



Betere rendabiliteit en prestaties bereiken met hybride AI

AUTEUR

Steven Dickens

Technologisch hoofdadviseur | The Futurum Group

Ron Westfall

Directeur onderzoek | The Futurum Group

AUGUSTUS 2024

IN PARTNERSCHAP MET





Samenvatting

De huidige IT- en bedrijfsbesluitvormers (ITBDM's) houden toezicht op de ongekende taak om het AI-traject van hun organisatie te begeleiden. De keuzes zijn breed en de snelheid van innovatie is intens. Bij het verkennen van de beste stappen voor het AI-traject, zien we dat besluitvormers steeds vaker kiezen voor een hybride AI-model als de beste algemene koers. Waarom hybride AI?

Hybride AI is de optimalisatie of uitbreiding van AI-modellen op basis van machine learning, deep learning, neurale netwerken met menselijke expertise om domeinspecifieke AI-modellen met een grotere nauwkeurigheid te creëren of case-specifieke modellen die AI gebruiken met maximale nauwkeurigheid of voorspellingswaarschijnlijkheid. Hybride AI is het idee om vast te houden aan de zwaartekracht van gegevens en AI-oplossingen en -mogelijkheden te brengen waar je die nodig hebt - van mobiele en desktopapparaten, privé-datacenters, edge computing of zelfs de cloud. Als zodanig kunnen de use-case en de strategie voor gegevensbeheer dicteren waar en hoe AI-mogelijkheden moeten worden ingezet, met inachtneming van factoren zoals beveiliging, kosten, netwerkinfrastructuur, gegevenszwaartekracht en betrouwbaarheid.

We zien dat deze aanpak de beperkingen van methoden met één model of één infrastructuur overwint en aansluit bij de vraag van klanten naar flexibel gegevensbeheer en applicatieondersteuning. Hybride AI integreert open-source AI en closed-source mogelijkheden, zoals OpenAI's ChatGPT, waardoor de productiviteit, flexibiliteit, gegevensbescherming en AI-democratisering worden verbeterd.

Voor belangrijke besluitvormers - waaronder CIO's, CTO's, Chief Data Officers, Chief AI/Data Officers en architecten - bestaat de eerste stap uit het definiëren van problemen, het identificeren van gegevensbronnen en het selecteren van use cases, AI-toepassingen, met de juiste mix van infrastructuur en modellen van de juiste omvang. Dit proces omvat meestal strategieontwikkeling, het bouwen van een AI-stack, een AI-fabriek of een center of excellence (COE) en het implementeren van gekozen oplossingen.

Het implementeren van hybride AI vereist het evalueren, testen en schalen van opkomende technologieën met een breed partnerecosysteem om een oplossing te kunnen ontwerpen met de juiste combinatie van AI-technieken, de juiste architectuur en het juiste model voor je bedrijf.

We onderzoeken waarom het portfolio en de visie van Lenovo Smarter AI for All kunnen voldoen aan de unieke eisen van het selecteren en implementeren van hybride AI in de hele organisatie. Lenovo Smarter AI for All geeft prioriteit aan een op bedrijfsresultaten gebaseerde aanpak door de juiste mix en grootte van AI-pc's, workstations, AI-klare oplossingen, expertise en voor AI geoptimaliseerde opslag, compute,

infrastructuur, software en partnermogelijkheden te bieden met een gevalideerd partnerecosysteem. Dit omvat de mogelijkheid om op maat gemaakte, flexibele, schaalbare en energie-efficiënte AI-oplossingen te leveren die zijn gebaseerd op de industrie- en use-cases, voor persoonlijke en industriële edge-omgevingen (pc's, workstations, bedrijfslocaties), datacenters (CoLo's, on-prem) en high-performance computing (HPC)-omgevingen, allemaal met integratie/synergieën in de openbare cloud.

Samen kwamen we erachter dat Lenovo en NVIDIA AI-fabrieken bouwen die AI-geoptimaliseerde en -ondersteunde bedrijfsomgevingen voor AI bieden, waarin AI-toepassingen veilig, privé en op schaal kunnen worden ontwikkeld, geïmplementeerd en beheerd.

Het snel evoluerende hybride AI-landschap

Organisaties en hun werknemers willen graag profiteren van de transformerende kracht van generatieve AI (GenAI) en verder gaan dan de eerste experimenten en AI/ML-pilots om van hun eigen gegevens te benutten. Wijdverspreide toepassing wordt echter vaak belemmerd door uitdagingen op het gebied van wetgeving, naleving en beheer, hiaten in de expertise, datasilo's, beveiligingsproblemen, ontoereikende AI-infrastructuur en de complexiteit van het ecosysteem. Budgetbeperkingen zorgen voor nog meer complicaties. Hoewel 96% van de CIO's verwacht dat de AI-investeringen het komende jaar zullen toenemen, verwacht slechts 20% dat de totale IT-budgetten met meer dan 10% zullen groeien (CIO-studie 2024).

Hybride AI is in opkomst als een cruciaal onderdeel van zakelijke AI-strategieën, waardoor organisaties AI- en gegevensoplossingen kunnen ontwikkelen die aansluiten bij hun doelen en AI dichterbij gegevensbronnen, realtime interacties en punten voor het creëren van ervaringen brengen. Hoewel gratis AI-tools, openbare basismodellen en bestaande cloudabbonementen de eerste pilots vergemakkelijken, zitten veel organisaties vast in eindeloze proof-of-concept-cycli en escalerende kosten. Dit komt vaak door het ontbreken van een robuuste hybride AI-strategie die AI integreert in private en publieke cloudomgevingen, de afwezigheid van aangepaste AI-modellen, beperkte toegang tot hoogwaardige organisatorische gegevens en ontoereikende AI-compute en -planning.

Vanuit ons perspectief wordt het uitvoeren van GenAI en AI lokaal aan de edge, ook op apparaten, van cruciaal belang voor organisaties met verspreide gegevens en edge-locaties (bijvoorbeeld detailhandel, productie, ziekenhuizen, hybride werkplekken), organisaties die realtime analyse en toegang nodig hebben, en entiteiten met niet-verbonden computebehoeften (bijvoorbeeld externe olie- en gasactiviteiten of sterk beschermde edge-omgevingen).

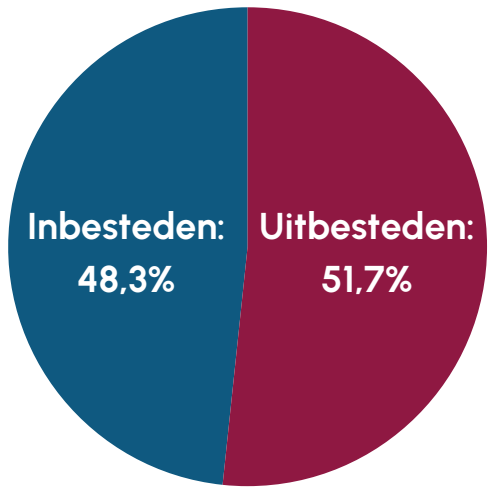
Edge-locaties en -apparaten genereren enorme hoeveelheden gegevens, waarvan de meeste niet worden geanalyseerd. AI en GenAI bieden een belangrijke kans om deze niet-gebruikte hulpbron te benutten. Met AI-toepassingen die zijn ontwikkeld voor edge-use cases, steeds efficiëntere kleinere modellen en AI-apparaten zoals AI-pc's, wordt edge computing werkelijkheid. Dit elimineert de noodzaak om alle gegevens of AI-modellen over te dragen naar openbare clouds of gecentraliseerde datacenters voor realtime analyse. Dit lost beveiligings-, latentie- en bandbreedtebeperkingen op.

Lokale GenAI-verwerking maakt gegevensanalyse mogelijk op het punt van creatie, waardoor de latentie wordt verminderd en de efficiëntie verbeterd, wat cruciaal is voor realtime toepassingen zoals productieprocessen en robotica. Daarnaast maken privacy-overwegingen lokale verwerking vaak noodzakelijk om ervoor te zorgen dat gevoelige gegevens op het apparaat blijven.

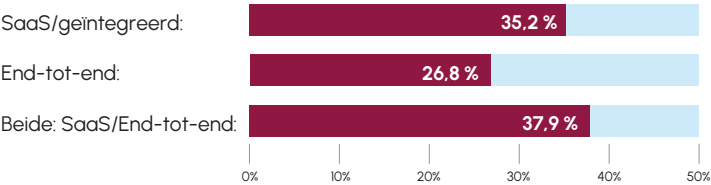
Cloud-gebaseerde grote taalmodellen (LLM's) brengen uitdagingen met zich mee die organisaties steeds meer proberen te vermijden. Deze omvatten het risico van hallucinaties - logisch inconsistente, maar plausibel klinkende resultaten - en de verspreiding van onnauwkeurigheden uit trainingsgegevens. Bovendien hebben organisaties beperkte controle over en inzicht in de infrastructuur en implementatie van cloud-managed services, waardoor beveiligingsproblemen ontstaan.

Een opmerkelijk voorbeeld deed zich voor in maart 2023 toen OpenAI's ChatGPT een storing ondervond als gevolg van een kwetsbaarheid in een open-source bibliotheek, waardoor mogelijk betalingsinformatie van klanten werd blootgesteld. Dit incident onderstreept het belang van rigoureuze LLM-beveiligingstesten en robuuste verdedigingsmechanismen voor cyberbedreigingen bij het gebruik van cloud-gebaseerde AI-services.

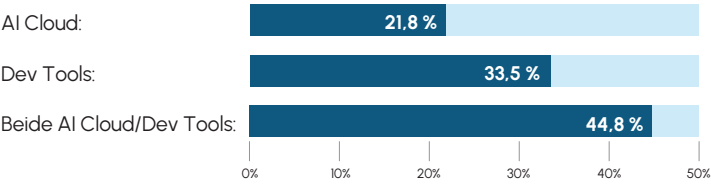
Bovendien zien we dat organisaties een evenwichtige aanpak volgen voor de manier waarop ze AI-services implementeren:



Uitbesteden



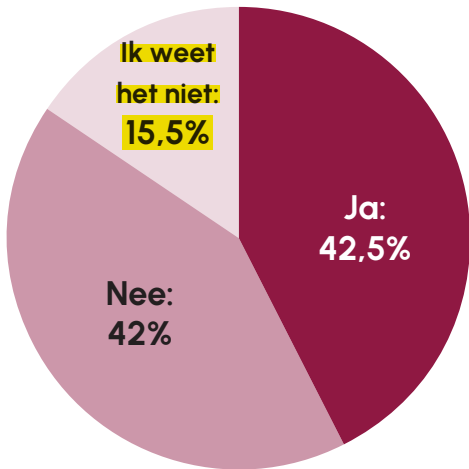
Inbesteden



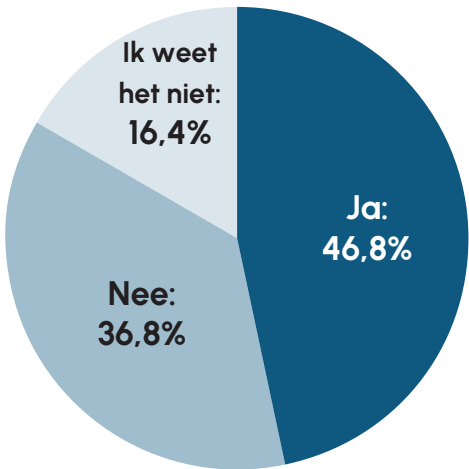
Bron: The Futurum Group

De snel groeiende interesse in hybride AI creëert een nieuwe concurrerende dynamiek, omdat organisaties steeds meer plannen verkennen om van leveranciers te veranderen of nieuwe toe te voegen (2023-24):

Uitbesteden



Inbesteden



Bron: The Futurum Group

Het succes van hybride AI hangt af van de cruciale rol van ecosystemen bij het samenstellen van AI-oplossingen die zijn afgestemd op de behoeften van de organisatie. Ecosystemen verenigen een breed scala aan belanghebbenden in de waardeketen, waardoor de naadloze integratie van AI-systemen in hybride omgevingen mogelijk wordt. Dit leidt tot de creatie van op maat gemaakte AI-oplossingen. AI-oplossingen vereisen vaak een combinatie van technologieën en expertise, afhankelijk van ecosystemen om toegang te bieden tot het brede scala aan software, toepassingen, tools en mogelijkheden om oplossingen te ontwikkelen die direct kunnen ingaan op zakelijke uitdagingen naarmate het AI-traject zich ontwikkelt.

Bovendien is de inherente flexibiliteit van hybride AI essentieel om te voldoen aan de uiteenlopende behoeften van de verschillende persona's die betrokken zijn bij AI binnen bedrijven. Ontwikkelaars van AI-toepassingen en datawetenschappers gebruiken liever bekende AI-tools en -frameworks, ongeacht hun distributie. Wat betreft de edge verwachten en vereisen sectoren, zoals winkels en productiebedrijven, reacties in realtime van AI-systemen, naast de mogelijkheid om gegevens snel te verwerken.

Het effectieve gebruik van hybride AI vereist echter een zorgvuldige planning en uitvoering, rekening houdend met technologie, mensen, processen en verantwoorde AI-benaderingen. Bedrijven worstelen met silo-initiatieven, verkeerd afgestemde keuzes en budgetten voor AI-oplossingen, terwijl IT-teams moeite hebben om de infrastructuur te moderniseren en te voldoen aan diverse AI-computebehoeften, naleving, beveiligingsnormen en de noodzaak om een hybride operatie uit te voeren. Naarmate gegevensbeheer en het gebruik van de openbare cloud ongecontroleerd worden, komen beveiliging, beheer en escalerende cloudkosten naar voren als dringende zorgen. Deze uitdagingen moeten worden afgewogen tegen de vaardigheden en het change management die nodig zijn om zinvolle bedrijfsresultaten te behalen met AI.

Datawetenschappers en ML-ingenieurs zouden bijvoorbeeld niet het grootste deel van hun tijd moeten besteden aan gegevenskwaliteit, regelgeving en naleving. Bedrijven moeten IT&O-teams betrekken bij AI-gerelateerde IT-beslissingen. Op dezelfde manier kan een bottom-up benadering door IT-teams om AI te introduceren in de onderneming leiden tot over- of ondervoorzieningen zonder een grondig begrip van de zakelijke behoeften en use cases.



De gezamenlijk ontwikkelde portfolio van Lenovo en NVIDIA

De groei van de AI-industrie wordt gedreven door meerdere factoren, waarbij de toenemende vraag naar automatisering en optimalisatie in alle sectoren van het grootste belang is. Zoals beschreven in ons artikel, benutten bedrijven de vooruitgang in compute en cloud-infrastructuur om efficiëntere AI-toepassingen te implementeren. Deze verschuiving heeft geleid tot een toename van het aantal aanvragen voor AI-oplossingen op maat, waaronder advies-, ontwerp-, implementatie- en onderhoudsdiensten die hen kunnen helpen dat proces te vereenvoudigen.

De portfolio Lenovo Smarter AI for All komt tegemoet aan deze veranderende behoeften met AI-apparaten, slimme infrastructuur en robuuste cloudoplossingen, waardoor naadloze AI-integratie op verschillende platforms mogelijk wordt. Het strategische partnerschap tussen Lenovo en NVIDIA versnelt innovatie in de AI-sector. Hun gezamenlijk ontwikkelde portfolio vereenvoudigt de invoering van AI, levert snellere resultaten en verbetert de prestaties en efficiëntie in hybride AI-omgevingen, wat de groei van de sector verder stimuleert.

Belangrijke informatie over dit rapport

MEDEWERKERS

Steven Dickens

Technologisch hoofdadviser | The Futurum Group

Ron Westfall

Directeur onderzoek | The Futurum Group

UITGEVER

Daniel Newman

CEO | The Futurum Group

VRAGEN

Neem contact met ons op als je dit rapport wilt bespreken. The Futurum Group zal snel antwoorden.

CITATEN

Dit artikel kan worden geciteerd door de geaccrediteerde pers en analisten, maar moet in de context worden geciteerd, met vermelding van de naam van de auteur, de titel van de auteur en 'The Futurum Group'. Niet-pers- en niet-analisten moeten vooraf schriftelijke toestemming krijgen van The Futurum Group voor eventuele citaten

LICENTIES

Dit document, inclusief eventueel ondersteunend materiaal, is eigendom van The Futurum Group. Deze publicatie mag niet worden gereproduceerd, verspreid of gedeeld in welke vorm dan ook zonder voorafgaande schriftelijke toestemming van The Futurum Group.

OPENBAARMAKING

De Futurum Group levert onderzoek, analyse, advies en consultancy aan veel hightechbedrijven, waaronder die welke in dit artikel worden genoemd. Er zijn geen werknemers van het bedrijf die aandelen bezitten in de bedrijven die in dit document worden genoemd.



OVER LENOVO EN NVIDIA

Lenovo brengt het nieuwe tijdperk van AI-aangedreven innovatie voor iedereen. Onze full-stack portfolio levert krachtige, flexibele en verantwoorde AI-oplossingen om sectoren te transformeren en individuen te versterken. We creëren een toekomst van slimmere AI voor iedereen. Bij Lenovo geloven we dat de toekomst van AI bestaat uit het naast elkaar bestaan van publieke en zakelijke AI. Lenovo brengt AI naar jou en je gegevens.

In samenwerking met NVIDIA worden hybride AI-oplossingen speciaal ontwikkeld door technische samenwerking om AI efficiënt naar klantgegevens te brengen, waar en wanneer gebruikers deze het meest nodig hebben. Dit bevordert de visie van Lenovo om AI voor iedereen mogelijk te maken en time-to-market ondersteuning te bieden van baanbrekende technologieën en architectuur voor de volgende generatie generatieve AI. Hybride oplossingen van Lenovo, die al zijn geoptimaliseerd voor het uitvoeren van NVIDIA AI Enterprise-software voor veilige, ondersteunde en stabiele productie-AI, bieden ontwikkelaars ook toegang tot NVIDIA-microservices, waaronder NVIDIA NIM's en NeMo Retriever.



OVER THE FUTURUM GROUP

[The Futurum Group](#) is een onafhankelijk onderzoeks-, analyse- en adviesbureau dat zich richt op digitale innovatie en marktverstorende technologieën en trends. Elke dag helpen onze analisten, onderzoekers en adviseurs bedrijfsleiders uit de hele wereld bij het anticiperen op tektonische verschuivingen in hun sectoren en het benutten van ingrijpende innovatie om een concurrentievoordeel in hun markten te behalen of te behouden.



CONTACTGEGEVENS

The Futurum Group LLC | futurumgroup.com | (833) 722-5337 |