

Demystifying data de-duplication: choosing the right solution

Author: Alexandre Delcayre, technical director EMEA, FalconStor (www.falconstor.com)

The complexity and prevalence of IT systems continues to cause storage volumes to grow exponentially; regulatory regimes are also having a significant impact as businesses are bound to store and protect data for longer. Although compression technology can deliver an average 2:1 data volume reduction, this is only a fraction of what is required to deal with the data deluge most companies now face.

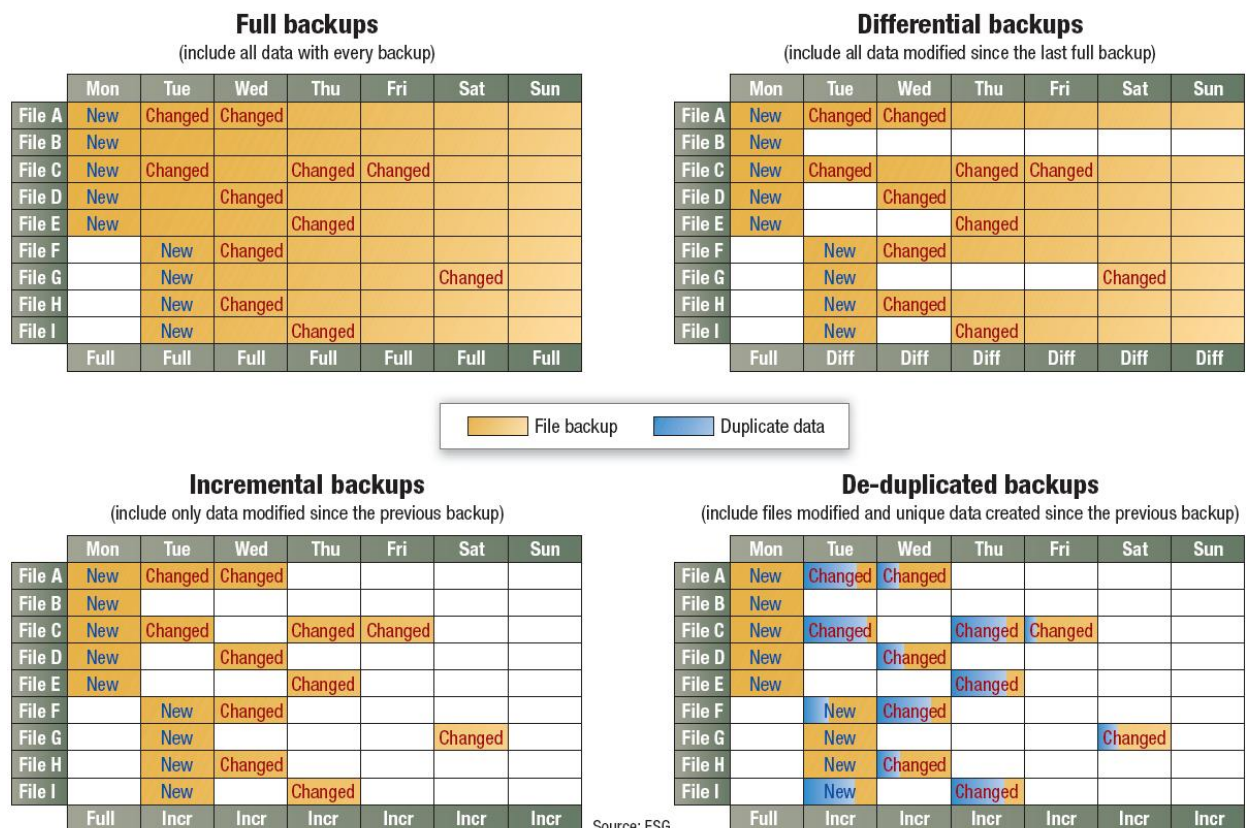
Only data de-duplication technology can meet the requirements companies have for far greater reductions in data volumes and, accordingly, data de-duplication is fast becoming a required technology for any company wanting to optimise the cost-effectiveness and performance of its data storage environment.

In this article we examine the eight key criteria to use when evaluating data de-duplication solutions:

- Focus on the largest problem
- Integration with current environment
- VTL capability
- Impact of de-duplication on backup performance
- Scalability
- Distributed topology support
- Real-time repository protection
- Efficiency and effectiveness

Focus on the largest problem

Are you going after the right one? The largest problem is backup data in secondary storage. The following graphic, courtesy of Enterprise Strategy Group, 2007, illustrates why a new technology evolution in backup is necessary:



Incremental and differential backups were introduced to decrease the amount of data required compared to a full backup. However, even with incremental backups, there is significant duplication of data when protection is based on file-level changes. Across multiple servers at multiple sites, the opportunity for storage reduction by implementing a data de-duplication solution becomes huge.

Integration with current environment

Who wants disruption? Many companies are turning to virtual tape libraries (VTL) as a non-disruptive way to improve the quality of their backup without changes to policies, procedures, or software. This makes VTL-based data de-duplication the least disruptive way to implement this technology. It also focuses on the largest pool of duplicated data: backups. Solutions requiring proprietary appliances are also less versatile than those providing more deployment flexibility.

VTL capability

Is your VTL up to the task? If data de-duplication technology is implemented around a VTL, the capabilities of the VTL itself must be considered as part of the evaluation process. It is unlikely that the savings from data de-duplication will override the difficulties caused by using a sub-standard VTL.

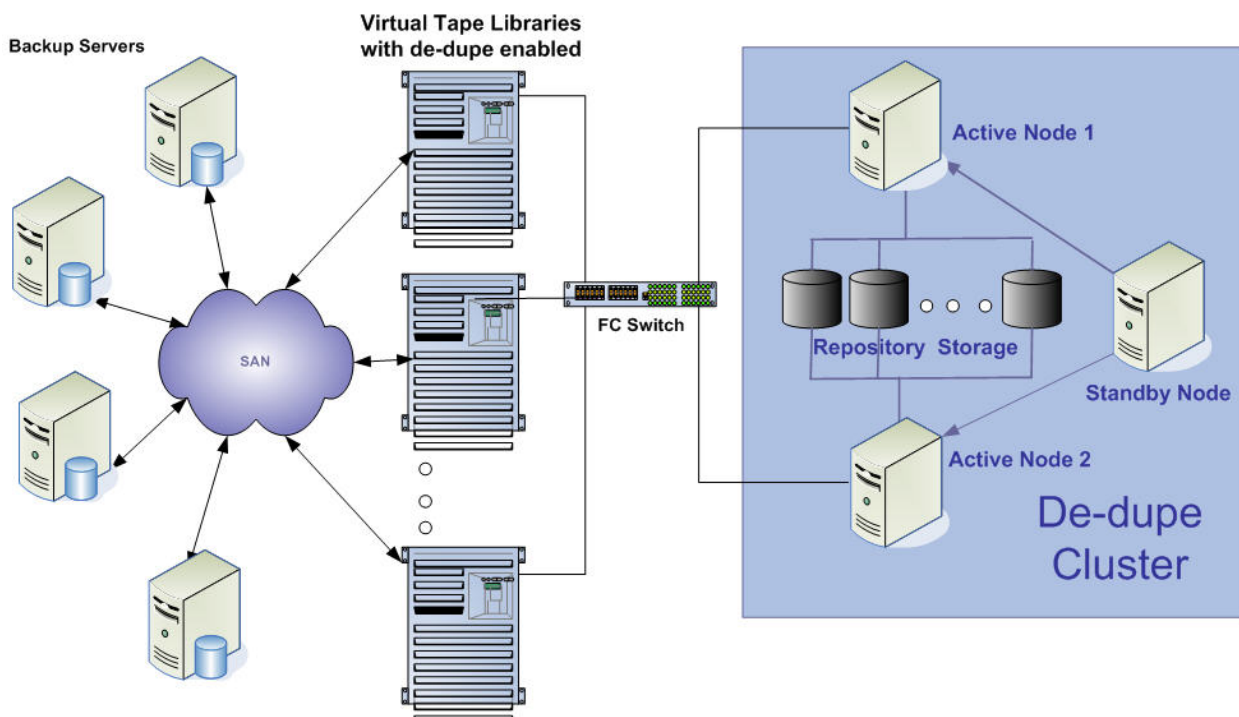
Impact of de-duplication on backup performance

Where and when does de-duplication take place in relation to the backup process? Although some solutions attempt de-duplication while data is being backed up, this approach can, over time, degrade VTL performance by as much as 60 percent. By comparison, data de-duplication solutions that run after backup jobs complete eliminate this problem and have no adverse impact on backup performance.

Scalability

Consider growth expectations over five years or more. How much data will you want to keep on disk for fast access? How will the data index system scale to your requirements? Because a de-dupe solution is chosen for longer-term data storage, scalability, in terms of both capacity and performance, is an important consideration.

The solution should provide an architecture that allows economic 'right-sizing' for both the initial implementation and the long-term growth of the system. A clustering approach provides the ability to scale to meet growing capacity requirements - even for environments with many petabytes of data - without degradation to the de-duplication efficiency or system performance.



This architecture also provides for failover support as described in the section on repository protection.

Distributed topology support

Where is your data? Data de-duplication is a technology that can deliver benefits throughout a distributed enterprise (e.g. with multiple offices) not just in a single data centre. A solution that includes replication and multiple levels of de-duplication can achieve maximum benefits from the technology.

A not unconnected consideration is bandwidth: a solution should require minimal bandwidth for the central site to determine if the remote data is contained in the central repository or not. Only the unique data across all sites should be replicated to the central site and hence to the disaster recovery site, otherwise bandwidth requirements swell.

Real-time repository protection

Can the resulting data store possibly be vulnerable? Access to the de-duplicated data repository is critical and should not be vulnerable to a single point of failure. A robust data de-duplication solution will include mirroring to protect against local storage failure as well as replication to protect against disaster. The solution should have failover capability in the event of a node failure. Even if multiple nodes in a cluster fail, the company must be able to continue to recover its data and respond to its business.

Efficiency and effectiveness

How rigorous are you prepared to be? Will your solution cope? File-based de-duplication approaches yield much less storage reduction than methods that analyze data at a sub-file or block level. Consider, for example, changing a single line in a 4MB presentation. In a file-based solution, the entire 4MB file must be stored, doubling the storage required. If the presentation is sent to multiple people, as presentations often are, the negative effects multiply.

Most sub-file level de-dupe processes use some sort of 'chunking' method to break up a large amount of data - such as a virtual tape cartridge - into smaller pieces to search for duplicate data. Larger chunks of data can be processed at a faster rate, but less duplication is detected. It is easier to detect more duplication in smaller chunks, but the overhead to scan the data is much higher.

Focus on the total solution

With stored data volumes continually increasing due to the demands of business applications and regulatory requirements, data de-duplication has become a vital technology. Data de-duplication is the only way to dramatically reduce data volumes, slash storage requirements, and minimise data protection costs and risks.

Although the benefits of data de-duplication are dramatic, organisations should not be seduced by the hype sometimes attributed to the technology. No matter what the approach, in the final analysis the amount of data de-duplication that can occur is driven by the nature of the data and the policies used to protect it.

In order to achieve the maximum benefit of de-duplication, organisations should choose a data de-duplication solution based on the total set of requirements described above - not just the biggest theoretical data reduction ratio they hear about.