White Paper

Dlaczego rozwijanie i wdrażanie technologii AI na stanowiskach pracy ma sens?

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

OPINIA IDC

We wszystkich branżach sztuczna inteligencja (SI) stała się charakterystyczną funkcją o istotnym znaczeniu, a wymagany do jej obsługi sprzęt jest szybko rozwijany. Branża technologiczna jest często mocno skupiona na wykładniczym wzroście rozmiaru większości najbardziej zaawansowanych modeli SI. Toczące się rozmowy dotyczą dziesiątek miliardów parametrów, kwestii optymalizacji, możliwości rozszerzania pamięci oraz potrzeb w zakresie wydajnych obliczeń (HPC) na potrzeby trenowania i stosowania modeli SI, a także stojaków z serwerami przyspieszającymi obliczenia. W rzeczywistości obecność tak ogromnej infrastruktury do obsługi SI jest wyjątkiem, zwłaszcza w przedsiębiorstwach.

Obecnie w wielu przedsiębiorstwach trwają intensywne prace nad inicjatywami związanymi ze SI, w tym generatywną sztuczną inteligencją (GSI), przy czym realizacja tych inicjatyw nie wymaga dostępu do superkomputera. W rzeczywistości wiele takich prac nad tworzeniem modeli SI (i coraz częściej nad ich wdrażaniem, zwłaszcza na brzegu sieci) odbywa się na bardzo wydajnych stacjach roboczych. W kontekście tworzenia i wdrażania modeli SI stacje robocze mają wiele zalet. Uwalniają one zajmujących się SI naukowców lub deweloperów od konieczności konkurowania o zasoby serwerowe, zapewniają akcelerację GPU (nawet jeśli serwerowe procesory GPU nadal nie są łatwo dostępne w centrach przetwarzania danych), są niezwykle przystępne cenowo w porównaniu z serwerami, stanowią niższy jednorazowy wydatek niż szybko kumulujące się opłaty za instancje chmurowe oraz dają poczucie komfortu wynikające ze świadomości, że dane wrażliwe są bezpiecznie przechowywane w obiektach przedsiębiorstwa. W ten sposób uwalniają również naukowców lub deweloperów od obaw związanych z ponoszeniem kosztów podczas samego eksperymentowania na modelach SI.

Według IDC wdrażanie rozwiązań SI na brzegu sieci postępuje szybciej niż w chmurze lub w obiektach przedsiębiorstwa. Również w tym przypadku stacje robocze odgrywają coraz ważniejszą rolę jako platformy SI, często nie wymagając nawet procesorów GPU, ale obsługując wnioskowanie SI na bazie zoptymalizowanych programowo procesorów CPU. Zastosowania SI na stacjach roboczych na brzegu sieci szybko stają się coraz liczniejsze i obejmują takie obszary, jak operacje IT, reakcje na katastrofy, radiologia, wydobycie ropy naftowej i gazu, zarządzanie gruntami, telezdrowie, zarządzanie ruchem, monitorowanie zakładów produkcyjnych i obsługa dronów.

W opracowaniu tym przeanalizowano rosnącą rolę stacji roboczych w tworzeniu i wdrażaniu modeli SI oraz w skrócie omówiono ofertę Dell w zakresie stacji roboczych do obsługi SI.

July 2023, IDC #US51017223

OMÓWIENIE SYTUACJI

Wpływ eksplozji SI na infrastrukturę

Liczba związanych ze sztuczną inteligencją projektów, w których uczestniczą organizacje na całym świecie, szybko rośnie. Już teraz wiele zadań we wszystkich branżach wykonuje oprogramowanie w całości lub w części oparte na modelach SI. IDC śledzi postępy w zakresie SI na wielu poziomach, a jednym z wartych uwagi wskaźników jest prognozowana kwota, jaką przedsiębiorstwa i dostawcy usług chmurowych planują wydać na serwery służące do tworzenia i uruchamiania modeli SI. Do 2026 r. kwota ta wyniesie 34,6 mld USD, co stanowi niemal 22% całkowitych wydatków na serwery na całym świecie.

Serwery stanowią jednak tylko element całej układanki. Ogromna część prac związanych z przygotowaniem, tworzeniem, budową prototypów i coraz częściej *wdrażaniem* modeli SI ma miejsce na stacjach roboczych. W miarę jak małe i duże przedsiębiorstwa doszły do wniosku, że nowe możliwości biznesowe mogą być realizowane poprzez wzbogacenie posiadanych aplikacji o funkcje SI, prace eksperymentalne nad modelami SI gwałtownie przyspieszyły, przy czym okazało się, że zaawansowane stacje robocze idealnie się do takich prac nadają ze względu na ich natychmiastową dostępność i bliskość do danych.

W jaki sposób sztuczna inteligencja stała się nagle tak powszechna, biorąc pod uwagę fakt, że algorytmy SI są stosowane od dekad? Dzieje się tak przede wszystkim dlatego, że w ciągu ostatnich kilku lat spełnione zostały dwa podstawowe warunki odpowiedniego działania szczególnie skutecznego typu algorytmu SI, czyli sieci neuronowej. Warunki te obejmowały łatwą dostępność ogromnych, tanich i różnorodnych rodzajów danych, w tym danych nieuporządkowanych i na wpół uporządkowanych, oraz rozszerzenie obliczeń liniowych o model równoległy umożliwiający przetwarzanie danych w tych sieciach neuronowych w akceptowalnych ramach czasowych. Po spełnieniu obu tych warunków danetycy (data scientists) poczynili ogromne postępy w rozwoju sieci neuronowych, które automatycznie uczą się, jak wykonywać coraz bardziej efektowne zadania. O ile tradycyjne mechanizmy samouczenia się maszyn (ML) są nadal przydatne w odniesieniu do danych tekstowych lub liczbowych, o tyle uczenie głębokie (DL) jest skuteczniejsze w przypadku filmów, nagrań dźwiękowych, tłumaczeń itp.

Tradycyjne modele samouczenia się maszyn można wdrażać na procesorach CPU stacji roboczej, które mają co najwyżej kilkadziesiąt rdzeni, natomiast sieci neuronowe wymagają koprocesorów na potrzeby równoległego przetwarzania danych na tysiącach rdzeni. Główną tego przyczyną jest fakt, że w modelu ML wyodrębnianie i klasyfikacja cech jest procesem wykonywanym ręcznie, podczas gdy w modelu DL jest to proces zautomatyzowany, który wymaga trenowania modelu w drodze ciągłego powtarzania przy użyciu dużych zbiorów danych. Obecnie najpopularniejszym koprocesorem jest GPU, ale do użytku wchodzą także opracowane przez start-upy nowe procesory właściwe dla SI. Tego typu akceleracja wykorzystująca oddzielny koprocesor do przetwarzania równoległego zrewolucjonizowała rynek serwerów i stacji roboczych i zapoczątkowała coś, co IDC nazywa obliczeniami masowo równoległymi (massively parallel compute).

W 2022 r. wartość światowego rynku serwerów przyspieszających obliczenia wynosiła 21,8 mld USD i do 2026 r. ma wzrosnąć do 43,4 mld USD, z czego 57% przypada na serwery przyspieszające obliczenia na potrzeby sztucznej inteligencji. Jednocześnie liczba oddzielnych procesorów GPU sprzedanych w 2022 r. do użycia w stacjach roboczych wzrosła do 6,4 mln sztuk. IDC szacuje, że wartość rynku stacji roboczych używanych do celów naukowych lub celów związanych z inżynierią oprogramowania (w coraz większym stopniu inspirowanych rozwojem SI) wzrośnie do 2026 r. do blisko 2 mld USD.

Etapy rozwoju sztucznej inteligencji

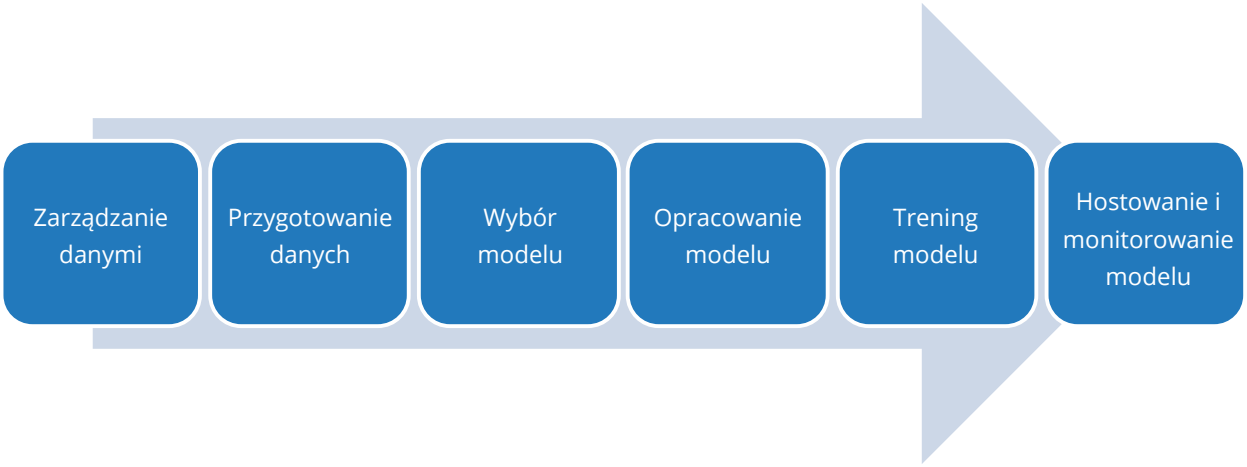
Jak już wspomniano, sieci neuronowe stały się realne ze względu na rosnące ilości i typy danych oraz nowe podejścia do obliczeń. Pierwszy element, czyli ilość i typ danych, jest całkiem istotny, ponieważ według niektórych raportów aż 80% nakładu pracy związanego z inicjatywą uczenia głębokiego w ramach trenowania sztucznej inteligencji skupia się na przygotowaniu danych i zarządzaniu nimi. Dane muszą zostać odpowiednio pozyskane, zapisane i przygotowane przed zaprojektowaniem i rozpoczęciem treningu modelu. Etapy rozwoju sztucznej inteligencji według IDC przedstawiają się następująco (zob. rys. 1):

- **Zarządzanie danymi.** Identyfikacja odpowiednich danych do modelu SI (pozyskiwanych, generowanych lub nabywanych przez organizację) oraz zarządzanie nimi na podstawie ogromnych ilości danych z centrum przetwarzania danych, brzegu sieci i chmury. Dane te mogą być dowolnego typu, na przykład mogą być generowane w odpowiedzi na konkretne zdarzenia lub przesyłane strumieniowo, a wiele z nich może wymagać określonego sposobu zarządzania.
- **Przygotowanie danych.** Przechowywanie danych (plików, bloków lub obiektów) w hurtowni lub jeziorze danych, czyszczenie ich, potwierdzanie ich kompletności i wysokiej jakości, a następnie konwertowanie ich do postaci, która umożliwi wykorzystanie ich w modelu SI, na przykład za pomocą narzędzi takich jak Spark lub Pandas.
- **Wybór modelu.** Określenie (pod względem poziomu błędów lub wydajności) modelu optymalnie wykonującego zadanie SI, dla którego został zaprogramowany.
- **Opracowanie modelu.** Zaprojektowanie modelu SI przy użyciu struktur takich jak XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn lub H2O.
- **Trening modelu.** Trening modelu za pomocą infrastruktury obliczeniowej z wystarczającą liczbą rdzeni procesorów lub koprocesorów na potrzeby równoległego przetwarzania danych (w coraz większym stopniu obejmujący również możliwość wyjaśniania, walidacji i dokumentowania decyzji podejmowanych przez model w celu zapewnienia uczciwości, odpowiedzialności i przejrzystości). Etap ten obejmuje przygotowanie prototypu na potrzeby przetestowania wytrenowanego modelu poprzez przeprowadzenie na nim wnioskowania.
- **Hostowanie i monitorowanie modelu.** Produkcyjne wdrożenie modelu w celu wykonania zadania, do którego został zaprojektowany, zazwyczaj określanego jako „wnioskowanie na bazie SI” (AI inferencing), oraz monitorowanie jego wydajności.

W połączeniu z centrum przetwarzania danych, chmurą lub infrastrukturą brzegową stacje robocze mogą odgrywać ważną rolę na każdym z tych sześciu etapów.

RYSUNEK 1

Etapy rozwoju sztucznej inteligencji



Źródło: IDC, 2023 r.

TWORZENIE MODELI SI NA STACJACH ROBOCZYCH

Stacje robocze a komputery osobiste

Powszechnie wiadomo, że komputery osobiste (PC) nie są wystarczająco wydajne na potrzeby tworzenia modeli SI. Danetycy i deweloperzy SI są zazwyczaj zaangażowani w projekty o strategicznym znaczeniu dla swoich organizacji, a niezakłócona produktywność ma w tym przypadku ogromne znaczenie. Stacje robocze zazwyczaj działają bardziej przewidywalnie niż komputery osobiste, ponieważ mają na ogół bardziej wydajne komponenty i są zoptymalizowane pod kątem działającego na nich oprogramowania.

Wśród tych komponentów można wymienić:

- **Wysokiej klasy procesory CPU –** na przykład skalowalne Intel Xeon.
- **Wydajne procesory GPU –** na przykład NVIDIA RTX 6000 Ada.
- **Więcej miejsca na dysku –** niektóre stacje robocze oferują nawet 60 TB miejsca na dysku, a prędkości we/wy są zwykle znacznie wyższe niż w przypadku komputerów osobistych.
- **Więcej pamięci –** stacje robocze oferują nawet 6 TB pamięci.
- **Chłodzenie –** wydajne komponenty generują dużo ciepła, dlatego danetycy potrzebują stacji roboczych z odpowiednim chłodzeniem, aby zapobiec ich przegrzaniu i utrzymać optymalną wydajność.
- **Karta sieciowa –** dla danetyków, którzy operują na dużych zbiorach danych przechowywanych na zdalnych serwerach, szybka karta sieciowa jest niezbędna do szybkiego i wydajnego przesyłania danych.
- **Wyświetlacz –** wysokiej jakości wyświetlacz pozwala na odpowiednią wizualizację danych, dlatego danetycy powinni szukać monitorów o wysokiej rozdzielczości, dokładnym odwzorowaniu kolorów i dużym rozmiarze ekranu.

- **Pamięć z systemem kodowania korekcyjnego (ECC) –** wykrywa i koryguje najczęstsze rodzaje wewnętrznych uszkodzeń danych, zapobiegając pojawianiu się podczas długiego treningu SI niebieskich ekranów będących wynikiem błędu krytycznego (zły bit) lub innego błędu oprogramowania (odwrócony bit powodujący złe wartości). Pamięć ECC zapewnia również dokładność wyników, co jest niezbędnym wymogiem w przypadku zastosowań o krytycznym znaczeniu dla życia człowieka, na przykład w opiece zdrowotnej.
- **Specjalne układy scalone –** na przykład procesory VPU Intel Movidius, które są koprocesorami przetwarzania równoległego do zastosowań związanych z rozpoznawaniem obrazów i aplikacjami SI na brzegu sieci, wykorzystywanymi w takich branżach, jak handel detaliczny, bezpieczeństwo i automatyka przemysłowa. W stacjach roboczych można również znaleźć układy FPGA używane na przykład do zastosowań finansowych.
- **Oprogramowanie optymalizacyjne –** na przykład OneAPI, czyli oparty na standardach model programowania firmy Intel, który upraszcza tworzenie i wdrażanie obciążeń roboczych zorientowanych na dane w środowiskach z procesorami CPU i GPU, układami FPGA i innymi akceleratorami, lub CUDA, czyli platforma obliczeń równoległych i interfejs programowania aplikacji firmy NVIDIA służący do uruchamiania ogólnych obciążeń roboczych w środowiskach z procesorami GPU.

Procesory CPU a procesory GPU w kontekście SI

Stacje robocze mogą być wykorzystywane na różnych etapach rozwoju SI i są zazwyczaj wyposażone w różne odpowiednie funkcje. Pomimo dużego znaczenia procesorów GPU dla przetwarzania równoległego, procesory CPU odgrywają kluczową rolę podczas tworzenia modelu SI na stacji roboczej. Podobnie jak procesory GPU, procesory CPU mogą być również używane do manipulowania danymi i rzecz jasna do tworzenia tradycyjnych modeli ML. Procesory CPU są również używane do eksploracji danych, czyli w celu korzystania z wizualnych reprezentacji zestawu danych na potrzeby zrozumienia charakterystyki danych.

W przypadku uczenia głębokiego rola procesorów CPU hosta jest nieco mniejsza, ponieważ procesory GPU przejmują kontrolę podczas rzeczywistego procesu uczenia. Nawet jednak wtedy procesory CPU nadal służą jako warstwa przetwarzania danych dla krytycznego oprogramowania takiego jak system operacyjny lub architektura CUDA, a także do orkiestracji procesów między procesorami GPU lub z innymi układami scalonymi. Ponadto procesory CPU coraz częściej przyjmują nową rolę jako silniki wnioskowania SI w sytuacjach, gdy stacja robocza jest używana do produkcyjnego uruchamiania modelu SI. IDC przewiduje, że do 2024 r. kwota wydatków na infrastrukturę do obsługi wnioskowania na bazie SI przekroczy kwotę wydatków na infrastrukturę SI do szkolenia SI, a znaczna część (39%) tego wnioskowania będzie odbywać pod kontrolą procesorów CPU hosta.

Stacje robocze a serwery — relacja symbiotyczna

Dla większości organizacji wdrożenie stacji roboczej, serwera lokalnego, instancji w chmurze lub dowolnej kombinacji tych trzech rozwiązań na potrzeby tworzenia modeli SI wynika z czystego pragmatyzmu. Na różnych etapach projektu SI istnieje swoista relacja symbiotyczna między stacjami roboczymi, serwerami i instancjami w chmurze.

Zaletą stacji roboczych w porównaniu z serwerami w centrach przetwarzania danych jest to, że w przypadku stacji roboczych danetycy mogą pracować z dowolnego miejsca, co jest ważnym czynnikiem w kontekście obecnej pandemii, ale także w zwykłych okolicznościach. Ponadto danetycy mogą wówczas swobodnie eksperymentować na swoich modelach SI, dokonując iteracji tak często, jak uznają to za konieczne, ponieważ moc obliczeniowa nowoczesnych stacji roboczych z wydajnymi

procesorami GPU często pozwala na bardziej interaktywny proces iteracji, zapewniając natychmiastowe informacje zwrotne i wyniki bez konieczności składania wniosków o dostęp do zasobów serwerowych lub borykania się z innymi ograniczeniami centrów przetwarzania danych. Ponadto stacje robocze zapewniają elastyczność w zakresie przenoszenia obliczeń bliżej danych, a nie odwrotnie, co pozwala zaoszczędzić przepustowość, zmniejszyć zagęszczenie i zwiększyć przepływność danych. Stacje robocze można również skonfigurować pod kątem różnych potrzeb, na przykład tradycyjnych zadań ML lub bardziej intensywnych zadań DL.

Warto zauważyć, że pomimo znacznego wzrostu wartości rynku serwerów przyspieszających obliczenia, serwery te nadal nie są powszechnie dostępne w korporacyjnych centrach przetwarzania danych. W momencie powstawania tego opracowania średnio 4% serwerów w korporacyjnych centrach przetwarzania danych było serwerami przyspieszającymi obliczenia, co oznacza, że wiele organizacji nie dysponuje środkami umożliwiającymi tworzenie lub uruchamianie aplikacji SI na łatwo dostępnych, zainstalowanych lokalnie procesorach GPU. Również z tego powodu stacje robocze przyspieszające obliczenia są przydatną alternatywą w kontekście tworzenia modeli SI.

Stacje robocze znacząco przyspieszające obliczenia są obecnie wydajne na tyle, że mogą przeprowadzać trening DL, o ile model SI nie jest zbyt duży, eliminując tym samym potrzebę prowadzenia treningu na serwerach. Modele trenowane na stacjach roboczych z procesorami GPU mogą być natomiast wdrażane zarówno na stacjach roboczych, jak i na serwerach bez procesorów GPU i wykorzystywać jedynie potencjał procesorów CPU na potrzeby wnioskowania na bazie SI. Specjalne oprogramowanie (na przykład Intel DL Boost lub oneAPI) może obsługiwać wspomniane wnioskowanie przy pomocy procesorów CPU, umożliwiając wdrożonym już w centrach przetwarzania danych serwerom nieprzyspieszającym obliczeń obsługę aplikacji SI.

Stacje robocze a chmura

Chmura obliczeniowa zrewolucjonizowała sposób postrzegania infrastruktury, danych i aplikacji przez organizacje. Dzięki niemal nieograniczonej skalowalności chmura obliczeniowa umożliwia deweloperom przydzielanie zasobów na żądanie, potencjalnie zwiększając tempo innowacji przy mniejszej liczbie ograniczeń. Na pierwszy rzut oka chmura wydaje się idealnym narzędziem do tworzenia modeli SI.

Nie zawsze jednak tak jest. Z badań IDC wynika, że organizacje coraz częściej przenoszą określone obciążenia robocze z chmury publicznej do infrastruktury zainstalowanej lokalnie. Wynika to z następujących czynników:

- **Dostępność chmury.** Każdy użytkownik usług chmurowych doświadczył przestojów w ich świadczeniu, czy to z powodu problemów u samego dostawcy rozwiązań chmurowych, czy też z powodu utraty łączności sieciowej między hiperskalowym centrum przetwarzania danych a użytkownikiem końcowym. W takich sytuacjach praca ustaje, a użytkownik jest zdany na łaskę i niełaskę dostawcy usług chmurowych.
- **Bezpieczeństwo i zgodność z przepisami.** W wielu branżach zasady ładu korporacyjnego określają, gdzie dane mogą być przekazywane i przechowywane, co ogranicza zakres korzystania z usług chmurowych. Przepisy prawa, takie jak przyjęte w Unii Europejskiej RODO i przyjęta w Kalifornii ustawa o ochronie prywatności konsumentów, również narzucają określone zasady dotyczące suwerenności danych.

- **Koszty.** Organizacje często nie zdają sobie sprawy z tempa wzrostu opłat za usługi chmurowe, zwłaszcza w kontekście obciążeń roboczych wymagających dużej mocy obliczeniowej i dużej ilości pamięci masowej. Opłacalność chmury opiera się na pomiarach wszystkich rodzajów zużycia zasobów, w tym kosztów danych przesyłanych z powrotem do infrastruktury lokalnej.
- **Presja związana z eksperymentowaniem.** Większość inicjatyw SI rozpoczyna się od znacznej liczby eksperymentów, w przypadku których pojawianie się modeli, które zawodzą, jest nieodłączną częścią całego procesu. Przy takiej metodzie prób i błędów można zatem dostrzec pewien psychologiczny podatek, którym danetycy i deweloperzy SI są obciążani, gdy płatności za korzystanie z chmury kumulują się, a oni nie są jeszcze w stanie przedstawić zadowalających wyników pracy.

Stacje robocze mogą być odpowiedzią na te ograniczenia, a jednocześnie nadal korzystać z działających w chmurze technologii takich jak architektury oparte na mikrousługach i mechanizmy automatyzacji oparte na interfejsach API. Przynosi to część takich samych korzyści, jak w przypadku porównania stacji roboczych z serwerami w centrach przetwarzania danych:

- **Praca w dowolnym miejscu.** Eliminacja zależności od chmury publicznej umożliwia teraz realizację scenariuszy odłączenia się od sieci. Wiele szczególnie zabezpieczonych środowisk jest odizolowanych od sieci publicznych, a stacje robocze korzystające z SI mogą w wyjątkowy sposób stać się elementem takiego zabezpieczenia. Zainstalowane lokalnie zasoby zmniejszają również zapotrzebowanie na kosztowną łączność sieciową.
- **Lokalność danych.** Upowszechnianie się urządzeń Internetu rzeczy (IoT) i innego zintegrowanego sprzętu przyczynia się do wykładniczego wzrostu ilości danych w lokalizacjach brzegowych. W wielu sytuacjach sensowne jest lokalizowanie zasobów obliczeniowych w jednym miejscu z dedykowaną stacją roboczą. Taka kolokacja rozwiązuje również wiele problemów związanych ze zgodnością z przepisami, ponieważ przyczynia się do ograniczenia przepływu danych.
- **Swobodne eksperymentowanie.** Trening i optymalizacja modeli SI to proces iteracyjny, który często obejmuje próby i błędy. Deweloperzy potrzebują swobody przeprowadzania eksperymentów bez ograniczeń wynikających z potencjalnych dodatkowych opłat za usługi. Stacje robocze zapewniają również większą elastyczność w zakresie niestandardowych narzędzi.

W przypadku tej ostatniej kwestii porównanie ceny stacji roboczej z ceną wdrożenia w chmurze jest stosunkowo łatwe, ponieważ większość dostawców rozwiązań chmurowych podaje natychmiastowe szacunki kosztów dowolnej konfiguracji, którą użytkownik końcowy chce wdrożyć. Na przykład koszt zwykłej maszyny wirtualnej z kartą NVIDIA T4 i instancją pamięci masowej SSD o pojemności 375 GB, która jest używana przez osiem godzin dziennie, pięć dni w tygodniu, wynosi u jednego z głównych dostawców rozwiązań chmurowych 140 USD. Podwojenie liczby maszyn wirtualnych, kart T4 i dysków SSD zwiększy koszty do 365 USD miesięcznie. Pozostanie przy dwóch maszynach wirtualnych, ale podwojenie liczby kart T4 i dysków SSD o pojemności 375 GB do czterech sztuk oraz przeprowadzenie pełnowymiarowego treningu w tym środowisku zwiększy koszty do 2700 USD miesięcznie. Można więc śmiało stwierdzić, że koszty tworzenia modelu SI w chmurze mogą łatwo wzrosnąć do dziesiątek tysięcy dolarów rocznie, czyli wynieść znacznie więcej niż wynosi roczna amortyzacja stacji roboczej klasy wyższej.

PRZYGOTOWANIE PROTOTYPU MODELU SI NA STACJACH ROBOCZYCH

W porównaniu zarówno z serwerami zainstalowanymi lokalnie, jak i chmurą stacje robocze zapewniają wyraźną przewagę w kontekście przygotowywania prototypów modeli SI. Serwery w centrum przetwarzania danych mogą być w pełni wykorzystane lub mieć zbyt krytyczne znaczenie dla wspomnianego tworzenia prototypów i testowania modeli SI, a jak już wspomniano, dopuszczalne koszty instancji w chmurze mogą zostać szybko przekroczone, gdy instancje te będą swobodnie używane w charakterze środowiska testowego. Dzięki zastosowaniu stacji roboczych danetycy lub deweloperzy zajmujący się SI nie muszą już konkurować o zasoby serwerowe ani przejmować się rosnącymi opłatami za korzystanie z chmury na etapie przygotowywania prototypów. Niskie koszty jednorazowe zapewniają bowiem pełną swobodę tworzenia prototypów w dowolnym miejscu i czasie i nie wiążą się z ponoszeniem innych dodatkowych kosztów.

WDRAŻANIE MODELI SI NA STACJACH ROBOCZYCH

Tworzenie modeli SI na stacjach roboczych jest powszechną strategią od lat, IDC obserwuje jednak coraz więcej przypadków *wdrażania* modeli SI na stacjach roboczych, zazwyczaj na brzegu sieci. Innymi słowy, model SI jest wdrażany produkcyjnie na stacjach roboczych, a zadaniem stacji jest obsługa wnioskowania na bazie takiego modelu. Modele SI są coraz częściej wdrażane na serwerach na brzegu sieci. W latach 2020–2024 roczne wydatki na sprzęt w tym obszarze wzrosły ponad trzykrotnie. Stacje robocze nie pozostają daleko w tyle, ponieważ użytkownicy końcowi zaczęli doceniać ich zalety na brzegu sieci.

IDC definiuje brzeg sieci jako rozproszony model obliczeniowy, który obejmuje wdrażanie infrastruktury i aplikacji poza scentralizowaną chmurą i lokalnymi centrami przetwarzania danych, tak blisko, jak to konieczne, do miejsca, w którym dane są generowane i przetwarzane. Obejmuje to oddziały i oddalone lokalizacje, a także lokalizacje właściwe dla danej branży, takie jak zakłady produkcyjne, magazyny, szpitale i punkty sprzedaży detalicznej.

Obciążenia robocze wymagające dużej ilości danych i mocy obliczeniowej są coraz częściej przenoszone do lokalizacji lokalnych lub brzegowych. Ma to na celu zminimalizowanie ograniczeń związanych z korzystaniem z chmur publicznych, takich jak czas potrzebny na przesłanie dużych zbiorów danych i zmienne koszty treningu SI, zwłaszcza w sytuacjach, które wymagają znacznej liczby eksperymentów danetycznych.

Z badań IDC wynika, że brzeg sieci jest szybko rozwijającą się lokalizacją wdrażania SI, a w 2023 r. organizacje zainwestują 2,9 mld USD w obliczenia związane z SI na brzegu sieci, przy czym w 2026 r. kwota ta wzrośnie do 6,9 mld USD (zob. *Worldwide AI Hardware Forecast, 2022-2026: Strong Market Growth for AI Compute and Storage*, IDC #US49671722, wrzesień 2022 r.). Ponadto brzeg sieci zyskuje na popularności jako miejsce wdrażania obciążeń roboczych HPC, w tym obciążeń o charakterze inżynieryjnym i technicznym, a przedsiębiorstwa inwestują obecnie prawie 1 mld USD w takie obciążenia, przy czym do 2027 r. kwota ta wzrośnie do 2,4 mld USD (zob. *Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs*, IDC #US50525123, kwiecień 2023 r.). Są to obszary, w których warto wdrożyć stację roboczą SI.

Podczas wdrażania modelu SI na stacji roboczej na brzegu sieci nie zawsze istnieje potrzeba stosowania procesorów GPU klasy wyższej, tak jak ma to miejsce w przypadku tworzenia modelu SI. Procesory GPU klasy niższej mogą obsługiwać wnioskowanie na bazie SI, a w niektórych przypadkach procesory te nie są w ogóle potrzebne. W takich przypadkach procesory CPU

mogą odpowiednio obsłużyć wspomniane zadanie wnioskowania, zwłaszcza gdy są zoptymalizowane przy użyciu takich narzędzi jak Intel DL Boost, czyli zestawu instrukcji mających na celu przyspieszenia przetwarzania przez mikroprocesory Intel obciążeń roboczych, w tym obciążeń związanych z wnioskowaniem na bazie SI. Firma Intel twierdzi, że dzięki funkcji Intel DL Boost odnotowała o 1,45 raza wyższą przepływność wnioskowania INT8 w czasie rzeczywistym przy użyciu obsługujących tę funkcję skalowanych procesorów 4. generacji Intel Xeon w porównaniu z poprzednią generacją tych procesorów (BERT-Large SQuAD). Sprawia to że stacja robocza jest bardziej odpowiednia do wdrożenia na brzegu sieci, gdzie względy takie jak moc obliczeniowa, mobilność i zarządzanie temperaturą wymagają mniejszej mocy w watach. Procesor Intel Movidius Myriad (M2) dobrze wpisuje się w ten przedział poboru mocy, dzięki swojemu niewielkiemu zużyciu energii na poziomie 12 W.

Wdrożenie SI na stacjach roboczych

W kilku sytuacjach wdrożenie SI na zainstalowanych lokalnie stacjach roboczych jest całkiem naturalne. Cechami wspólnymi tych sytuacji są duże ilości generowanych maszynowo danych szeregów czasowych i danych nieuporządkowanych, takich jak strumienie danych wideo i obrazów. Istnieją również sytuacje, w których specjaliści merytoryczni muszą rozszerzyć modele SI o interpretacje człowieka.

Przykłady:

- **Operacje IT na bazie sztucznej inteligencji (AIOps).** Wraz ze wzrostem skali i złożoności systemów IT pojawia się coraz większa potrzeba przejścia od reaktywnego zarządzania zdarzeniami do proaktywnego ich monitorowania. Jest to szczególnie prawdziwe wtedy, gdy elementy infrastruktury i aplikacje są rozmieszczane w lokalizacjach brzegowych, w których personel techniczny jest nieliczny lub nie ma go wcale. Modelując stan odniesienia dla zwykłej wydajności, można identyfikować anomalie i automatyzować działania naprawcze.
- **Reakcja na katastrofy.** W nagłych wypadkach osoby udzielające pierwszej pomocy muszą szybko ocenić sytuację, śledzić krytyczny sprzęt i rozmieścić zasoby tak, aby pomóc najbardziej potrzebującym. Zważywszy na fakt, że często działania te muszą się odbywać w środowisku pozbawionym łączności sieciowej, przydatna jest wówczas lokalna stacja robocza, która może agregować dane, obsługiwać wnioskowanie na bazie SI oraz automatyzować komunikację z kluczowym personelem.
- **Radiologia.** Postępy w technologii obrazowania doprowadziły do zwiększenia ilości danych generowanych na podstawie jednego zdjęcia. Wymaga to przechowywania tych danych na miejscu, aby mogły zostać przeanalizowane w odpowiednim czasie. Wytrenowane na milionach wcześniejszych zdjęć modele sztucznej inteligencji mogą identyfikować wzorce dokładniej niż ludzkie oko, przyczyniając się do lepszej jakości diagnoz.
- **Wydobycie ropy naftowej i gazu.** Zajmujące się wydobyciem ropy naftowej i gazu przedsiębiorstwa wykorzystują połączenie danych telemetrycznych, sejsmicznych i obrazowych do lokalizowania złóż surowców naturalnych, wybierania lokalizacji odwiertów i optymalizowania wydajności sprzętu w procesie produkcji. Działania te często wymagają analizy informacji w miejscach, w których dostępna jest jedynie kosztowna łączność satelitarna.
- **Badania nad nowotworami i opracowywanie leków.** Lekarze z klinik medycznych i naukowcy z ośrodków akademickich wykorzystują modele sztucznej inteligencji i przetwarzania języka naturalnego, aby pomóc onkologom w znalezieniu najskuteczniejszych, zindywidualizowanych metod leczenia nowotworów. Ponadto łączą mechanizmy samouczenia się maszyn z funkcjami rozpoznawania obrazu, aby w ten sposób umożliwić radiologom lepsze zrozumienie postępów choroby nowotworowej u pacjentów. Używają w tym celu odpowiednich algorytmów, aby lepiej poznać proces rozwoju nowotworów i określić najskuteczniejsze metody ich leczenia.

- **Ocena roszczeń ubezpieczeniowych.** Ręczne przetwarzanie roszczeń jest pracochłonne i podatne na błędy ludzkie. Model SI, który jest zdolny do oceny zasadności roszczeń, obniża koszty, pozwalając rzeczoznawcom skupić się na sprawach wymagających dokładniejszego zbadania. W ten sposób rośnie wydajność bez uszczerbku dla dokładności.
- **Telezdrowie.** Model SI poprawia wskaźniki rekonwalescencji pacjentów, dostosowując indywidualne plany leczenia do parametrów życiowych uzyskiwanych w czasie rzeczywistym z urządzeń noszonych przez pacjenta. Informacje te są łączone z historiami leczenia pacjentów i danymi z bazy wiedzy o podobnych przypadkach. Jest to szczególnie ważne na obszarach wiejskich, które w większym stopniu polegają na zdalnej opiece zdrowotnej.
- **Bezpieczeństwo w handlu detalicznym (ochrona przed kradzieżami).** Narzędzia analizujące strumienie wideo w czasie rzeczywistym są używane do przewidywania ludzkich zachowań mogących prowadzić do działań przestępczych. Zazwyczaj wymaga to połączenia obrazów z wielu kamer w celu śledzenia danej osoby w sklepie. Z uwagi na to, że czas identyfikacji istotnego zdarzenia ma bardzo duże znaczenie, jest to proces, który najlepiej przeprowadzać lokalnie.
- **Zarządzanie ruchem.** Organy administracji rządowej, które odpowiadają za operacje transportowe, coraz częściej wykorzystują modele SI na potrzeby koordynacji sygnalizacji świetlnej i cyfrowego oznakowania w celu poprawy ruchu pojazdów i zapewnienia bezpieczeństwa obywatelom. Wymaga to połączenia danych wejściowych, w tym danych z kamer wideo i danych telemetrycznych z czujników drogowych, w celu optymalizacji wzorców ruchu.
- **Monitorowanie zakładów produkcyjnych.** Dla dyrektora zakładu produkcyjnego zapewnienie nieprzerwanego działania krytycznych procesów i dotrzymanie harmonogramów produkcji ma kluczowe znaczenie. Wymaga to predykcyjnej konserwacji kluczowego sprzętu, automatycznego wykrywania usterek i optymalizacji zarówno w łańcuchu dostaw, jak i poza nim. Jest to obszar, w którym sztuczna inteligencja może pomóc pracownikom w zwiększeniu wydajności bez uszczerbku dla przestrzegania norm bezpieczeństwa.
- **Drony**. Zautomatyzowana analiza obrazów zarejestrowanych przez drony umożliwia monitorowanie szerokiego zakresu warunków na skalę wcześniej nieosiągalną. Istotnie wpływa to na inspekcje instalacji gazowych i elektrycznych, badania ubezpieczeniowe, działania poszukiwawczo-ratownicze, rolnictwo precyzyjne oraz utrzymanie łowisk i rezerwatów przyrody.
- **Typowe środowiska biurowe.** Typowe środowiska biurowe są w coraz większym stopniu ulepszane przy użyciu narzędzi zwiększających produktywność opartych na sztucznej inteligencji, takich jak Microsoft Copilot.
- **Energia odnawialna.** Miejsca pozyskiwania energii odnawialnej, takie jak farmy wiatrowe, zapory wodne i farmy słoneczne, wymagają monitorowania w czasie rzeczywistym, konserwacji i zbierania danych, które muszą być generowane i analizowane lokalnie.

STACJE ROBOCZE DELL DO OBSŁUGI MODELI SI

Dell oferuje pod marką Data Science Workstation (DSW) szeroką gamę stacji roboczych dla różnych poziomów tworzenia lub wdrażania modeli SI. W tej sekcji pokrótce przedstawiono specyfikacje tych stacji, a następnie omówiono różne funkcje i zastosowania związane ze sztuczną inteligencją (danetycy, korzyści z wdrożenia technologii Dell DSW itp.). Stacje te zostały zaprojektowane specjalnie dla danetyków zajmujących się SI. Najnowsze stacje robocze Precision Data Science wykorzystują funkcje SI do precyzyjnego dostrajania urządzeń pod kątem zapewnienia optymalnej wydajności aplikacji, z których najczęściej korzystają danetycy. W rezultacie mogą oni szybciej wykonywać swoją najważniejszą pracę. Ponadto stacje robocze Dell Precision są testowane i certyfikowane przez niezależnych dostawców oprogramowania, aby zapewnić odpowiednią obsługę wydajnych aplikacji potrzebnych klientom Dell do wykonywania codziennych zadań.

Stacje robocze Dell na tle konkurencji

Stacje robocze Dell Precision z procesorami GPU NVIDIA RTX mają zapewniać wysoką skalowalność i wydajność inicjatyw związanych z analizą danych i sztuczną inteligencją. Dell Technologies oferuje kompleksowe rozwiązania sprzętowe zoptymalizowane pod kątem obsługi najnowszego oprogramowania SI, które wyróżnia:

- **Świetna konfiguracja sprzętu.** Stacje robocze Dell Precision oferują szereg wydajnych konfiguracji sprzętowych obejmujących m.in. procesory wielordzeniowe, pamięć operacyjną o dużej pojemności i wiele opcji GPU. Komponenty te zapewniają niezbędne zasoby obliczeniowe do wykonywania zadań SI, pozwalając na efektywny trening modeli SI, a następnie efektywne ich stosowanie.
- **Skalowalność i możliwości dostosowania.** Stacje robocze Dell Precision można skalować i dostosowywać do swoich specyficznych wymagań związanych ze sztuczną inteligencją. Elastyczność ta zapewnia, że stacja robocza może być zoptymalizowana pod kątem konkretnych potrzeb związanych z obciążeniami roboczymi SI.
- **Certyfikacja i optymalizacja.** We współpracy z firmą NVIDIA Dell certyfikuje stacje robocze Precision pod kątem wydajności i zgodności z procesorami GPU NVIDIA RTX, w tym kartami NVIDIA RTX 6000 Ada Generation. Certyfikacja ta potwierdza płynność integracji i zoptymalizowaną wydajność podczas używania stacji roboczych Dell Precision z procesorami GPU NVIDIA RTX do zadań związanych ze sztuczną inteligencją.
- **Wydajne przetwarzanie danych.** Stacje robocze Dell Precision z procesorami Intel zapewniają moc obliczeniową niezbędną do wykonywania zadań związanych ze sztuczną inteligencją. Dzięki wielordzeniowym procesorom i dużym szybkościom zegara stacje te oferują wydajność niezbędną do efektywnego treningu i stosowania modeli SI.
- **Wsparcie w zakresie oprogramowania i narzędzi.** Stacje robocze Dell Precision są fabrycznie wyposażone w oprogramowanie i narzędzia wspierające tworzenie i wdrażanie modeli SI Obejmują one zoptymalizowane pakiety oprogramowania, struktury sztucznej inteligencji i biblioteki, które wykorzystują procesory GPU NVIDIA RTX, ułatwiając użytkownikom pierwsze kroki w projektach związanych ze sztuczną inteligencją.

Omówione w kolejnych sekcjach technologie to kolejne ważne obszary, w których stacje robocze Dell wyróżniają się na tle konkurencji.

Reliable Memory Technology

Dell uzupełnia pamięci ECC o technologię RMT Pro (Reliable Memory Technology Pro), której celem jest maksymalizacja czasu pracy bez przestojów. We współpracy z pamięcią ECC technologia ta wykrywa i koryguje błędy pamięci w czasie rzeczywistym. Według Dell technologia RTM Pro praktycznie eliminuje błędy pamięci, zapobiegając ponownemu wykorzystaniu uszkodzonej pamięci, nawet jeśli moduł DIMM jest w pełni używany. Po ponownym uruchomieniu systemu technologia RTM Pro izoluje wadliwy obszar pamięci i ukrywa go przed systemem operacyjnym. W rezultacie zajmujący się SI danetycy i deweloperzy nie będą już narażeni na ciągłe awarie wynikające z możliwości adresowania złej pamięci, co znacznie zwiększy ich produktywność.

Dell Optimizer for Precision

W większości swoich stacji roboczych Dell oferuje również oprogramowanie Dell Optimizer for Precision, które automatycznie dostosowuje ustawienia systemowe tak, aby stacja robocza uruchamiała różne popularne aplikacje komercyjne z największą możliwą szybkością. Zwiększa to produktywność danetyków i deweloperów. Oprogramowanie to tworzy również raporty o wydajności w czasie rzeczywistym dla działu IT, które dotyczą wykorzystania procesora, pamięci masowej, pamięci operacyjnej i grafiki. Dell Optimizer for Precision nie działa jeszcze na Linuksie i dlatego jest przydatny głównie do wdrażania modeli SI, ponieważ tworzenie modeli SI zwykle odbywa się za pomocą oprogramowania open source opartego na Linuksie. Ponadto Dell Optimizer for Precision oferuje funkcje ExpressSign-in, Express Charge (na telefonach komórkowych) i Intelligent Audio oraz narzędzia raportujące i analityczne, które pomagają precyzyjnie dostroić działanie danej stacji roboczej.

WYZWANIA I MOŻLIWOŚCI

Dla przedsiębiorstw

Na rynku SI firma IDC obserwuje swoiste rozwidlenie. Z jednej strony przedsiębiorstwa wdrażają strategie dotyczące danych, aby pozostać konkurencyjnymi, w tym na szeroką skalę wykorzystują modele SI. Przedsiębiorstwom tym są na przykład przedstawiane inne firmy z danej branży, które po wdrożeniu oferowanej im infrastruktury SI osiągnęły fantastyczne wyniki, przy czym infrastruktura ta jest obsługiwana przez sto najlepszych superkomputerów na świecie. Z drugiej strony codziennością dla przedsiębiorstw są małe projekty związane ze sztuczną inteligencją, obsługiwane na dostępnych serwerach w centrum przetwarzania danych lub w chmurze, często z niewystarczającym budżetem i nieefektywnym sprzętem.

Dla wielu przedsiębiorstw pierwszy scenariusz nie jest realny, a drugi jest aż zanadto realny. Wyzwaniem jest dla nich zapewnienie danetykom lub deweloperom zajmującym się SI odpowiednich narzędzi do trenowania modeli SI w odpowiednim czasie bez wydawania ogromnych sum pieniędzy na instancje serwerów w chmurze lub serwery przyspieszające obliczenia z akceleracją GPU w centrach przetwarzania danych. Zdaniem IDC wystarczy, jeśli przedsiębiorstwa te wyposażą swoich naukowców i programistów w wydajne stacje robocze z akceleracją GPU.

Dla Dell

Na rynku panuje błędne przekonanie, że tworzenie i wdrażanie modeli SI wymaga drogich serwerów przyspieszających obliczenia, często nawet w klastrze. Może to być prawdą w przypadku najbardziej zaawansowanych algorytmów SI z miliardami parametrów, większość przedsiębiorstw nie tworzy jednak tak rozbudowanych algorytmów. Tworzą one natomiast modele użyteczne, wartościowe i łatwe

w zarządzaniu w ramach swoich inicjatyw SI, przy czym wiele z nich nie zdaje sobie sprawy z tego, że takie zwykłe modele SI można tworzyć i wdrażać na stacjach roboczych. Wyzwaniem dla Dell jest przełamanie tych stereotypów i uświadomienie uczestnikom rynku możliwości oferowanych przez stacje robocze Dell.

Jednocześnie firma Dell musi dopilnować, aby jej stacje robocze spełniały swoje zadania i nie stały się z czasem wąskimi gardłami technologicznymi. Oznacza to szybkie wprowadzanie innowacji na bieżąco, aby nigdy nie zawieść użytkowników końcowych używających stacji roboczych do wspomnianych zastosowań SI (innymi słowy, użytkowników, którzy nie próbują uruchamiać algorytmów o wielu miliardach parametrów). Oznacza to również, że dla klientów, którzy nagle zaczynają bardzo szybko skalować swoje rozwiązania lub których algorytmy rzeczywiście stają się bardzo rozbudowane, musi być zapewniona łatwa ścieżka przejścia ze stacji roboczej na serwery Dell do obsługi sztucznej inteligencji. W tym oczywiście tkwią szanse sprzedażowe dla firmy Dell, która musi mieć odpowiednie rozwiązanie dla każdego klienta, niezależnie od wielkości realizowanej przez niego w danym momencie inicjatywy SI.

WNIOSKI

Zdaniem IDC stacje robocze są obecnie niedoceniane jako motory napędowe tworzenia i wdrażania modeli SI w różnych zastosowaniach. Zapewniają one zajmującym się SI naukowcom i deweloperom zaawansowane platformy z akceleracją GPU, które wymagają mniejszych nakładów inwestycyjnych niż serwery, mają znacznie niższe koszty operacyjne niż instancje w chmurze i dają znacznie większą swobodę eksperymentowania z modelami SI. Przedsiębiorstwa rozwijające inicjatywy SI, które nie wymagają algorytmów o miliardach parametrów (czyli większość), powinny rozważyć wyposażenie swoich zespołów SI w stacje robocze umożliwiające nieskrępowany rozwój modeli SI oraz łatwe ich wdrażanie na urządzeniach brzegowych.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com