Rekomendacje Dell Technologies w zakresie konfiguracji systemów do składowana dokumentów elektronicznych (macierze plikowe).

Wstęp:

Środowiska dedykowane do składowania danych plikowych (czyli leżących poza strukturą baz danych) charakteryzuje kilka popularnych parametrów:

- Wydajność – mierzona jako IOPS i czas odpowiedzi lub przepustowość strumienia danych;
- Wymagana pojemność użyteczna i efektywność przechowywania danych (pojemność dysków vs pojemność użyteczna na dane);
- Możliwości i koszt rozbudowy wydajności i pojemności;
- Zakres i czas trwania wsparcia technicznego (SLA).

W niniejszym dokumencie skupimy się wyłącznie na kwestiach związanych z zapewnieniem efektywnego udostępniania danych plikowych za pomocą dedykowanych macierzy Scale-Out.

Typy rozwiązań macierzowych:

Systemy składowania danych są podzielone w zależności od sposobu korzystania z danych:

- Macierze blokowe:
  dla obsługi danych systemowych serwerów lub baz danych (dostęp na poziomie bloków danych, tj. zarządzanie danymi na poziomie serwera lub bazy danych). W takich systemach dostęp odbywa się za pomocą protokołu iSCSI lub sieci FC;
- Macierze plikowe:
  dla obsługi plików wykorzystywanych przez użytkowników lub aplikacje (czyli Macierz plikowa jest niezależnym serwerem plików). Dostęp do danych za pomocą protokołów: NFS, SMB/CIFS, FTP, HDFS i sieci Ethernet;
- Macierze unified:
  pozwalające na uruchomienie dwóch typów usług – blokowych i plikowych na jednym systemie;
- Macierze obiektowe:
  dla obsługi obiektów (tj. plików + dodatkowych informacji opisujących je – tzw. metadanych), głownie za pomocą protokołu S3, interfejsu REST-API i sieci Ethernet.

Korzyści z zastosowania rozwiązań dedykowanych (macierze plikowe Scale-OUT):

Dane plikowe są wykorzystywane do różnych zastosowań (dane multimedialne, logi z aplikacji, archiwizacja, backup długoterminowy, dane użytkowników – tzw. katalogi domowe lub udziały sieciowe, file share, itp.). Statystyki pokazują, że ok. 80% powierzchni danych globalnie jest zajęta przez dane płaskie – leżące poza bazami danych. Ich przyrost szacuje się na 50-100% rocznie, ze względu na ciągle rosnące wymagania biznesowe (multimedia w większej rozdzielczości, obieg dokumentów elektronicznych, logowanie danych z systemów IoT pracujących 24x7, większe pojemności plików, dłuższy czas ich archiwizacji itp.). Ilości przetwarzanych danych plikowych w ramach jednej organizacji są liczone w dziesiątkach lub setkach Terabajtów, a w największych z nich również w Petabajtach. W związku z powyższym kluczowe dla systemów plikowych są poniższe funkcjonalności:

- liniowa skalowalność wydajności, pojemności i kosztów rozbudowy;
- wielostopniowe zabezpieczenie architektury wewnętrznej i możliwość replikacji do drugiej lokalizacji;
- prostota zarządzania niezależnie od pojemności macierzy;
- brak konieczności migracji danych przy zmianie/odświeżeniu platformy sprzętowej macierzy;
- możliwość uruchamiania wielu wirtualnych macierzy na jednej platformie.

Powyższe funkcjonalności są realizowane wyłącznie przez macierze typu Scale-Out, czyli zbudowane z węzłów kontrolerowo-dyskowych (architektura modułowa), pozwalających na elastyczną rozbudowę o kolejne moduły (węzły) dodające pojemność i wydajność do obecnego systemu. Taka architektura pozwala na planowanie inwestycji i dopasowanie macierzy do zmieniających się potrzeb, dzięki możliwości stosowania modułów o różnej wydajności i pojemności.

Podstawowe funkcje wiodących macierzy plikowych

Współczesne systemy macierzy plikowych scale-out (tzw. Scale-OUT NAS) posiadają kluczowe funkcjonalności niezbędne do efektywnego wykorzystania ich w dowolnej organizacji:

- jednoczesny dostęp do tych samych danych wieloma protokołami: czyli dowolnymi protokołami plikowymi (NFS v3, v4, SMB/CIFS v2, v3, FTP), dostępem w ramach architektury analitycznej Hadoop – protokół HDFS, dostęp za pomocą kontenerów aplikacyjnych (architektury Kubernets, Open Shift) – protokół CSI 1.0 oraz opcjonalnie możliwość dostępu po standardzie obiektowym – S3;
- zarządzanie ruchem użytkowników (load balancing): architektura wielokontrolerowa (wiele węzłów kontrolerowo-dyskowych) pozwala na rozłożenie ruchu użytkowników na wiele węzłów i maksymalne wykorzystanie możliwości macierzy. Współczesne systemy mają wbudowane mechanizmy zarządzania rozłożeniem ruchu na poszczególne węzły w zależności od potrzeb aplikacji i użytkowników;
- kopie migawkowe: są to dodatkowe kopie danych z wybranego momentu w czasie, pozwalające na ich odtworzenie w przypadku utraty źródłowego materiału lub chęci

powrócenia do poprzedniej wersji dokumentu. Współczesne systemy pozwalając na integrację z usługami domenowymi, np. Microsoft VSS;

- limity powierzchni dla użytkowników (tzw Quoty): skuteczne zarządzanie macierzą wymaga możliwości elastycznego zakładania limitów na wykorzystanie powierzchni macierzy dla użytkowników lub per folder/udział macierzy;

- tiering danych (wewnętrzny i do zewnętrznych zasobów – chmury publicznej): sposób wykorzystania danych zależy od momentu cyklu życia, w którym się znajdują, w związku z tym niezbędne jest posiadanie mechanizmów, które automatycznie przesuną mniej używane dane na tańszy (wolniejszy) nośnik, zwalniając miejsce dla danych wymagających szybkiego dostępu. Temu służy mechanizm tzw. tieringu – realizowany wewnątrz macierzy lub przy wykorzystaniu zewnętrznych komponentów (np. tiering do usługi z chmury publicznej za pomocą standardu S3);

- niezaprzeczalność danych (WORM SEC17a-4): 100% danych powstaje w wersji elektronicznej i większość materiału źródłowego jest składowana na dyskach, w związku z tym krytyczne staje się zapewnienie niezaprzeczalności danych w zadanym horyzoncie czasu. Funkcjonalność WORM (Write Once Read Many) pozwala na zablokowanie wybranych danych do edycji w wybranym okresie czasu. Dzięki temu mamy pewność, że dostajemy się do oryginalnych danych i zapewniamy dodatkową odporność na ataki typu Ransomware (czyli szyfrowanie danych). Standard SEC17a-4 leży u podstaw międzynarodowych i lokalnych regulacji prawnych w tym zakresie i jest gwarantem niezaprzeczalności danych składowanych elektronicznie;

- replikacja: druga kopia danych jest niezbędnym elementem każdego systemu macierzowego. Systemy plikowe muszą umożliwiać szybką replikację asynchroniczną, obsługującą przyrosty danych (w celu ograniczenia zajętości łączy) oraz szyfrowanie (zapewnienie bezpieczeństwa). Replikacja powinna być możliwa w obu kierunkach dla wybranych folderów/udziałów, na bazie polityk wybranych przez administratora. Replikacja może być realizowana do drugiego ośrodka (zapasowe DC), lub do systemu wirtualnego – realizowanego jako usługa w ramach chmury publicznej. Alternatywnie, można stosować mniej efektywny sposób zabezpieczenia danych plikowych czyli backup na nośnik zewnętrzny – przy wykorzystaniu standardu NDMP (obsługiwanego przez wiodące systemy backupowe na rynku);

- multi-tenancy: macierz typu Scale-OUT doskonale realizują konsolidację danych, tj. pozwalają na wykorzystanie ich jednocześnie do różnych zastosowań/aplikacji. Jest to możliwe dzięki jednoczesnej skalowalności pojemności i wydajności oraz opcji multi-tenancy, czyli możliwości stworzenia wielu wirtualnych macierzy w ramach jednej platformy sprzętowej. Multi-tenancy zapewnia pełne odseparowanie danych, możliwość odrębnej adresacji/metod kontroli dostępu dla każdej z wirtualnych macierzy oraz pełną separację administratorów poszczególnych udziałów;

- RBAC (Role Based Access Control): efektywne zarządzanie IT wymaga elastycznego podejścia do uprawnień administratorów macierzy plikowych i elastycznego zarządzania nimi – w zależności od potrzeb;

- opcja redukcji danych: dla wybranych danych warto stosować algorytmy kompresji (czyli redukcji zajętości na poziomie plików) lub deduplikacji (czyli redukcji zajętości na poziomie bloków danych na dyskach);

- szybka indeksacja i przeszukiwanie danych: zarządzanie danymi i możliwość ich indeksacji staje się kluczowa wraz ze wzrostem pojemności tych systemów – szczególnie w skali powyżej 100TB. Współczesne systemy pozwalają na szybkie zindeksowanie danych, wyszukiwanie wybranych plików (lub np. ich duplikatów), nadawanie własnych atrybutów/opisów danych oraz możliwości przenoszenia wybranych danych do innych systemów.

Zarządzanie i monitoring:

Poniżej przedstawiamy 3 podstawowe funkcje dotyczące monitorowania stanu macierzy plikowych i danych, które się na nich znajdują:

- auto-monitoring producenta (usługa call-home): współczesne macierze automatycznie rejestrują awarię lub anomalie i automatycznie rejestrują zgłoszenie w systemie serwisowym producenta, rozpoczynając proces naprawczy;

- monitoring graficzny podstawowych funkcji systemu: wszystkie witalne parametry systemu powinny być dostępne za pomocą przejrzystego interfejsu graficznego dostępnego z poziomu stacji roboczej lub smartfona;

- możliwość rozliczania użytkowników systemu za jego wykorzystanie (charge-back): w przypadku konsolidacji dużej ilości danych (setki TB) kluczowa staje się możliwość rozliczania użytkowników z wykorzystanej powierzchni (np. w ramach poszczególnych działów firmy, korzystających z jednego zasobu plikowego).