

Wstęp

Tomasz Lewicki

WWSIS, Wrocław

kwiecień 2007

Terminologia

- informacja i dane
- archiwizacja
- *backup*
- kopia zapasowa
- replikacja
- nośnik i urządzenie przechowywania informacji
- kompresja
- digitalizacja

Dane to zbiory informacji zebrane w plikach w najróżniejszej formie, np. tekstowej, liczbowej, graficznej. Dobrym przykładem są bazy danych, np. systemy biblioteczne lub bankowe.

Rodzaje dokumentów i danych

Nieliczne przykłady

- biznesowe: umowy, faktury, zamówienia, wyciągi bankowe, sprawozdania finansowe
- publiczne i urzędowe: akty prawne, księgi wieczyste i katastralne, protokoły, notatki służbowe, wnioski, patenty
- prywatne: listy, zdjęcia, muzyka, filmy
- inne: dokumentacja medyczna, archiwa prasowe, mapy geodezyjne, dokumentacja techniczna

Powody ochrony danych

Dane i dokumenty są chronione ze względu na różne, mniej lub bardziej przewidywalne zdarzenia, które mogą doprowadzić do ich nieodwracalnej utraty bądź znacznie utrudnić ich odtworzenie. Przyczyny utraty danych mogą być rozmaite:

- przypadkowe, ale możliwe do przewidzenia: pożar, powódź, zalanie, huragan, zanik zasilania, uszkodzenie sprzętu, kradzież z włamaniem, działania wojenne, zamachy terrorystyczne
- przewidywalne, ale trudniej im zapobiec: błędy w oprogramowaniu, błędy operatorów (użytkowników i administratorów), celowe działanie (np. usunięcie danych), złośliwe oprogramowanie

Według raportu organizacji *IT Policy Compliance* z lutego 2007 r. błąd człowieka jest najczęstszą przyczyną utraty informacji (75%, wliczając w to naruszenia polityki bezpieczeństwa). Złośliwe oprogramowanie jest odpowiedzialne za około 20% przypadków utraty danych.

Skutki utraty danych

Utrata danych wiąże się najczęściej z licznymi problemami, o ile nie dysponuje się dobrze wykonaną kopią utraconej informacji. „Kaliber” straty jest różny w zależności od tego, kto był właścicielem utraconych danych i jakie informacje zawierały. Skutki utraty danych można podzielić na:

- bezpośrednie: koszty spowodowane koniecznością odtworzenia danych, utrata zysków w czasie przestoju przedsiębiorstwa, koszty związane z opłaceniem kar umownych dla kontrahentów w przypadku niewywiązania się z umów
- pośrednie: utrata zaufania klientów i kontrahentów, utrata zadowolenia klientów, utrata marki i prestiżu, obniżenie pozycji rynkowej firmy, spadek produktywności, spadek wartości przedsiębiorstwa np. wskutek obniżki cen akcji na giełdach

Niektórych danych nie da się odtworzyć (dotyczy to szczególnie osób prywatnych).

Archiwizacja jest procesem wykonywania kopii danych w celu zmniejszenia obciążenia systemu komputerowego danymi zbędnymi lub rzadko używanymi. Archiwizacja jest często mylona z **backupem**.

Archiwizacji poddaje się rzadko przetwarzane lub przestarzałe dane, które mają niewielkie znaczenie dla użytkowników, ale mogą być potencjalnie przydatne. Archiwizacja może mieć postać *analogową* (wydruki) lub *cyfrową* (kopie na nośnikach optycznych, magnetycznych bądź magnetoptycznych). Nośniki optyczne i magnetoptyczne, a także niektóre magnetyczne (taśmy) są stosowane ze względu na niewielki koszt jednostkowy oraz znaczną pojemność. Ich wadą jest długi (w porównaniu np. z dyskami twardymi) czas dostępu do danych.

Archiwizacja

Archiwizacja jest procesem wykonywania kopii danych w celu zmniejszenia obciążenia systemu komputerowego danymi zbędnymi lub rzadko używanymi. Archiwizacja jest często mylona z **backupem**.

Archiwizacji poddaje się rzadko przetwarzane lub przestarzałe dane, które mają niewielkie znaczenie dla użytkowników, ale mogą być potencjalnie przydatne. Archiwizacja może mieć postać *analogową* (wydruki) lub *cyfrową* (kopie na nośnikach optycznych, magnetycznych bądź magnetoptycznych). Nośniki optyczne i magnetoptyczne, a także niektóre magnetyczne (taśmy) są stosowane ze względu na niewielki koszt jednostkowy oraz znaczną pojemność. Ich wadą jest długi (w porównaniu np. z dyskami twardymi) czas dostępu do danych.

Backup

Backup to kopia danych przetwarzanych na bieżąco, czasem także kopia całego systemu operacyjnego i zainstalowanych na nim aplikacji. Kopia taka służy jako zabezpieczenie w przypadku utraty bądź uszkodzenia danych spowodowanego zdarzeniem losowym lub błędem ludzkim. Kopie mogą być wykonywane z serwera lub bezpośrednio ze stacji roboczych, w zależności od rodzaju zabezpieczanych danych.

Dobrze wykonany backup pozwala na szybkie odtworzenie danych bądź całego systemu i przywrócenie funkcjonalności, jakiej żądają użytkownicy. Backup jest tylko jednym ze środków ochrony systemu informatycznego przed utratą lub uszkodzeniem danych.

Backup

Backup to kopia danych przetwarzanych na bieżąco, czasem także kopia całego systemu operacyjnego i zainstalowanych na nim aplikacji. Kopia taka służy jako zabezpieczenie w przypadku utraty bądź uszkodzenia danych spowodowanego zdarzeniem losowym lub błędem ludzkim. Kopie mogą być wykonywane z serwera lub bezpośrednio ze stacji roboczych, w zależności od rodzaju zabezpieczanych danych.

Dobrze wykonany backup pozwala na szybkie odtworzenie danych bądź całego systemu i przywrócenie funkcjonalności, jakiej żądają użytkownicy. Backup jest tylko jednym ze środków ochrony systemu informatycznego przed utratą lub uszkodzeniem danych.

Backup — c.d.

Rodzaje backupu

Wyróżniamy trzy rodzaje backupu:

- całościowy (pełny, *full*) — wykonywana jest kopia wszystkich danych, bit archive pliku otrzymuje wartość 0
- przyrostowy (*incremental*) — kopiowane są jedynie te dane, które uległy zmianie bądź zostały utworzone od momentu wykonania ostatniego backupu pełnego lub przyrostowego, tzn. kopiowane są pliki, których bit archive ma wartość 1; po wykonaniu backupu bit ten otrzymuje wartość 0
- różnicowy (*differential*) — wykonywana jest kopia tylko tych danych, które zmieniły się po przeprowadzeniu ostatniego backupu całościowego lub przyrostowego, tzn. kopiowane są pliki z wartością bitu archive równą 1; po wykonaniu backupu pozostaje ona niezmieniona

Backup — c.d.

Backup całościowy (pełny)

Zalety:

- łatwość odtworzenia danych — najczęściej wszystkie znajdują się na jednym nośniku
- krótki czas odtworzenia danych lub przywrócenia systemu do pełnej funkcjonalności

Wady:

- długi czas tworzenia backupu
- nieoptymalne i nieefektywne wykorzystanie nośników — dane rzadko wykorzystywane są traktowane na równi z danymi często ulegającymi zmianom

Backup — c.d.

Backup przyrostowy

Zalety:

- backup jest tworzony szybciej niż w przypadku backupu pełnego
- nośniki są wykorzystywane efektywniej

Wady:

- do odtworzenia danych potrzebny jest komplet nośników z backupami przyrostowymi (w porządku chronologicznym) oraz ostatni backup pełny

Backup — c.d.

Backup różnicowy

Zalety:

- odtworzenie danych jest łatwiejsze niż w przypadku backupu przyrostowego — trzeba mieć do dyspozycji ostatnią pełną kopię oraz ostatnią kopię różnicową
- czas odtworzenia danych jest krótszy niż w przypadku korzystania z kopii przyrostowych
- wykonanie backupu zajmuje mniej czasu niż w przypadku backupu pełnego

Wady:

- nieefektywne wykorzystanie nośników — dane nie ulegające zmianom są backupowane kolejny raz
- czas wykonania backupu jest dłuższy niż w przypadku backupu przyrostowego

Kopia zapasowa (bezpieczeństwa)

Kopia zapasowa to kolejna kopia nośnika z backupem lub danymi zarchiwizowanymi, realizowana jako dodatkowe zabezpieczenie szczególnie ważnych danych. Jeśli nośniki są niepewnej jakości (tanie płyty CD lub DVD) bądź są mocno eksploatowane (taśmy magnetyczne), istnieje ryzyko, że odtworzenie danych zapisanych na takim nośniku będzie trudne lub wręcz niemożliwe.

Macierze RAID (*Redundant Array of Independent Disks*) nie są w założeniu jedynym sposobem zabezpieczenia danych przed uszkodzeniem lub utratą, ale ich stosowanie znacząco podnosi poziom bezpieczeństwa lokalnej kopii danych zawartej na zwielokrotnionych twardych dyskach dostępnych dla systemu komputerowego.

Replikacja

Replikacja zdalna

Ogólnie mianem **replikacji** określamy proces powielania danych, czyli ich zabezpieczania już na etapie tworzenia bądź modyfikacji danych. Wyróżnia się kilka rodzajów replikacji zdalnej:

- synchroniczna — zabezpieczanie danych w czasie rzeczywistym poprzez ich kopiowanie do innej lokalizacji fizycznej (np. zapasowego centrum danych)
- asynchroniczna — j.w., ale kopiowanie odbywa się z niewielkim opóźnieniem, zazwyczaj kilkunastominutowym
- okresowa — od replikacji asynchronicznej różni się zdefiniowaniem pór kopiowania danych

Replikacja — c.d.

Replikacja lokalna

Replikacja lokalna pozwala na skrócenie czasu przygotowania danych do zabezpieczenia, zdejmując z głównego systemu ciężar przygotowania i wykonania replikacji zdalnej. W ramach replikacji lokalnej wyróżniamy:

- migawkową kopię danych (*snapshot*) — replikacji poddawane są jedynie wskaźniki danych, dzięki czemu powstaje wirtualna kopia danych w stanie takim, w jakim dane były w momencie rozpoczęcia replikacji. Taka kopia ma niewielkie rozmiary
- migawkowy duplikat danych (*clone*) — w odróżnieniu od *snapshot*-u, wykonywana jest pełna kopia danych

Nośniki i urządzenia przechowywania danych

W informatyce **nośnikiem** nazywamy przedmiot, na którym można trwale przechowywać informacje i odtwarzać je wielokrotnie. **Urządzeniem przechowywania danych** można określić urządzenie elektroniczne zdolne zapisać informacje na nośnik i odczytać te informacje w późniejszym czasie.

Mianem **pamięci masowej** określa się te rodzaje pamięci, które umożliwiają trwałe przechowanie informacji. Do pamięci masowych zaliczają się nośniki danych.

W niektórych przypadkach słowo „urządzenie” jest używane zamiennie ze słowem „nośnik” (np. pamięci typu *flash*), a „nośnik” bywa zamieniany na „pamięć masową”.

Rodzaje nośników i pamięci masowych oraz odpowiadających im urządzeń do zapisu i odczytu

- karty i taśmy dziurkowane → czytniki i dziurkarki kart oraz taśm
- taśmy magnetyczne → *streamery*, magnetofony
- bębny magnetyczne → pamięci bębnowe
- dyskietki → stacje dyskietek
- dyski (talerze) magnetyczne → dyski twarde
- dyski optyczne → czytniki i nagrywarki CD, DVD, Blu-Ray
- dyski holograficzne → czytniki i nagrywarki dysków holograficznych
- dyski magnetoptyczne → czytniki i nagrywarki dysków magnetoptycznych
- pamięci typu *flash* (błyskowe): karty pamięci, *pendrives*

Mówiąc o **kompresji** w znaczeniu informatycznym mamy zazwyczaj na myśli zmniejszenie objętości danych przy zachowaniu „ładunku informacyjnego”, czyli sensu tych danych. Kompresja ma zazwyczaj na celu oszczędność nośnika i/lub łącza sieciowego, którym są przesyłane dane.

Algorytmem kompresji nazywamy schemat postępowania przy zmniejszaniu objętości pliku, najczęściej zależny od charakteru kompresowanych danych. Proces odwrotny do kompresji nazywamy **dekompresją**.

Algorytmy kompresji

Algorytmy kompresji można podzielić ze względu na różne kryteria, np. stopień zmiany „ładunku informacyjnego” lub rodzaj danych poddawanych kompresji:

- kompresja bezstratna — w procesie dekompresji uzyskuje się dane w postaci identycznej z tą, jaką miały przed poddaniem ich kompresji (przykład: tekst, baza danych)
- kompresja stratna — w procesie kompresji „gubione” są pewne informacje, jednak ich znaczenie nie jest kluczowe dla całości danych, tzn. po dekompresji sens danych zostaje zachowany (przykład: zdjęcie, dźwięk)

Przykłady algorytmów bezstratnych: Deflate, Huffman, LZW, RLE, BZIP2.
Przykłady algorytmów stratnych: DCT, metoda falkowa, JPEG, MPEG, Vorbis, MP3.

Digitalizacja

Najogólniejsza definicja **digitalizacji** to przekształcanie informacji w postaci analogowej na odpowiadającą jej postać cyfrową. W węższym znaczeniu pod pojęciem digitalizacji będziemy rozumieć zamianę dokumentu papierowego w jedną z form cyfrowych (np. poprzez skanowanie).

Digitalizacja ma zazwyczaj na celu udostępnienie danego dokumentu większej grupie odbiorców dzięki rozpowszechnianiu ich szybszymi niż tradycyjne kanałami informacyjnymi (np. za pośrednictwem Internetu). Drugim celem digitalizacji jest zwiększenie trwałości („czasu życia”) dokumentów papierowych, czyli ich archiwizację w postaci cyfrowej.

Technologia i formaty zapisu danych

Do różnych zastosowań zaprojektowano różne, czasami pokrywające się, rozwiązania technologiczne i formaty zapisu informacji. Niektóre z nich w zmodyfikowanej formie przetrwały przez całe dekady (np. taśma magnetyczna), inne były zjawiskiem chwilowym (np. napędy Zip). Podobna sytuacja ma miejsce wśród formatów zapisu danych.

Przy projektowaniu kompleksowego systemu ochrony danych w przedsiębiorstwie, urzędzie, a nawet w zastosowaniach prywatnych należy wziąć pod uwagę dostępność i rozpowszechnienie danej technologii lub formatu, wsteczną zgodność (tzn. możliwość odczytania starszych technologicznie nośników lub formatów przez nowy sprzęt lub oprogramowanie), wsparcie ze strony producenta (deklarowane i rzeczywiste), bezpośrednie i pośrednie koszty (zakup urządzeń, materiałów eksploatacyjnych i nośników) oraz opinie innych użytkowników.