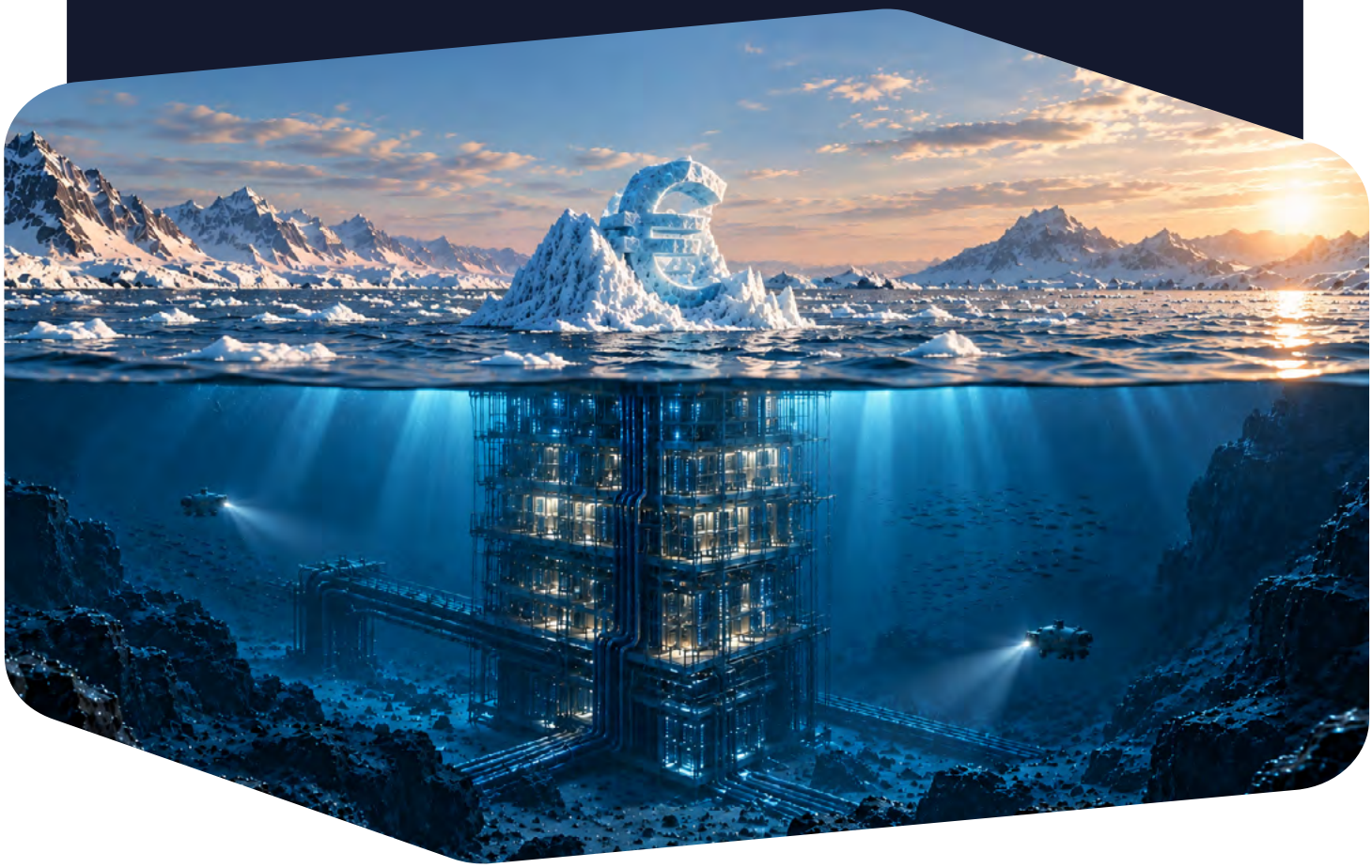
PQR

Wat AI je werkelijk kost

Zonder de juiste infrastructuur wordt AI al snel AI AI AI op je factuur.



HPE

De echte kosten van AI blijven nog te vaak onzichtbaar. Op de factuur zie je hoeveel tokens je verbruikt en wat je daarvoor betaalt. Maar tokens tonen alleen het resultaat, niet de oorzaak van jouw verbruik. Om die te begrijpen, moet je kijken naar de infrastructuur waarop jouw AI draait. Elke workload stelt andere eisen aan rekenkracht en capaciteit. Waar je die workload laat draaien, bepaalt mede hoe efficiënt en tegen welke kosten je aan die eisen kunt voldoen. Tegelijk heeft die keuze gevolgen voor waar jouw data wordt verwerkt, welke partijen erbij betrokken zijn en hoeveel controle je zelf houdt.

Die keuzes worden belangrijker naarmate je AI-gebruik groeit. Meer toepassingen, grotere volumes en intensiever gebruik zorgen ervoor dat de kosten verder oplopen. En als de infrastructuur daar niet op is afgestemd, groeien bestaande inefficiënties gewoon mee. Wie vandaag niet weet waar die kosten ontstaan, loopt het risico er straks op grotere schaal voor te betalen. De echte rekensom begint daarom bij wat er onder jouw AI draait.

Wat je leert

- Is jouw IT-infrastructuur klaar voor groeiend AI-gebruik?
- Wat jouw AI-factuur niet vertelt
- Waarom AI-kosten naar verwachting verder zullen stijgen
- Welke ontwikkelingen de eisen aan jouw AI-omgeving veranderen
- Welke risico's een inefficiënte infrastructuur met zich meebrengt voor kosten, soevereiniteit en beheer
- Hoe je bepaalt welke infrastructuur bij welke AI-workload past
- Hoe je jouw AI-omgeving efficiënter inricht én beheert

Is jouw IT-infrastructuur klaar voor AI?

Veel banken, verzekeraars en vermogensbeheerders bouwen vandaag verder op een IT-omgeving die jaren geleden is ingericht. Nieuwe AI-toepassingen vragen daar steeds meer rekenkracht en capaciteit van. Bij een testfase is dat vaak geen probleem, maar zodra een toepassing breder wordt gebruikt of grotere volumes verwerkt, groeit ook de impact van een inefficiënte inrichting mee. Wat op kleine schaal misschien weinig kost, kan bij opschaling een structurele kostenpost worden die met elk nieuw AI-project verder groeit.

Neem bijvoorbeeld een AI-toepassing die documenten bij de onboarding van nieuwe klanten verwerkt. In een testfase blijft het gebruik beperkt.

Zodra dezelfde toepassing alle nieuwe klantdossiers verwerkt, gaan de keuzes erachter zwaarder doorwegen: welk model wordt gebruikt, hoeveel capaciteit ervoor beschikbaar is en hoeveel data bij elke verwerking wordt opgehaald. Een inrichting die op kleine schaal weinig verschil maakt, kan bij grotere volumes een merkbare impact op de kosten hebben.

Tegelijk wordt ook een tweede vraag belangrijker: waar draait die toepassing en wat gebeurt er met de data die ze verwerkt? Zeker wanneer het om klant-, transactie- of andere gevoelige gegevens gaat, wil je niet alleen weten wat een workload kost, maar ook waar de data terecht-komt en hoeveel controle je daar zelf over houdt.





Wat je AI-factuur niet vertelt

Op het eerste gezicht lijken de kosten van AI behoorlijk transparant. Je ziet hoeveel tokens je verbruikt en hoeveel je daarvoor betaalt. Tokens zijn de eenheden waarin een AI-model informatie verwerkt en genereert en waarop het gebruik doorgaans wordt afgerekend. Stijgt het gebruik, dan stijgt de factuur. Dat is logisch. Maar die factuur vertelt je niet waarom de ene AI-toepassing duurder is dan de andere. Of waarom de kosten sneller stijgen dan het gebruik. En al helemaal niet waar je moet ingrijpen als je die kosten naar beneden wilt krijgen.

Vergelijk het met je energiefactuur. Je kent jouw verbruik in kilowattuur en de prijs die je per kilowattuur betaalt. Toch weet je daarmee nog niet waarom uw energierekening zo hoog is. Misschien is je gebouw slecht geïsoleerd of verbruikt een installatie meer dan nodig. Met AI is dat vergelijkbaar. De kosten die je op de factuur ziet, wordt mede bepaald door het model dat je gebruikt, de infrastructuur waarop het draait, de hoeveelheid data die wordt verwerkt en de manier waarop de omgeving wordt beheerd.

Net daar ontstaat een belangrijke vraag voor IT en finance: kan je verklaren waarom je AI-kosten stijgen en waar je kunt ingrijpen? Uit de CFO Benchmark van PQR en The Blue Hour blijkt hoe cruciaal die vraag is. 86% van de ondervraagde CFO's werkt actief aan AI-initiatieven, terwijl geen van hen aangeeft al over een duidelijk gedefinieerde AI-strategie te beschikken.[1] De onderzoekers stellen bovendien vast dat organisaties de ROI van hun cloudactiviteiten niet structureel monitoren. Wie die kosten wil beheersen, moet dus eerst kunnen verklaren waar ze ontstaan voordat groeiend AI-gebruik bestaande inefficiënties verder uitvergroot.

Drie ontwikkelingen die de eisen aan jouw AI-omgeving veranderen

Naarmate AI structureel wordt ingezet, veranderen ook de eisen aan de omgeving eronder. Drie ontwikkelingen spelen daarbij een belangrijke rol:



Soevereiniteit komt hoger op de agenda

Waar je AI-workloads draaien, bepaalt mede waar data wordt verwerkt, wie er toegang toe kan hebben en van welke externe partijen je afhankelijk bent. Die vragen zijn niet nieuw, maar de druk om er een duidelijk antwoord op te hebben is de voorbije jaren wel toegenomen.

DORA is daar een duidelijk voorbeeld van. Sinds 2025 gelden strengere eisen rond digitale weerbaarheid en het beheersen van risico's bij externe ICT-leveranciers. Ook het onderzoek van de EDPS naar het gebruik van Microsoft 365 door de Europese Commissie zette vragen rond internationale gegevensdoorgifte en controle over persoonsgegevens op scherp.

Daarnaast waarschuwden DNB en AFM in 2025 voor de groeiende afhankelijkheid van een klein aantal niet-Europese IT-leveranciers. Ze roepen financiële instellingen onder meer op zich voor te bereiden op ontwrichtende scenario's en hun weerbaarheid te testen.[2]

Het zijn maar enkele voorbeelden van dezelfde ontwikkeling: waar data en workloads draaien, van welke leveranciers je afhankelijk bent en hoeveel controle je zelf houdt, wordt steeds belangrijker.





AI wordt een structurele en steeds hogere kostenpost

Zodra AI-toepassingen dagelijks door medewerkers of klanten worden gebruikt, worden ook de kosten structureel. En alles wijst erop dat die verder zullen oplopen. Organisaties zetten meer AI-toepassingen in, bestaande toepassingen worden intensiever gebruikt en de hoeveelheid data die ermee wordt verwerkt groeit. Daardoor neemt niet alleen het totale AI-verbruik toe, maar gaan ook de keuzes achter elke workload financieel zwaarder doorwegen.

Een infrastructuur die vandaag nog voldoende efficiënt lijkt, kan daardoor bij groeiend gebruik steeds duurder uitvallen. Ongebruikte capaciteit, een te zwaar model of een workload die op een duurdere omgeving draait dan nodig, wordt immers telkens opnieuw aangerekend. Hoe groter het gebruik, hoe groter de financiële impact van die inefficiënties. Dan volstaat het niet meer om te weten dát een toepassing werkt. Je wil ook weten waarom de kosten stijgen en waar je kunt ingrijpen voordat ze verder oplopen.

Voor kleinere financiële instellingen komt daar nog een uitdaging bij. De IT-capaciteit is beperkt en een groot deel ervan gaat al naar het draaiende houden van de bestaande omgeving. Tegelijk vragen AI, data en nieuwe regelgeving rond AI steeds meer aandacht. Meer AI betekent dus niet alleen meer infrastructuurkosten, maar ook meer druk op de mensen die de omgeving moeten beheren.



Jouw infrastructuur krijgt een andere rol

Veel organisaties in de financiële sector werken met een IT-omgeving die jaren geleden is ingericht en sindsdien vooral is onderhouden. Zolang die omgeving stabiel draait, is er weinig reden om ze grondig te herbekijken.

Met AI komen daar andere soorten workloads bij. Denk aan fraudedetectie die grote hoeveelheden transactiegegevens analyseert, document-verwerking bij onboarding en KYC, chatbots voor klantenservice of toepassingen die vermogens-beheerders ondersteunen bij advies. Elke toepassing stelt andere eisen aan rekenkracht, data en beschikbaarheid. En naarmate ze breder worden ingezet, wegen die eisen steeds zwaarder op de infrastructuur eronder.

Een omgeving die technisch nog prima functioneert, kan daardoor minder efficiënt worden voor de AI-workloads die erop terechtkomen. De vraag wordt dan welke infrastructuur elke workload nodig heeft en waar die het meest efficiënt kan draaien.

Wat gebeurt er als je infrastructuur niet mee evolueert?

Minder grip op waar je data terechtkomt

Als je onvoldoende overzicht hebt over je workloads, data en leveranciers, wordt het moeilijk om de controle die van je wordt verwacht ook daadwerkelijk aan te tonen. Vragen van een toezichthouder beantwoorden, afhankelijkheden beoordelen of een overstap naar een andere leverancier voorbereiden, kost dan meer tijd en wordt complexer. Je loopt bovendien het risico dat afhankelijkheden pas zichtbaar worden wanneer een leverancier uitvalt of je ervan weg wilt.

Kosten die blijven oplopen

De bestaande omgeving laten zoals ze is, betekent dat ook inefficiënties blijven bestaan. Je behoudt bijvoorbeeld capaciteit die je nauwelijks gebruikt of laat workloads draaien op infrastructuur die zwaarder of duurder is dan nodig.

Zolang het gebruik beperkt blijft, hoeft dat financieel niet meteen op te vallen. Maar naarmate AI breder wordt ingezet, worden ook die inefficiënties steeds duurder. Zonder bijsturing groeit dus niet alleen je AI-gebruik, maar betaal je ook steeds vaker en op grotere schaal voor capaciteit en infrastructuur die niet optimaal worden benut. De kosten groeien mee, zonder dat duidelijk is welk deel van die stijging werkelijk nodig is.

Meer beheer voor een IT-team dat al vol zit

Elke extra omgeving moet worden gemonitord, beveiligd, bijgewerkt en ondersteund. Voor kleinere organisaties waar IT-teams vaak al verantwoordelijk zijn voor een brede bestaande omgeving, komt dat werk boven op de dagelijkse beheerlast. Ook het aantal leveranciers telt mee. Elke extra partij betekent een contract, een aanspreekpunt en een facturatiestroom die beheerd moeten worden. Een infrastructuur die op papier voordelig lijkt, kan daardoor alsnog duurder uitvallen wanneer je ook de interne tijd en expertise meetelt die nodig zijn om ze draaiende te houden.

Hoeveel grip heb je vandaag?

Kosten, soevereiniteit en beheer hangen nauw samen met de manier waarop je AI-omgeving is ingericht. Maar hoeveel zicht heb je daar vandaag werkelijk op? Deze vragen geven een eerste indicatie:

- Kan je verklaren waar je AI-kosten vandaan komen en waarom ze stijgen?
- Kan je voor jouw belangrijkste AI-workloads aantonen waar ze draaien, waar de data wordt verwerkt en onder welke regelgeving?
- Weet je wat er gebeurt wanneer een kritieke leverancier of clouddienst tijdelijk niet beschikbaar is?

Kan je een of meer van deze vragen niet eenduidig beantwoorden? Dan is het verstandig om jouw AI-omgeving breder te bekijken: van verbruik en infrastructuur tot datacontrole, leveranciers en beheer.

Hoe infrastructuur jouw kosten én soevereiniteit bepaalt

Niet elke AI-workload stelt dezelfde eisen aan je infrastructuur. De ene vraagt veel en voorspelbare rekenkracht, de andere kent sterke pieken. Sommige toepassingen verwerken gevoelige klant- of transactiedata, terwijl dat bij andere veel minder speelt. Een fraudedetectiemodel dat continu transacties analyseert, stelt bijvoorbeeld andere eisen dan een chatbot die vooral piekt wanneer veel klanten tegelijk contact opnemen. De juiste infrastructuurkeuze verschilt daardoor per workload.

Niet elke workload hoort op dezelfde plek

Niet elke workload vraagt dezelfde infrastructuur. Om te bepalen waar een toepassing het best draait, kijk je daarom niet naar één criterium, maar naar drie aspecten die elkaar beïnvloeden:

Wat moet je bekijken?

Kosten

Hoeveel capaciteit vraagt de workload? Hoe voorspelbaar is het gebruik? Betaal je voor capaciteit die je niet gebruikt?

Soevereiniteit

Welke data wordt verwerkt? Waar mag die worden verwerkt? Wie heeft er toegang toe? Van welke externe partijen ben je afhankelijk?

Beheer

Wat vraagt de omgeving van je IT-team? Welke kennis is nodig? Hoeveel leveranciers en omgevingen moet je beheren?

Wat betekent dit voor de keuze van je IT-infrastructuur?

Bij wisselend gebruik kan de publieke cloud financieel interessant zijn doordat je capaciteit kan meeschalen. Bij een hoge, voorspelbare belasting kan private infrastructuur voordeliger uitvallen.

Voor workloads met gevoelige data kan een private of afgeschermde omgeving meer controle bieden over waar data wordt verwerkt, wie er toegang toe heeft en van welke leveranciers je afhankelijk bent.

Hoe meer omgevingen en leveranciers je combineert, hoe meer monitoring, onderhoud en afstemming jouw IT-team moet opvangen.

Die drie kolommen hierboven kan je niet los van elkaar bekijken. De goedkoopste omgeving is niet noodzakelijk de beste keuze wanneer ze meer afhankelijkheid of beheer met zich meebrengt. Omgekeerd hoeft meer controle niet automatisch duurder te zijn. De juiste balans verschilt per workload.



Beweeg naar een omgeving die je kunt beheersen

Voor banken, verzekeraars en vermogensbeheerders gaat een efficiënte AI-omgeving niet alleen over meer rekenkracht. Je wil kunnen sturen op wat een workload verbruikt, waar ze draait en hoeveel capaciteit je ervoor beschikbaar houdt. Dat wordt alleen belangrijker naarmate het aantal AI-workloads en het gebruik ervan groeit. Tegelijk moet de omgeving beheersbaar blijven voor een IT-team dat vaak al veel verantwoordelijkheden draagt. **Daarvoor zijn drie zaken belangrijk:**

01

AI-modellen efficiënt laten draaien

Naarmate meer AI-toepassingen in gebruik komen, wordt het belangrijker om modellen efficiënt uit te rollen, te schalen en te monitoren. Zo kun je de capaciteit beter afstemmen op de werkelijke belasting en krijgt IT meer zicht op prestaties en gebruik. HPE Machine Learning Inference Software ondersteunt het uitrollen, schalen en monitoren van AI- en ML-modellen over cloud-, hybride en on-premises omgevingen.

02

Zicht houden op workloads over verschillende omgevingen

Wanneer workloads verspreid draaien over public cloud en private infrastructuur, wil je niet voor elke omgeving afzonderlijk uitzoeken wat er gebeurt. Je hebt overzicht nodig om kosten, gebruik en plaatsing van workloads naast elkaar te kunnen leggen. HPE Morpheus brengt het beheer van die verschillende omgevingen samen. Onze experts bepalen samen met jou waar workloads het best passen op basis van de criteria die voor jouw organisatie belangrijk zijn.

03

Capaciteit voorzien zonder alles vooraf vast te leggen

Een eigen omgeving hoeft niet automatisch te betekenen dat je alle capaciteit vooraf moet aankopen voor het hoogste verwachte gebruik. Met HPE GreenLake betaal je voor de capaciteit die je daadwerkelijk gebruikt. Zo stem je jouw infrastructuur af op je verbruik en hoeft meer controle over je omgeving niet automatisch een grote investering vooraf te betekenen.

Van inzicht naar een beheerde AI-omgeving

01 Assessment

PQR brengt je huidige AI-omgeving, workloads en kosten in kaart en bepaalt waar optimalisatie mogelijk is.

02 Implementatie

Na de assessment kan PQR de gekozen aanpassingen implementeren en de omgeving opnieuw inrichten.

03 Beheer

Als managed service partner kan PQR de omgeving vervolgens monitoren, onderhouden en ondersteunen, zodat het dagelijkse beheer niet boven op het werk van je eigen IT-team komt.



Je krijgt zo niet alleen advies over je infrastructuur, maar een partner die ook verantwoordelijkheid neemt voor de werking en continuïteit ervan. Als partner van HPE combineren we daarvoor de juiste technologie met de expertise om ze in te richten en te beheren.

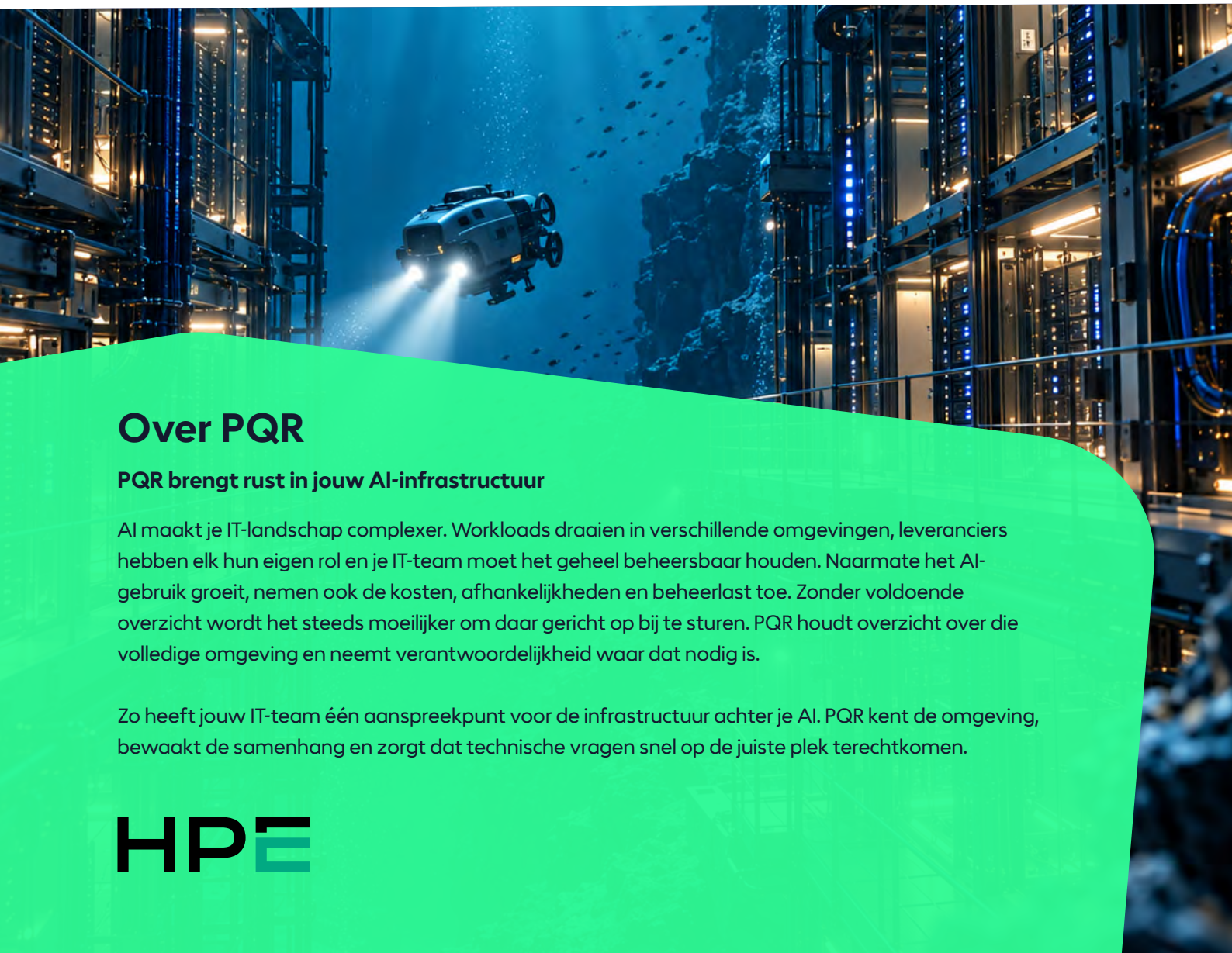
Inzicht krijgen in je huidige AI-omgeving?

Ontdek waar je AI-kosten ontstaan en welke infrastructuur het best past bij elke workload. Een assessment kan vervolgens concreet maken waar je kunt bijsturen om groeiende kosten te beheersen, zonder grip op beheer en datacontrole te verliezen.

Alles begint met een goed gesprek en een kop koffie. Bekijk wat PQR al doet voor organisaties in finance en verzekeringen, of plan een afspraak met onze specialist in IT voor de financiële sector.

Ontdek PQR voor Finance & Insurance

Bronvermelding:
[1]PQR, HPE & The Blue Hour. (2024). CFO Benchmark 2024: De impact van digitale transformatie op de CFO. PQR B.V.
[2]AFM & DNB, Digitale afhankelijkheid in de financiële sector: risico's, weerbaarheid en Europese autonomie, 20 oktober 2025



Over PQR

PQR brengt rust in jouw AI-infrastructuur

AI maakt je IT-landschap complexer. Workloads draaien in verschillende omgevingen, leveranciers hebben elk hun eigen rol en je IT-team moet het geheel beheersbaar houden. Naarmate het AI-gebruik groeit, nemen ook de kosten, afhankelijkheden en beheerlast toe. Zonder voldoende overzicht wordt het steeds moeilijker om daar gericht op bij te sturen. PQR houdt overzicht over die volledige omgeving en neemt verantwoordelijkheid waar dat nodig is.

Zo heeft jouw IT-team één aanspreekpunt voor de infrastructuur achter je AI. PQR kent de omgeving, bewaakt de samenhang en zorgt dat technische vragen snel op de juiste plek terechtkomen.

HPE