Archiwizacja i kopie zapasowe

Witold Paluszyński
Katedra Cybernetyki i Robotyki
Politechnika Wrocławska
http://www.kcir.pwr.edu.pl/~witold/

Zagrożenia bezpieczeństwa systemu komputerowego

- Bezpieczeństwo fizyczne systemu: kradzież, zniszczenie, uszkodzenie, np. pożar, zalanie, skok napięcia.
- Bezpieczeństwo danych: utrata, uszkodzenie, sabotaż, kradzież, ujawnienie.
- Ciągłość dostępu do systemu: awarie sprzętu, oprogramowania, awarie zasilania, wentylacji, klimatyzacji, ogrzewania, ciągłości dostaw (np. prądu, paliwa do generatorów, albo tonera do drukarki), ataki typu DOS.

Większości z tych zagrożeń można zapobiegać, jednak nigdy ze 100%-wym skutkiem. Dlatego wykonuje się **kopie zapasowe** (*backup*).

Terminologia

W informatyce istnieje nieścisłość terminologiczna zwiazana z procesami tworzenia kopii zapasowych. Często te procesy nazywa się archiwizacją.

Ściśle rzecz biorąc, **archiwizacja** oznacza przenoszenie danego obiektu do archiwum, czyli wykonanie kopii archiwalnej obiektu, która ma go zastąpić. W języku polskim mieszanie tych pojęć jest dodatkowo uzasadnione, ponieważ nie istnieje zgrabny czasownik oznaczający tworzenie kopii zapasowych (*backup*-owanie??).

Dlatego w tej prezentacji pojęcia archiwizacji i tworzenia kopii zapasowych również będą stosowane zamiennie.

Nie należy natomiast mylić tworzenia kopii zapasowych z tworzeniem kopii dodatkowych, jako elementu **redundancji**, mającej na celu zabezpieczenie przed awarią. Na przykład, macierz RAID-1 zapisuje dane jednocześnie na dwóch dyskach. Po awarii jednego dysku macierz pracuje dalej korzystając z drugiego.

Kopie zapasowe nie chronią przed awarią, tylko pomagają przywrócić stan systemu po będącej skutkiem awarii utracie danych.

Tworzenie kopii zapasowych

- element planowania kryzysowego
- zabezpieczenie przed mikro-awariami (przypadkowe skasowanie pliku/ów) i prawdziwymi katastrofami (awaria dysku, włamanie, pożar)
- również pozwala odzyskać poprzednią wersję niedawno zmienionego pliku
- wykonywana na mediach ładowalnych, tzn. takich, które wyciąga się z komputera w czasie pracy, co zabezpiecza przed takimi awariami, jak awaria kontrolera dysku (wyłączającego nie jeden ale wiele dysków), jak również przed jednoczesną utratą lub uszkodzeniem wszystkich dysków
- wymaga planowania:
  - cykl nagrywania i rotacja nośników
  - sposób przechowywania i zabezpieczenie
  - format danych i oprogramowanie
- wymaga testowania — nie wystarczy zaplanować i uruchomić ambitne procedury tworzenie kopii zapasowych, a potem spać spokojnie; trzeba testować nie tylko możliwość odczytania poszczególnych taśm, ale pełne scenariusze awarii, i możliwość odbudowania i ponownego uruchomienia całego systemu lub centrum obliczeniowego

Zasady archiwizacji

- Archiwizacja danych
  - ośrodka obliczeniowego
    - regularna, globalna, duże ilości danych, dobrze działająca machina
  - indywidualnego użytkownika
    - mniejsze ilości danych, inteligentne podejście, często zaniedbywana
- Archiwizacja systemu
  - ośrodka obliczeniowego
    - przechowywanie nośników instalacyjnych
    - rozwiązywanie problemów w przypadku reinstalacji na innym sprzęcie (sterowniki, licencje, numery seryjne)
  - indywidualnego użytkownika
    - często zastępowana nową instalacją; możliwe problemy, np. sterowniki
- Archiwizacja struktur (całych partycji/dysków)
  - szybka, wygodna metoda archiwizacji systemu + danych
  - możliwa tylko przy odbudowie systemu na tym samym sprzęcie
  - wymaga wstępnej instalacji systemu, lub odtwarzania dysków na innym
  - podobna w przypadku ośrodka IT oraz indywidualnego użytkownika

Planowanie harmonogramu archiwizacji

- dyski, partycje dyskowe krytyczne — muszą być archiwizowane
  - szybkozmienne: np. katalogi użytkowników
  - wolnozmienne: np. systemowe
  - inne, np. pojedyncze zmienione, lub zmieniające się pliki na ogólnie niezmiennej strukturze, np. plik /etc/passwd, lub pliki konfiguracyjne jakiegoś programu, lub podsystemu
- dyski, partycje dyskowe niekrytyczne — niekoniecznie wymagające archiwizacji: /tmp, /var/log, /var/spool

Rodzaje kopii zapasowych

- **Pełna** — kompletna kopia zawierająca wszystkie dane
- **Przyrostowa** — archiwum zawierające wszystkie pliki zmodyfikowane / utworzone od czasu ostatniego ich nagrania na kopii zapasowej.

  Tworzenie kopii przyrostowych znacznie oszczędza koszt i nakład pracy, co w efekcie pozwala zaimplementować skuteczniejszy schemat archiwizacji.

  Na przykład, wykonanie raz w tygodniu kopii pełnej, a przez pozostałe dni kopii przyrostowych daje bardzo dobry poziom bezpieczeństwa, typowo znacznie niższym kosztem, niż codzienne nagrywanie kopii pełnej.

  Jednak wadą tej procedury jest dłuższy czas odtwarzania danych.
- **Różnicowa** — archiwum zawierające wszystkie pliki zmodyfikowane / utworzone od czasu nagrania ostatniej pełnej kopii zapasowej.

  Rozwiązuje problem odtwarzania z kopii pełnej+przyrostowych. Odtworzenie wszystkich danych wymaga użycia tylko kopii pełnej i ostatniej różnicowej.

  W systemach, gdzie zmiany nie są bardzo intensywne, schemat nagrywania pełnej kopii raz w tygodniu, a różnicowych archiwów w dni pozostałe, może zużyć tyle samo taśm co w przypadku metody przyrostowej.

Nośniki

- taśma magnetyczna
  - wady: wolny dostęp, zużywa się, podatna na uszkodzenia mechaniczne, magnetyczne, rozproszone promieniowanie elektromagnetyczne
  - czas zachowania danych ograniczony do ok. 2 lat
  - użytkowanie ograniczone do ok. 50 przebiegów (zapisu lub odczytu)
  - wysoka pojemność
  - historycznie najczęściej używane i nadal powszechnie stosowane
- inne nośniki magnetyczne (dyskietki, dyski wymienne)
- magneto-optyczne i optyczne
- nośniki jednorazowego nagrywania: WORM, CD-R, DVD-R

  Nowoczesne, wygodne i tanie (ale tylko w cenie za urządzenie, a niekoniecznie za megabajt). Zwykle mają mniejsze pojemności i są głównie stosowane do archiwizacji na mniejszą skalę, np. do zarchiwizowania dysku prywatnego komputera.
- dyski twarde ???

Wymagania procedur archiwizacji

- Archiwizacja plików systemowych:
  - wymaga opracowania planu odbudowy systemu z taśm
    - taki plan musi zawierać szczegółowe komendy do wykonania w czasie odbudowy i musi być przechowywany poza systemem;
  - cykl i częstotliwość archiwizacji muszą być dostosowane do częstotliwości zmian w konfiguracji systemu
    - w razie konieczności zachowania logów systemowych (`/var`) częstotliwość ich archiwizacji może być inna niż pozostałych plików systemowych.
- Archiwizacja plików użytkowników:
  - częstotliwość archiwizacji z reguły wyższa niż systemu, może być wymagana np. wiele razy dziennie
    $\rightarrow$ często celowa jest archiwizacja przyrostowa na zmianę z całościową;
  - konieczne jest zapewnienie poufności danych.
- Archiwizacja baz danych:
  - wymaga zastopowania serwera b.d. i/lub wykonania snapshota.

Wymagania procedur archiwizacji (cd.)

Dodatkowe procedury stosowane przy tworzeniu kopii zapasowych:

- kompresja kopii zapasowych
  - zalety:
    - lepsze upakowanie danych na nośnikach, oszczędności
    - napedy taśm mają często wbudowane algorytmy kompresji
  - wady:
    - utrudniony dostęp do poszczególnych plików w archiwum; w przypadku uszkodzenia nośnika konsekwencje mogą być poważniejsze niż utrata danych jednego pliku
    - trudna do oszacowania pojemność nośnika
    - dodatkowe obciążenie CPU, spowolnienie nagrywania
- szyfrowanie kopii zapasowych
  - czasami konieczne dla zabezpieczenia danych na nośnikach
  - obciążenie CPU, spowolnienie

Schemat i harmonogram archiwizacji

- określenie częstotliwości archiwizacji poszczególnych struktur; z niego wynika maksymalny czas zanim zostanie utworzona kopia zapasowa zapisanego pliku
- wybór archiwizacji pełnej i liczby poziomów archiwizacji przyrostowej
- nadanie taśmom etykiet (nalepek na szpulach lub kasetach), i przypisanie konkretnych taśm do konkretnych dysków lub systemów plików i poziomów archiwizacji
- określenie sposobu recyrkulacji taśm, tzn. które taśmy i kiedy można nagrywać ponownie; z niego wynika okres zachowania danych (horyzont kopii zapasowych) określający jak dawna kopia danych jest dostępna

Przykładowy schemat archiwizacji

Wykorzystujemy 10 kompletów taśm, 6 wystarczających do zapisania kompletnego archiwum, i 4 mieszczące dzienne archiwa przyrostowe.

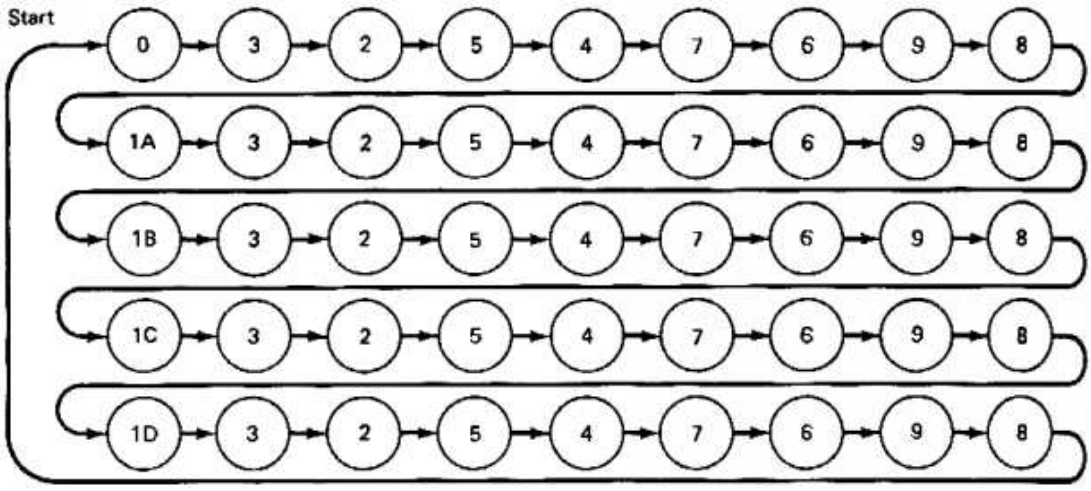
Taśmy etykietujemy: Poniedziałek, Wtorek, Środa, Czwartek (taśmy archiwizacji przyrostowej poziomu 1), i Piątek 1, Piątek 2, Piątek 3, Miesiąc 1, Miesiąc 2, Miesiąc 3 (taśmy archiwizacji pełnej poziomu 0).

Cykl rozpoczynamy w piątek wykonując pełną archiwizację na taśmie (taśmach) Piątek 1. We wszystkie dni powszednie: poniedziałek, wtorek, środę, i czwartek wykonujemy archiwizację przyrostową na odpowiednich taśmach.

W kolejne piątki nagrywamy pełną archiwizację na taśmach: Piatek 2 i Piątek 3, a co czwarty piątek wykorzystujemy taśmy: Miesiąc 1, itd.

Ten schemat zapewnia codzienną (dni powszednie) archiwizację przez 12 tygodni, z pełną archiwizacją w każdy piątek, po czym taśmy zostają ponownie użyte.

Bardziej rozbudowany schemat archiwizacji



Nagrywamy pełne archiwum poziomu 0, po czym wykonujemy archiwa przyrostowe poziomów 2–9 w cyklu podwójnym (każdy plik archiwizowany dwa razy). Po nagraniu ośmiu taśm zaczynamy cykl przyrostowy od nowa nagrywając wszystko od poziomu 0, i cztery razy powtarzamy taki duży cykl używając czterech oddzielnych kompletów taśm poziomu 1, ale jednego kompletu taśm poziomów 2–9. Do archiwizacji poziomu 0 używamy za każdym razem nowych taśm.

Zasady nagrywania nośników

- Nośniki nagrywa się raz, a następnie zabezpiecza przed zapisem.
- Nagrany nośnik musi być od razu precyzyjnie zaetykietowany, najlepiej przed nagraniem.
- Nagrane nośniki muszą być przechowywane w bezpiecznym miejscu.
- Wskazane jest prowadzenie dziennika archiwizacji i ewentualnie spisów zawartości taśm.
- Należy śledzić liczbę błędów zapisu (skorygowanych) aby wymieniać nośniki na nowe na długo zanim zaczną pojawiać się błędy niekorygowalne.
- Archiwizację należy wykonywać na nieczynnych systemach plików (zasada generalnie nieprzestrzegana).
- Sprawdzać, sprawdzać, sprawdzać:
  - sprawdzać czytelność nagranego archiwum,
  - sprawdzać pracę operatora nagrywającego taśmy,
  - sprawdzać poprawność całej procedury archiwizacji i zdolność odtworzenia pojedynczych plików i całego systemu.
- Sensem archiwizacji jest czarny scenariusz; nie ma żadnego znaczenia, że przy sprzyjających warunkach czasami uda się coś odtworzyć.

Narzędzia archiwizacji

Istnieje wiele narzędzi archiwizacji, zarówno pojedynczych programów, jak i złożonych systemów do pracy w instalacji sieciowej na różnych platformach i systemach operacyjnych. Istnieją narzędzia komercyjne i wolnego oprogramowania (*open source*, FOSS).

W świecie IT, gdzie typowo ośrodek obliczeniowy wykonuje swoją własną archiwizację (zamiast zlecić ją zewnętrznej firmie), narzędzia archiwizacji oparte na formatach uniksowych (tar, cpio), są bardzo popularne. Należą do nich rozbudowane systemy sieciowe, służące do archiwizowania wielu systemów, często z pojedynczym serwerem archiwizacji.

W świecie użytkownika indywidualnego ważna jest łatwość użytkowania (GUI), których ani narzędzia tradycyjne, ani rozbudowane systemy sieciowe nie posiadają. Powstają więc nowe narzędzia, i jest ich wiele.

Innym rodzajem użytkownika są użytkownicy komercyjni, którzy posiadają rozbudowane instalacje komputerowe i bazy danych, ale nie są firmami informatycznymi ani nie posiadają własnego zaplecza informatycznego. Tacy użytkownicy mogą zlecać archiwizację swoich danych firmom usługowym.

Narzędzia archiwizacji — lista

- archiwizacja plików: nagrywanie plików w postaci archiwum z katalogami, plikami, nagłówkami, atrybutami, itp.
  - tradycyjne — prosta funkcjonalność, nadal chętnie stosowane
    - tar, cpio, dump/restore — proszą o zmianę nośnika po jego zapełnieniu
  - sieciowe, open-source — rozwiązanie dla rozbudowanych instalacji
    - amanda, bacula, duplicity
  - inne — łączą funkcjonalność z łatwością użycia, typowo GUI
    - np. fwbackups, wiele innych
- archiwizacja struktur: nagrywanie całych systemów plików lub dysków
  - dd (robi kopię dysku 1:1), ufsdump (kopiuje tylko używane bloki)
- archiwizacja przez kopiowanie na zdalny system
  - cp, scp, rsync

Rsync

rsync jest programem do kopiowania plików przez sieć. Teoretycznie jest rozszerzeniem programu rcp, niezalecanego ze względu na swoje wady. Jednak rsync posiada zdolność współpracy z programem scp, a jednocześnie tak dużo dodatkowych przydatnych możliwości, że w rzeczywistości jest systemem samym w sobie.

Poza opcjami wygody rsync posiada własny protokół komunikacyjny, i jest możliwe uruchomienie jego własnego serwera (zwykle ten sam program rsync z opcją --daemon). Protokół ten posiada wiele dodatkowych możliwości, jak np. funkcja kopiowania przyrostowego, i z tego względu można traktować rsync jako narzędzie do tworzenia i utrzymywania kopii zapasowych.

Protokół rsync nie stosuje szyfrowania, ale jest możliwe użycie rsync z wykorzystaniem programu ssh jako warstwy transportowej, która zapewnia szyfrowanie.

Serwer rsync standardowo używa portu 873.

rsync — zwykłe kopiowanie plików

Przykłady z Unix_Toolbox:

Copy the directories with full content:

```
# rsync -a /home/colin/ /backup/colin/                # "archive" mode. e.g keep the sa
# rsync -a /var/ /var_bak/
# rsync -aR --delete-during /home/user/ /backup/      # use relative (see below)
```

Same as before but over the network and with compression. Rsync uses SSH for the transport per default and will use the ssh key if they are set. Use „:" as with SCP. A typical remote copy:

```
# rsync -axSRzv /home/user/ user@server:/backup/user/ # Copy to remote
# rsync -a 'user@server:My\ Documents' My\ Documents  # Quote AND escape spaces for the
```

Exclude any directory tmp within /home/user/ and keep the relative folders hierarchy, that is the remote directory will have the structure /backup/home/user/. This is typically used for backups.

```
# rsync -azR --exclude=tmp/ /home/user/ user@server:/backup/
```

Use port 20022 for the ssh connection:

```
# rsync -az -e 'ssh -p 20022' /home/colin/ user@server:/backup/colin/
```

rsync — użycie protokołu rsync

Przykładowy plik konfiguracyjny serwera `rsync` (`/etc/rsyncd.conf`):

```
uid = nobody
gid = nobody
use chroot = no
max connections = 4
syslog facility = local5
pid file = /var/run/rsyncd.pid

[uucppublic]
        comment = Traditional anonymous dumping place
        path = /usr/spool/uucppublic
        read only = false

[weatherdata]
        comment = Weather station data files
        path = /var/tmp/weather
        read only = true
```


Using the rsync daemon (used with „::") is much faster, but not encrypted over ssh. The location of /backup is defined by the configuration in /etc/rsyncd.conf. The variable `RSYNC_PASSWORD` can be set to avoid the need to enter the password manually.

```
# rsync -axSRz /home/ ruser@hostname::rmodule/backup/
# rsync -axSRz ruser@hostname::rmodule/backup/ /home/    # To copy back
```

Archiwizacja w skali mikro

Podstawowe założenia archiwizacji w środowisku pojedynczego komputera, sieci domowej, lub małej firmy (SOHO), są inne niż w przypadku ośrodka IT. W skali mikro nie ma tak wyśrubowanych wymagań, i w konsekwencji chciałoby się, żeby nakłady czasowe i finansowe były małe.

Jednak praktycznie wymagania w tych dwóch światach (mikro i makro) są podobne — użytkownik chciałby nie stracić swoich cennych danych, a w przypadku awarii być w stanie odzyskać działający system i dane w miarę łatwo.

Co z tego wynika?

Wydaje się, że pewne zasady i procedury archiwizacji w skali makro można i warto przenieść do świata mikro. Na przykład:

- przemyślana strategia
- systematyczność, osiągana przez stosowanie automatycznych procedur (w tym procedur manualnych, nie tylko zadań crona)
- przechowywanie nośników w bezpiecznym miejscu (off-site)
- okresowe testowanie

Tworzenie kopii zapasowych — praktyka

Dlaczego nowoczesny serwer z macierzą RAID nie stanowi dobrego medium do archiwizacji?

Hard Lessons in the Importance of Backups: JournalSpace Wiped Out
Jason Fitzpatrick, lifehacker.com, artykuł z 2009
`http://lifehacker.com/5122848/hard-lessons-in-the-importance-of-backups-journalspace-wiped-out`

Porównanie różnych programów kopiujących w zastosowaniu do archiwizacji

Torture-testing Backup and Archive Programs: Things You Ought to Know But Probably Would Rather Not
Elizabeth D. Zwicky, SRI International, artykul z 1991
`http://www.coredumps.de/doc/dump/zwicky/testdump.doc.html`

Podstawowa terminologia i procedury tworzenia kopii zapasowych:
`http://en.wikipedia.org/wiki/Backup`

Wykorzystanie prostych technik i narzędzi do backup-owania:
`http://www.softpanorama.org/Admin/backup.shtml`