

Big Data A Big Backup Challenge

Backing up Big Data requires a system that is fast, cost effective, and reliable. These are conflicting terms in the world of storage.

By George Crump, [InformationWeek](#)

June 10, 2011

URL: <http://www.informationweek.com/news/storage/230500250>

Big Data is, well, big, and size is not the only challenge it places on backup. It also is a backup application's worst nightmare because many Big Data environments consist of millions or even billions of small files. How do you design a backup infrastructure that will support the Big Data realities?

First, examine what data does not have to be backed up at all because it can be easily regenerated from another system that is already being backed up. A good example is report data generated from a database.

Turn the mobile device management challenge into a business opportunity.

[Discover four strategies to secure your mobile environment.](#)

Once this data is identified, exclude it. Next, move on to the real problem at hand--unique data that can't be re-created. This is often discrete file data that is feed into the environment via devices or sensors. It is essentially point-in-time data that can't be regenerated. This data is often copied within the Big Data environment so that it can be safely analyzed. As a result, there can be a fair amount of redundancy in the Big Data environment. This is an ideal role for disk backup devices. They are better suited for the small file transfers and, with deduplication, can eliminate redundancy and compress much of the data to optimize backup capacity.

Effective optimization is critical since Big Data environments are measured in the 100's of terabytes and will soon be measured in the dozens of petabytes. It is also important to consider just how far you want to extend disk backup's role in this environment.

Clearly deduplicated disk is needed, but it probably should be used in conjunction with tape--not in replacement of it. Again, often a large section of this data can't be regenerated. Loss of this data is permanent and potentially ruins the Big Data sample. You can't be too careful and, at the same time, you have to control capacity costs so that the value of the decisions that Big Data allows are not overshadowed by the expense of keeping the data that supports them. We suggest a Big Data backup strategy that includes a large tier of optimized backup disk to store the near-term data set for as long as possible, seven to 10 years worth of data being ideal, then using tape for the decades worth of less frequently accessed data.

Alternatively you could go with the suggestion we made in a recent article "[Tape's Role in Big Data](#)" and combine the two into a single active archive--essentially a single file system that seamlessly marries all of these

media types. This would consist of fast but low capacity (by Big Data standards) primary disk for data ingestion and active analytical processing, optimized disk for more near term data that is not being analyzed at that moment, and tape for long-term storage. In this environment data can be sent to all tiers of storage as it is created or modified so that less or even no backups need to be done.

Big Data is a big storage challenge, not only to store the data but to put it on a fast enough platform that meaningful analytics can be run while at the same time, being cost effective and reliable. These are conflicting terms in the world of storage. Resolving that conflict is going to require a new way of doing things.

Follow Storage Switzerland on [Twitter](#)

George Crump is lead analyst of [Storage Switzerland](#), an IT analyst firm focused on the storage and virtualization segments. Storage Switzerland's [disclosure statement](#).