we have already seen that because we do not need reversibility we can abandon the concept of one input one output (a 1-1 map) and the concept of determinism. Abandoning these two core principles of encryption allows for more secure data masking solutions.

In addition to the security implications, there are a few practical considerations to take into account. Primarily, encryption methods are very slow compared to pure random generation – so using encryption methods as masking solutions actually slows down your data masking (especially if you are using asymmetric encryption methods – where the encryption and decryption steps are different – such as RSA encryption). This becomes especially important when an enormous amount of data is masked – the cost savings of an efficient mask could be a factor.



Summary

To summarise our examination of encryption and masking: if you want to protect your production data from unauthorised entry but the data is important in its current context: use encryption; however if you require to use your production data in a test environment, where the actual content of the data is meaningless, then use masking. Not only is masking more secure than encryption, you may also find it to be a much more efficient process.

About Grid-Tools

Headquartered in Oxford, UK (with offices in Chicago and India), Grid-Tools specializes in data masking, test data generation and test data management tools, solutions and services. Our experienced personnel have been writing and developing solutions for large companies in both the private and public sectors for nearly 30 years.

Grid-Tools

11 Oasis Business Park

Eynsham

Oxfordshire

OX29 4TP

info@grid-tools.com

UK: +44 (0) 1865 884 600

USA: 1 866 563 3120

