

Consider apps when deploying CDP

Continuous data protection (CDP) should include data classification, retention policies, and integration with applications.

By Agnes Lamont and Julie Lockner



BUSINESS CONTINUITY IS consistently ranked among the top IT initiatives, according to end-user surveys. As a result, continuous data protection (CDP) as a function of business continuity is a growing market segment. Server consolidation, increased corporate governance, and system and application are also fueling interest in CDP as a complement, or alternative, to traditional data-protection technologies, such as backup and snapshots.

IT departments tasked with implementing a data-protection strategy or supporting business continuity should consider their data-classification policy and their applications to ensure effective deployments. By carefully classifying data, IT departments can make better choices about where and how to use CDP to meet their short- and long-term requirements for data protection and availability.

App-based data classification

Data classification, as defined by the Storage Networking Industry Association (SNIA), is “an organization of data into groups for management purposes. A purpose of a classification scheme is to associate service-level objectives with groups of data based on their value to the business.” According to the SNIA Data Management Forum (DMF), this requires understanding the source of the data (which business process it belongs to), the policy that determines when the business value of the data changes, and how to enforce the policy based on the type of data (e-mail, files, or database data).

For example, General Ledger data belongs to the financial general accounting business process and is managed by an integrated database and financial accounting application. Generally, this data is subject to monthly and quarterly booking periods, after which the books are closed and financial statements are generated. In

open booking periods, every transaction needs to be protected and the data readily available. Once periods are closed, the data is considered frozen and read-only. Data from closed periods is no longer changed but retains importance from a reporting and compliance perspective. Typically, financial data from the current and past year are most valuable for reporting and decision-making comparison. However, the data generally needs to be retained and available for exception reporting, audit, and compliance for seven years or more. Once the retention period passes, the data no longer needs to be protected and available and may be purged. General Ledger data typically consists of files and database data that are driven by a financial application that may be federated across several systems.

Simple classification of General Ledger data from financial database application

Booking period status	Time	Reporting/data availability requirements
Open	Current	High
Closed	< 2 years	Medium
Closed	2-7 years	Low
Closed	> 7 years	None

In the case of our General Ledger example, the value of capturing every transaction change as it occurs and maintaining high availability of the data is obvious. With CDP, every invoice, payment, etc., is captured as it occurs. In the event of a problem—even if it’s an accidental deletion—the lost or corrupted data can be recovered to the exact point in time before the mishap. Having the kind of complete change history that only true CDP can provide can ensure rapid recovery, high data availability, and timely reporting. In the case of financial data, this benefits the organization from a decision-making as well as a corporate governance perspective.

After the booking period is closed and the financial statements are prepared, the data doesn’t change. So the capture mechanism of the CDP solution is not triggered and no additional data is generated to the protected data repository.

However, by having CDP in place, it is possible to detect any attempts to change the historical data after the close point. From an audit, compliance, and security perspective it is extremely valuable to be able to ensure and validate the ongoing integrity of the data that is under CDP.

CDP and retention

Once data is properly classified, it is easier to establish and enforce appropriate policies for its protection and retention. By classifying and protecting data in logical content groups, rather than by the storage device on which the bits reside, retention management and recovery is easier to manage. By maintaining the database and related log files for a specific application in a single, manageable content group, protection and retention can be managed throughout the lifecycle of the data so that recovery is simplified.

As mentioned earlier, true CDP captures all changes to data as they occur. In a rapidly changing environment, it is natural to expect that the storage capacity requirements will be large. However, that may not be the case. Many CDP systems capture only the data changes. So, a 100GB database that experiences change rates of 200MB per day will only capture, transmit, and store the changed 200MB in its repository each day—not the entire 100GB. This contrasts with backup, where typically full backups are performed weekly in any case, regardless of the rate of change and the fact that incremental or differential backups are performed nightly. However, it should be noted that, in the case of smaller files, it may be appropriate to recapture an entire file because the overhead is minimal.

It is a mistake to think that just because a CDP system has captured data that the IT organization must then retain it forever. Some CDP systems enable intelligent data pruning based on desired retention policies, by content group. So, while it may be desirable to have the ability to recover any data changes within, say, 72 hours, after that it may be desirable to retain hourly data views for seven days, daily views for a month, and then monthly views for seven years. In the case of an order entry system where orders are generally dispatched within 48 hours, it is clearly beneficial to be able to view all transactions for a few days. This can help in troubleshooting, such as tracking what operator changed a particular order and recovering accidentally deleted fields if need be. However, as orders ship and data is verified and aged, keeping everything is unnecessary and may lead to unnecessarily large CDP repositories. Rather than incurring protection gaps by implementing periodic snapshots, IT organizations can implement a CDP solution that enables them to release data views when they are no longer necessary, thereby pruning their repositories and keeping their repository capacity within acceptable limits.

Because almost all CDP systems today are interoperable with tape backup, it is possible for IT organizations to continue to leverage their tape systems even as they transition traditional data protection to CDP over time. Before deleting data from a CDP repository, it is possible to make copies to tape, perhaps for off-site archival purposes. However, there is increasing interest in maintaining CDP-captured data on disk, often with the use of de-duplication so that even greater storage efficiencies are maintained beyond pruning by the CDP application.

Increasingly, IT organizations are being held to more-stringent RTOs. The inability to access data is costly from financial, reputation, regulatory, and corporate governance perspectives. Getting to the raw data is insufficient. It is necessary to get to the right data in a way that can be understood and used by the application in the context of its business process. So recovery speed relies on the logical coherence, or application consistency, of the data.

In electronic data management, a set of data is said to be “consistent” when the data can be correctly and unambiguously interpreted by an application. In the case of backups and snapshots, applications and databases generally need to be quiesced or closed to flush queues, buffers, and caches and create a logically consistent point to protect, which will lend itself to later recovery. This is also true of some, but not all, CDP solutions.

For an IT organization planning its data-protection deployment strategy, whether using traditional methods or by implementing CDP, spending time understanding and testing the level of application-consistency provided by various alternatives are important. Block-based CDP, for instance, works at the device rather than the logical level of the application stack. File- and application-based CDP work at the logical level and are therefore architecturally closer to the application semantics. Different CDP solutions may use multiple interfaces to applications to ensure varying levels of application-awareness and consistency. In some cases, the application or database may need to be momentarily quiesced to capture and mark an application-consistent view. In other systems, all application-consistent views are recognized transparently to the application.

CDP is increasingly being deployed as part of IT organizations’ data-protection and business continuity strategies. The closer attention IT departments pay to the connection between their data and the associated applications, the more successful and efficient they will be. Classifying data

by application for logical grouping and life-cycle management will make it easier to manage data protection and retention and to ensure effective, efficient recovery.

For more information on CDP, see SNIA’s CDP Special Interest Group (SIG) at www.snia-dmf.org/cdp. □

Agnes Lamont is co-chair of the Storage Networking Industry Association’s Data Management Forum and vice president of marketing for TimeSpring Software. **Julie Lockner** is a member of the Data Management Forum Board of Directors, and vice president of sales for Solix Technologies.



[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

