Rekomendacje Dell Technologies w zakresie konfiguracji systemów do archiwizacji dokumentów elektronicznych (macierze obiektowe).

Wstęp:

Środowiska dedykowane do składowania danych obiektowych (czyli dokumentów leżących poza strukturą baz danych wraz z dodatkowymi informacjami opisowymi – tzw. metadanymi) są charakteryzowane przez 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) szczególnie w rozproszeniu geograficznym na kilka centrów danych;
- 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 obiektowych 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, z możliwością rozproszenia geograficznego macierzy (tj. instalacji w więcej niż dwóch ośrodkach danych).

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

Dane obiektowe są wykorzystywane do różnych zastosowań (archiwizacja dokumentów, składowanie logów z aplikacji i rozwiązań IoT, backup długoterminowy, itp.). Składowanie danych w formie obiektów pozwala na uzyskanie następujących korzyści:

- możliwość opisywania składowanych dokumentów własnymi atrybutami (tzw. metadanymi) zgodnie z systematyką Klienta;
- składowanie w jednym miejscu (na poziomie macierzy) danych oraz informacji je opisujących (metadanych), czyli możliwość wykorzystania dodatkowych narzędzi na poziomie aplikacji do katalogowania danych i ich uporządkowania;
- brak konieczności podawania ścieżki dostępu do plików na poziomie aplikacji – macierz obiektowa wymaga jedynie podania unikalnego identyfikatora (nazwy obiektu) i automatycznie wyszuka go w swojej bazie wewnętrznej niezależnie od jego aktualnego fizycznego położenia (np. w różnych centrach danych), jest to realizowane w ramach standardu dostępu do danych – S3;
- możliwość uruchomienia usługi „chmurowej" – tj. jednego spójnego, rozproszonego geograficznie systemu serwującego dane dla użytkowników i aplikacji;
- automatyczne wersjonowania danych – każda zmiana w dokumencie, powoduje zapis na macierzy nowego obiektu (nowego pliku i odpowiednich metadanych).

Statystyki pokazują, że ok. 80% powierzchni danych globalnie jest zajęta przez dane płaskie (plikowe lub obiektowe) tj. 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 płaskich 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 obiektowych są poniższe funkcjonalności:

- liniowa skalowalność wydajności, pojemności i kosztów rozbudowy;
- wielostopniowe zabezpieczenie architektury wewnętrznej i możliwość geograficznego rozproszenia systemu na więcej niż dwie lokalizacje z zachowaniem redundancji (nadmiarowości – czyli odporności na awarię jednej lub więcej lokalizacji oraz ich komponentów);
- 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 obiektowych Scale-OUT:

Współczesne systemy macierzy obiektowych 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 oprócz typowego dostępu obiektowego za pomocą standardu S3 oraz interfejsu REST-API, możliwość obsługi protokołów plikowych (do importu/eksportu danych) tj. NFS, SMB/CIFS, dostęp w ramach architektury analitycznej Hadoop – protokół HDFS;

- 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;

- wersjonowanie: są to dodatkowe kopie danych powstające każdorazowo przy edycji wybranego dokumentu, pozwalające na ich odtworzenie w przypadku utraty źródłowego materiału lub chęci powrócenia do poprzedniej wersji dokumentu;

- limity powierzchni dla użytkowników (tzw. Quoty): skuteczne zarządzanie macierzą wymaga możliwości elastycznego zakładania limitów na wykorzystanie powierzchni macierzy dla użytkowników macierzy;

- Geo-replikacja z efektywnym wykorzystaniem powierzchni dyskowej: możliwość wykorzystania rozproszenia danych oraz dodatkowych informacji (tzw. kontroli parzystości danych) w celu uniknięcia konieczności wielokrotnego replikowania całości danych w przypadku ich rozproszenia na więcej niż dwie lokalizacje geograficzne.

- 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;

- 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 obiektowych 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 obiektów lub grupy obiektów zawierających takie same atrybuty (metadane), np. dokumenty dotyczące wybranego projektu, grupy roboczej, z wybranego roku, itp.

Zarządzanie i monitoring:

Poniżej przedstawiamy 3 podstawowe funkcje dotyczące monitorowania stanu macierzy obiektowych 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 zasobów obiektowych).