TD SYNNEX

# MLOps dla sztucznej inteligencji

## Destination AI: Przypadek użycia

## Wprowadzenie

MLOps for AI integruje zasady DevOps ze specjalistycznymi potrzebami uczenia maszynowego, usprawniając cały cykl życia modelu - od eksperymentów i szkoleń po wdrażanie, monitorowanie i ciągłe doskonalenie. Podejście to wykorzystuje platformy orkiestracji kontenerów, takie jak Kubernetes, kontrolowane wersje notatników do eksperymentów i zautomatyzowane potoki do zarządzania odtwarzalnością, skalowalnością i zarządzaniem. MLOps umożliwia zespołom wydajną iterację modeli, płynne przenoszenie ich do produkcji i monitorowanie ich wydajności w czasie rzeczywistym. Włączając zautomatyzowane testowanie, CI/CD dla modeli i solidne rejestrowanie, organizacje zapewniają, że ich inicjatywy AI są zarówno niezawodne, jak i reagują na zmieniające się wzorce danych. Wiodące rozwiązania dostawców w tej przestrzeni obejmują AWS SageMaker, Microsoft Azure Machine Learning, Google AI Platform, IBM watsonx, Nvidia AI Enterprise Suite, Dell AI Factory i HPE Private Cloud AI, z których każdy oferuje zintegrowane środowiska do obsługi pełnego spektrum przepływów pracy uczenia maszynowego

## Punkty bólu

Opracowywanie modeli uczenia maszynowego i zarządzanie nimi wiąże się z licznymi wyzwaniami, od niespójnego śledzenia eksperymentów i kwestii kontroli wersji po złożone procedury wdrażania. Tradycyjne przepływy pracy często prowadzą do silosów między naukowcami zajmującymi się danymi a zespołami operacyjnymi, powodując opóźnienia w przenoszeniu modeli z badań do produkcji. Bez ustandaryzowanych procesów utrzymanie powtarzalności i śledzenie wydajności modelu w czasie staje się trudne. Brak zautomatyzowanych mechanizmów testowania i monitorowania może prowadzić do pogorszenia dokładności modelu, wydłużenia czasu przestojów i problemów ze zgodnością. Ponadto skalowanie modeli ML w różnych środowiskach - przy jednoczesnym zapewnieniu bezpieczeństwa danych i efektywnego wykorzystania zasobów - zwiększa złożoność, ostatecznie spowalniając innowacje i zmniejszając sprawność operacyjną.

## Korzyści z rozwiązania

Integracja praktyk MLOps z inicjatywami AI oferuje transformacyjne korzyści. Standaryzowane potoki i zautomatyzowane przepływy pracy skracają czas produkcji, wypełniając lukę między eksperymentowaniem a wdrażaniem. Ciągła integracja i dostarczanie (CI/CD) modeli umożliwiają szybką iterację i szybkie aktualizacje, zapewniając dokładność i trafność modeli. Ulepszona odtwarzalność i kontrola wersji usprawniają współpracę między zespołami, a zautomatyzowane testowanie i monitorowanie pomagają szybko zidentyfikować spadek wydajności lub anomalie. Usprawniając zarządzanie zasobami i zapewniając bezpieczne, zgodne z przepisami operacje, organizacje osiągają wyższą wydajność operacyjną, skracają czas

przestojów i utrzymują zgodność z przepisami. Ogólnie rzecz biorąc, MLOps umożliwia zespołom niezawodne skalowanie rozwiązań AI, wspieranie innowacji i dostarczanie spójnej wartości biznesowej.

## Wpływ na KPI

Kluczowe wskaźniki wydajności, które korzystają z MLOps dla AI obejmują:

**Częstotliwość wdrażania modeli**: Zwiększona częstotliwość przenoszenia modeli z fazy rozwojowej do produkcyjnej.

**Time-to-Market**: Skrócony czas od opracowania modelu do wdrożenia na żywo.

**Dokładność modelu i dryft**: Ciągłe monitorowanie zapewnia utrzymanie wysokiej wydajności modeli w czasie.

**Wydajność operacyjna**: Ulepszona przepustowość w potokach szkolenia i wdrażania modeli, redukująca ręczną interwencję.

**Częstotliwość przestojów i incydentów**: Niższa częstotliwość incydentów produkcyjnych dzięki zautomatyzowanym testom i proaktywnemu monitorowaniu.

## Wdrażanie przypadku użycia

Wdrożenie MLOps dla AI wymaga kompleksowego, wielofazowego podejścia, które integruje technologię, procesy i dostosowanie kulturowe

## Planowanie i dostosowanie do interesariuszy

**Definiowanie** celów: Jasno określ cele, takie jak skrócenie cykli wdrażania, zwiększenie dokładności modeli i poprawa współpracy między zespołami data science i IT. Odnieś je do mierzalnych wskaźników KPI.

**Zbierz wielofunkcyjny zespół**: Zaangażuj naukowców zajmujących się danymi, inżynierów ML, specjalistów DevOps, bezpieczeństwo IT i interesariuszy wykonawczych, aby zapewnić uwzględnienie wszystkich aspektów - od eksperymentów po produkcję.

**Ocena dostawców i narzędzi**: Oceń platformy MLOps, takie jak AWS SageMaker, Microsoft Azure Machine Learning, Google AI Platform, IBM watsonx, Nvidia AI Enterprise Suite, Dell AI

Factory i HPE Private Cloud AI, aby określić najlepsze dopasowanie do istniejącej infrastruktury**.**

## Alokacja zasobów i infrastruktura

**Integracja i zarządzanie** danymi: Konsolidacja zbiorów danych, dzienników eksperymentów i artefaktów modeli w scentralizowanym repozytorium z solidną kontrolą wersji i zarządzaniem dostępem.

**Planowanie budżetu**: Przydziel zasoby na licencjonowanie oprogramowania, aktualizacje infrastruktury w chmurze lub lokalnej, szkolenia i bieżące wsparcie. Oszacuj zwrot z inwestycji, porównując usprawniony czas wdrożenia i skrócony czas przestojów z kosztami inwestycji.

**Bezpieczeństwo i zgodność**: Wdrażaj silne środki bezpieczeństwa - w tym szyfrowanie, kontrolę dostępu opartą na rolach i ramy zgodności - w celu ochrony wrażliwych danych i własności intelektualnej w całym cyklu życia modelu.

## Projekt pilotażowy i dowód koncepcji

**Wybór docelowego modelu lub** przepływu pracy: Zidentyfikuj niekrytyczny model lub określony segment przepływu pracy ML, aby pilotować proces MLOps.

**Konfiguracja narzędzi i potoków**: Konfiguracja wybranej platformy, integracja notebooków do eksperymentów, orkiestracja kontenerów (np. Kubernetes), automatyczne testowanie i systemy monitorowania.

**Podstawowe wskaźniki**: Dokumentowanie bieżących wskaźników związanych z częstotliwością wdrażania, wydajnością modelu i wskaźnikami incydentów w celu ustalenia punktów odniesienia dla ulepszeń.

## Wykonanie i przyjęcie

**Wdrażanie i** szkolenie: Przeprowadzenie kompleksowego szkolenia dla wszystkich zaangażowanych zespołów, kładąc nacisk na nowe przepływy pracy, narzędzia i najlepsze praktyki w zakresie wersjonowania, wdrażania i monitorowania modeli.

**Iteracyjne pętle informacji zwrotnych**: Ustanowienie regularnych sesji przeglądowych w celu zbierania informacji zwrotnych, udoskonalania przepływów pracy i dostosowywania zautomatyzowanych potoków w oparciu o rzeczywiste dane dotyczące wydajności.

**Wdrożenie na pełną skalę**: Stopniowe rozszerzanie praktyk MLOps na dodatkowe modele i projekty, zapewniając płynną integrację z istniejącymi systemami produkcyjnymi.

## Monitorowanie i zarządzanie zmianami

**Ciągłe monitorowanie:** Wdrażaj pulpity nawigacyjne w czasie rzeczywistym, aby śledzić kluczowe wskaźniki wydajności, takie jak częstotliwość wdrażania, dokładność modeli i wskaźniki incydentów, zapewniając szybkie wykrywanie anomalii.

**Zarządzanie ryzykiem**: Opracowanie protokołów dotyczących rozwiązywania potencjalnych problemów, takich jak dryf modelu lub luki w zabezpieczeniach, zapewniających solidne plany awaryjne i ręczny nadzór w razie potrzeby.

**Skalowanie i integracja**: Zapewnienie skalowalności struktury MLOps wraz z rozwojem organizacji poprzez integrację z szerszymi systemami IT, potokami danych i infrastrukturą bezpieczeństwa.

## Konserwacja i ciągłe doskonalenie

**Aktualizacje modeli i potoków**: Regularnie aktualizuj i ponownie trenuj modele przy użyciu świeżych danych i udoskonalaj potoki wdrażania, aby dostosować się do zmieniających się wymagań i pojawiających się zagrożeń.

**Dokumentacja:** Prowadzenie szczegółowej dokumentacji przepływów pracy, najlepszych praktyk i procedur integracji w celu wspierania ciągłego doskonalenia i ułatwiania audytów.

**Zaangażowanie w społeczność i ekosystem**: Angażowanie się w fora branżowe, społeczności dostawców i grupy badawcze, aby być na bieżąco z pojawiającymi się trendami, narzędziami i najlepszymi praktykami w MLOps.

## Przegląd oferty dostawców

- AWS SageMaker: Oferuje zintegrowany pakiet do tworzenia, szkolenia i wdrażania modeli uczenia maszynowego z solidnymi możliwościami MLOps.
- Microsoft Azure Machine Learning: Zapewnia kompleksowe narzędzia do śledzenia eksperymentów, zautomatyzowane CI/CD dla modeli i bezpieczne struktury wdrażania.
- Google AI Platform: Wykorzystuje Kubernetes i zaawansowaną orkiestrację ML do skalowalnego szkolenia i wdrażania modeli.

- IBM watsonx: Zapewnia zarządzanie cyklem życia modelu AI klasy korporacyjnej i integrację generatywnych możliwości AI w ramach bezpiecznej, skalowalnej platformy.
- Nvidia AI Enterprise Suite: Oferuje kompletny zestaw narzędzi AI zoptymalizowanych pod kątem przyspieszonego szkolenia modeli, wdrażania i wnioskowania w różnych infrastrukturach.
- Dell AI Factory: Zapewnia zintegrowane rozwiązania sprzętowe i programowe w celu usprawnienia operacji ML i zapewnienia wydajnego, skalowalnego wdrażania modeli.
- HPE Private Cloud AI: Obsługuje lokalne lub hybrydowe rozwiązania MLOps z zaawansowanymi funkcjami bezpieczeństwa, zgodności i skalowalności.

## Wnioski

MLOps dla AI ma kluczowe znaczenie dla wypełnienia luki między rozwojem innowacyjnych modeli a niezawodnymi, skalowalnymi wdrożeniami produkcyjnymi. Standaryzując przepływy pracy, automatyzując testowanie i wdrażanie modeli oraz zapewniając solidne monitorowanie i zarządzanie, organizacje mogą przyspieszyć swoje inicjatywy w zakresie sztucznej inteligencji, zachowując jednocześnie wysokie standardy wydajności i bezpieczeństwa. Ustrukturyzowany proces wdrażania - obejmujący planowanie strategiczne, dostosowanie zasobów, testy pilotażowe i ciągłe doskonalenie - zapewnia, że przejście na MLOps zapewnia wymierną poprawę wydajności i jakości modeli. Wykorzystanie zintegrowanych rozwiązań dostawców, takich jak AWS SageMaker, Microsoft.