

NetApp

Make every token earn its place

The Economics of Intelligence

Ryan Barkwell
Senior AI Solution Sales Specialist
October 2026

© 2026 NetApp, Inc. All rights reserved.

ACROSS SEVERAL ERAS IN THE AGE OF DATA, NETAPP HAS LED IN DATA INFRASTRUCTURE INNOVATION



2002

Data Silos & Unification

NetApp becomes the first vendor to **unify file & block workloads, and structured & unstructured data**

Hybrid Cloud

NetApp creates the first **data fabric strategy** that eliminates silos & provides unified control across any environment

Hybrid Multiclouds

NetApp becomes the ONLY vendor to introduce **cloud ops and data services** as key data infrastructure pillars in addition to being the **only vendor natively embedded in all major clouds**

Today

Intelligence

NetApp delivers **silo-free infrastructure,** then harnesses **observability and AI** to enable best data management everywhere

NetApp

Building the world's best AI solutions

Partners since 2018

>1500 joint customers

Worldwide partner ecosystem

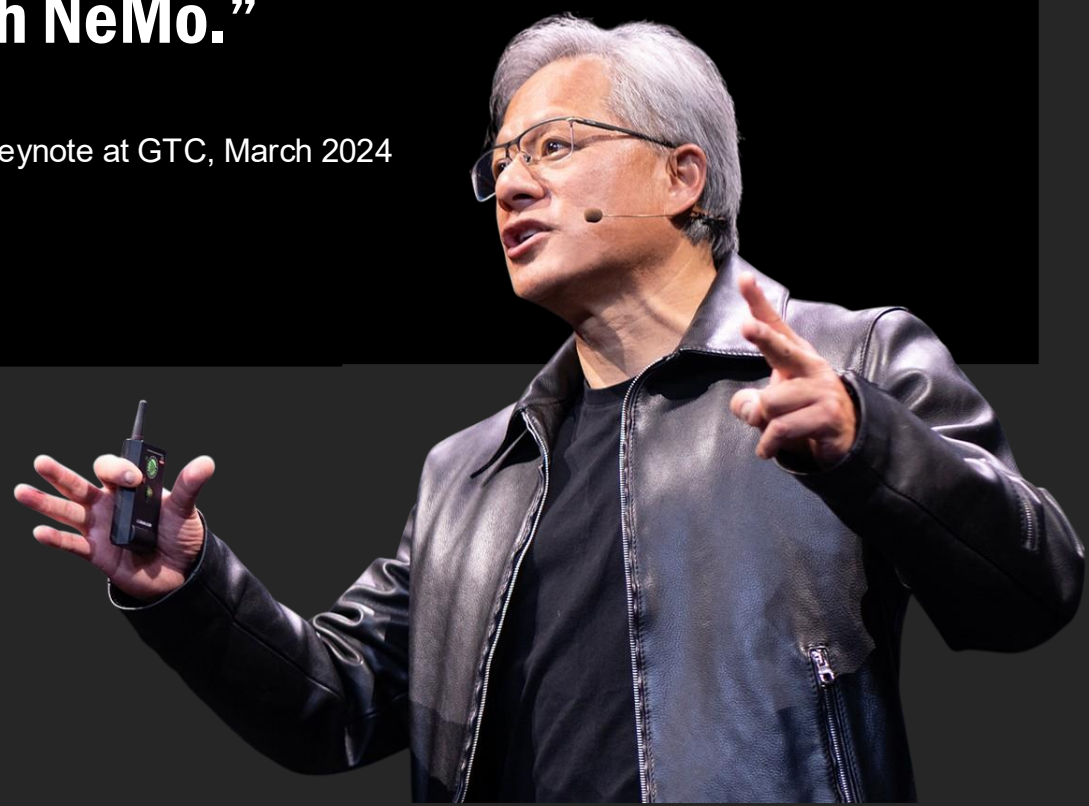
Joint solution development:

- Certified reference architectures for DGX SuperPOD
- NetApp® AIPod™ and FlexPod™ reference architectures
- NCP certification
- NVAIE in hyperscalers with NetApp
- NeMo/NIMS in ONTAP (soon)
- And more….

NetApp | NVIDIA

"Half of the world's files run on NetApp, and now you can securely talk to them with GenAI with NeMo."

—Jensen Huang, Keynote at GTC, March 2024



NVIDIA DGX SuperPOD + NetApp AFF/AFX

ONTAP for the highest-performance AI workloads



Certified for NVIDIA DGX SuperPOD AI infrastructure

Press release

© 2026 NetApp, Inc. All rights reserved.

The Agentic Multiplier

NetApp

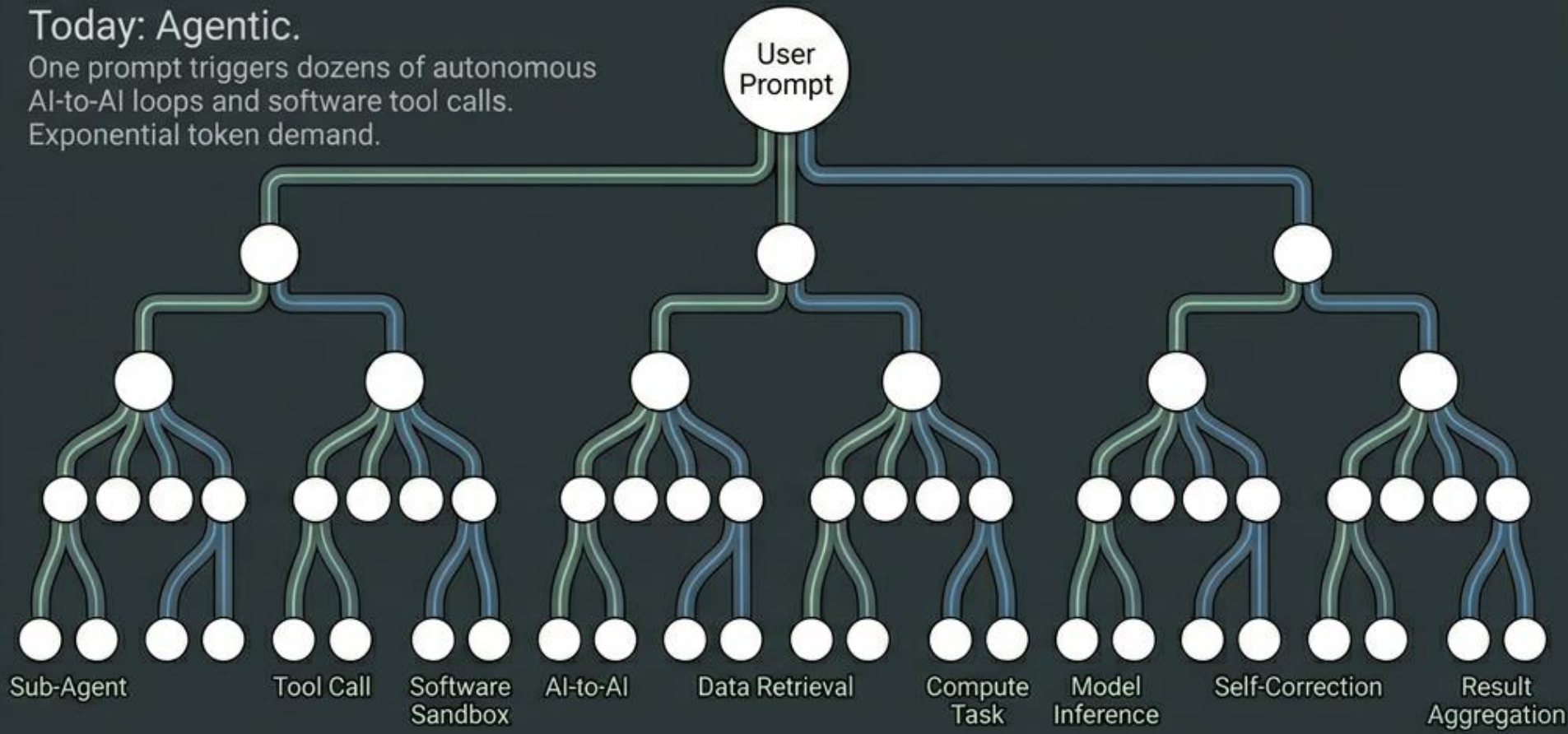
Yesterday: Conversational.

Human takes a turn,
AI takes a turn.
Linear token demand.



Today: Agentic.

One prompt triggers dozens of autonomous AI-to-AI loops and software tool calls.
Exponential token demand.



Agentic workflows demand ultra-low latency data access. Every millisecond of storage delay multiplies across dozens of agentic turns, destroying time-to-value.

Not all tokens are created equal.

NetApp

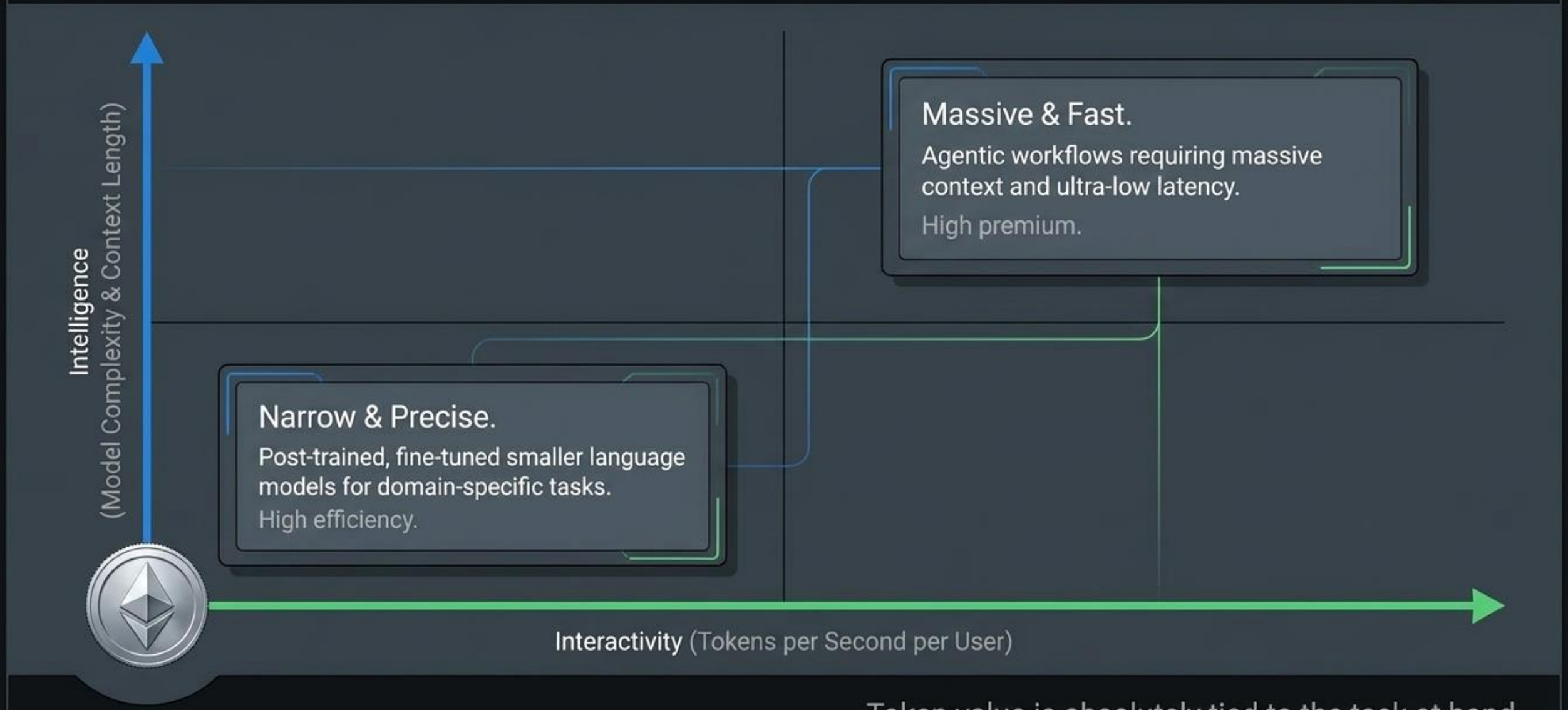
Interactivity
Tokens per second per user
(The speed of delivery)

Token Value = Intelligence × Interactivity

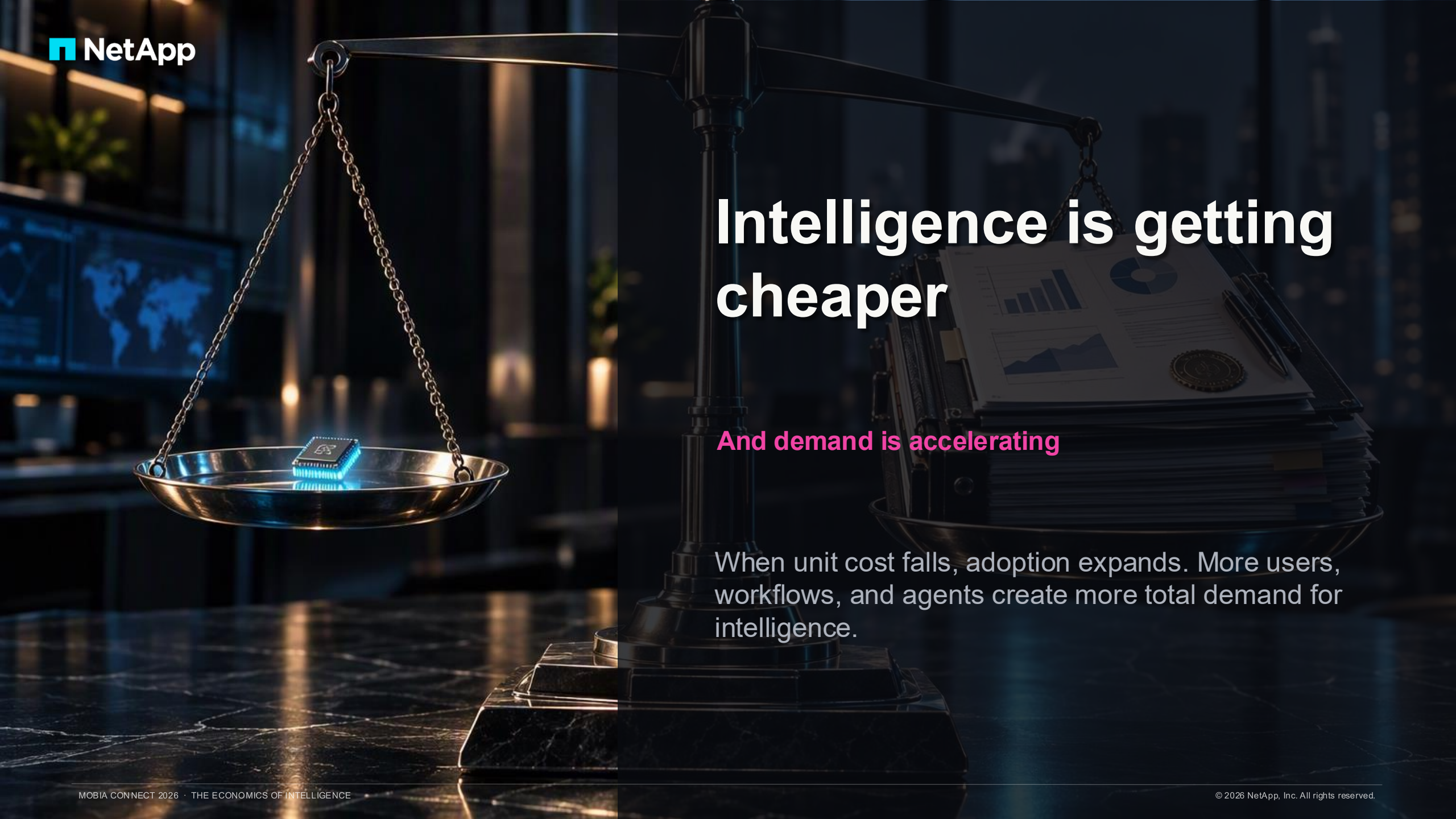
Intelligence
The embedded complexity of the model
(The raw brainpower)

Not all tokens are created equal.

NetApp



Token value is absolutely tied to the task at hand.



Intelligence is getting cheaper

And demand is accelerating

When unit cost falls, adoption expands. More users, workflows, and agents create more total demand for intelligence.

© 2026 NetApp, Inc. All rights reserved.

Cheap tokens do not mean cheap outcomes

Cost moves into the system

Model calls are only one line item. Retrieval, context, tools, validation, latency, and human rework all sit on the same bill.

© 2026 NetApp, Inc. All rights reserved.

NetApp

THE PARADOX

Cheaper tokens can still create bigger bills

UNIT COST ↓
TOTAL DEMAND ↑

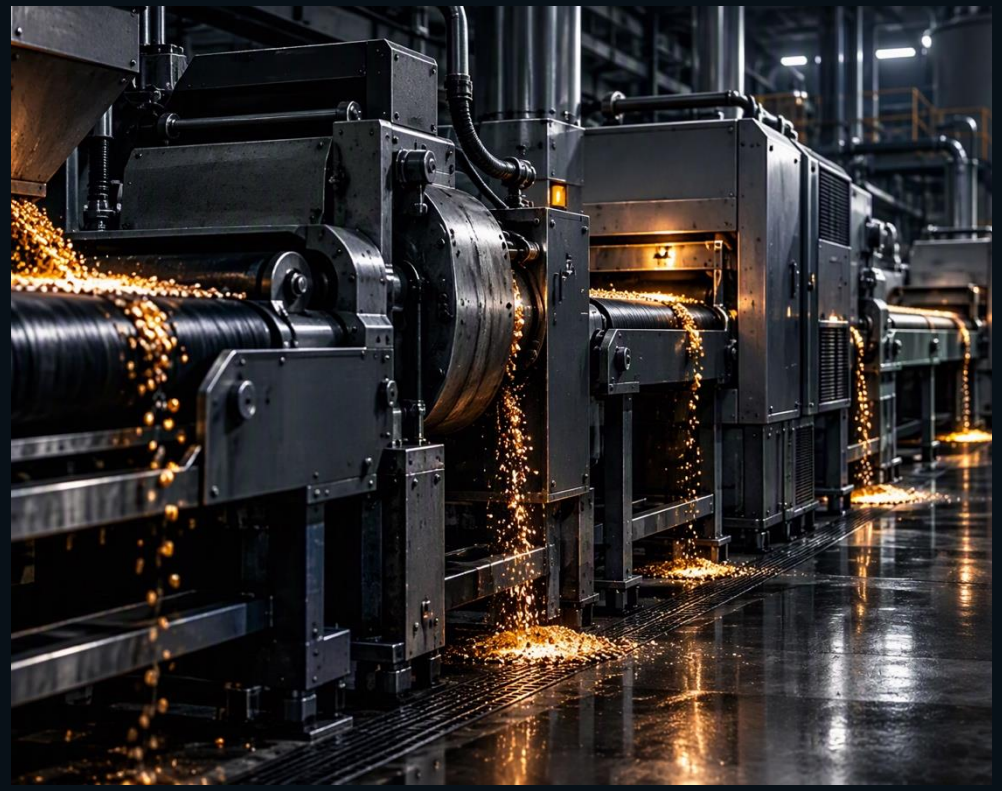
BUSINESS IMPACT Efficiency lowers the barrier to consumption; it does not cap demand.

NetApp

THE HIDDEN VILLAIN

Waste lives between data and decision

THE MODEL IS ONLY ONE COST CENTRE.



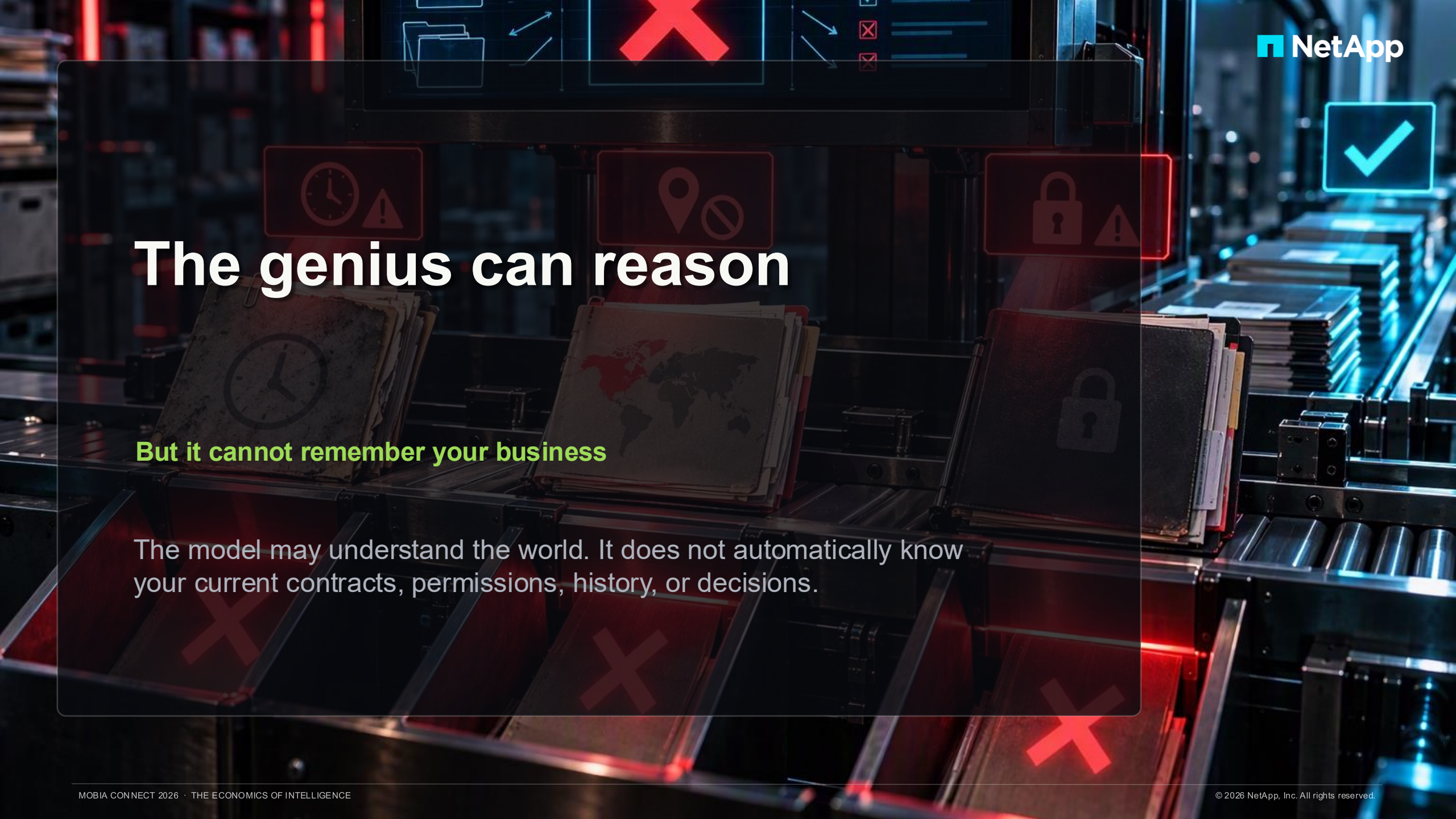
BUSINESS IMPACT Most avoidable AI spend is created between model calls.

Meet the genius

Extraordinary reasoning. No company memory.



© 2026 NetApp, Inc. All rights reserved.



The genius can reason

But it cannot remember your business

The model may understand the world. It does not automatically know your current contracts, permissions, history, or decisions.

© 2026 NetApp, Inc. All rights reserved.

The CEO asks one question

Should we modify the Q3 supply chain strategy?



© 2026 NetApp, Inc. All rights reserved.



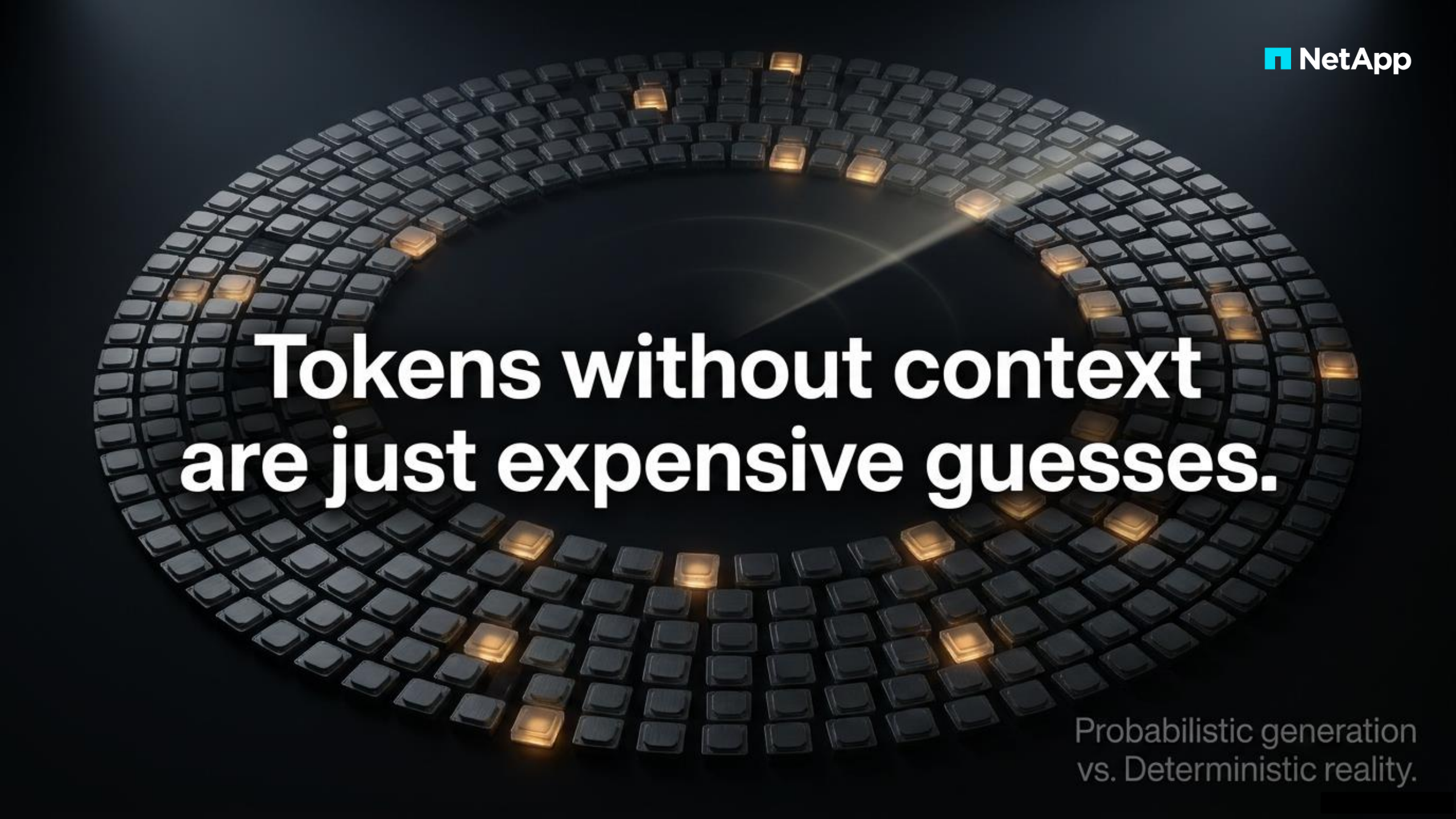
NetApp

Many generic questions come back

Which suppliers? Which policies? Which history? ????

The genius needs institutional memory before it can create a responsible enterprise recommendation.

© 2026 NetApp, Inc. All rights reserved.


NetApp
Tokens without context are just expensive guesses.
Probabilistic generation vs. Deterministic reality.

If you're tempted to push through this anyway

UnitedHealth uses faulty AI to deny elderly patients medically necessary coverage, lawsuit claims

Google's AI Overviews Caught Giving Dangerous "Health" Advice

Air Canada found liable for chatbot's bad advice on plane tickets

Airline's claim that online helper was responsible for its own actions was 'remarkable':

Vibe coding service Replit deleted user's production database, faked data, told fibs galore

Deloitte refunds Aussie gov after AI fabrications slip into $440K welfare report

Big Four consultancy billed Canberra top dollar, only for investigators to find bits written b

NYC's AI Chatbot Tells Businesses to Break the Law

Apple is pulling its AI-generated notifications for news after generating fake headlines

Top Law Firm Apologizes to Bankruptcy Judge for AI Hallucination

McDonald's scraps AI pilot after order mix-ups go viral

AI vs AI: Agent hacked McKinsey's chatbot and gained full read-write access in just two hours

Google shares drop $100 billion after its new AI chatbot makes a mistake

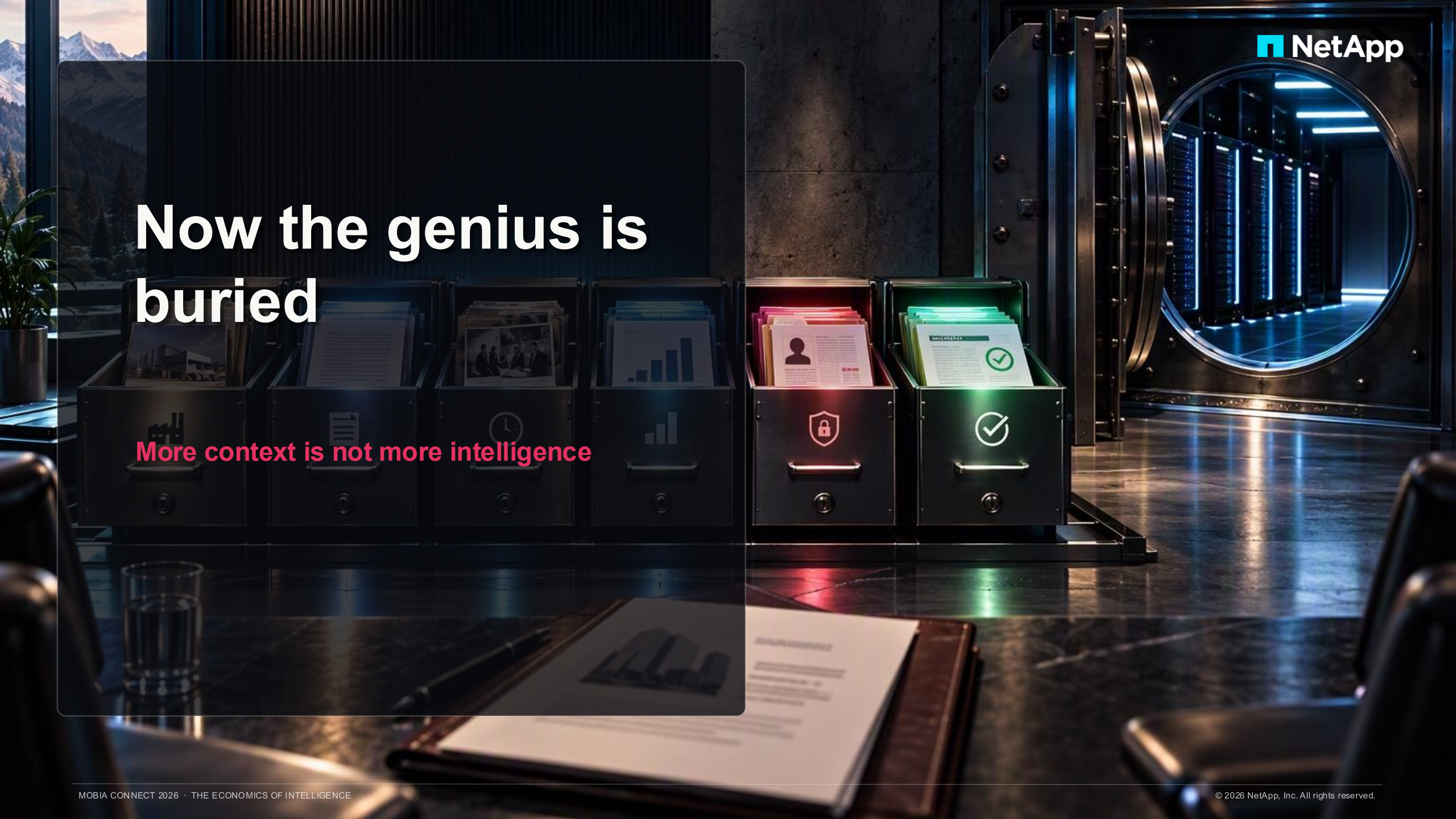
© 2026 NetApp, Inc. All rights reserved.



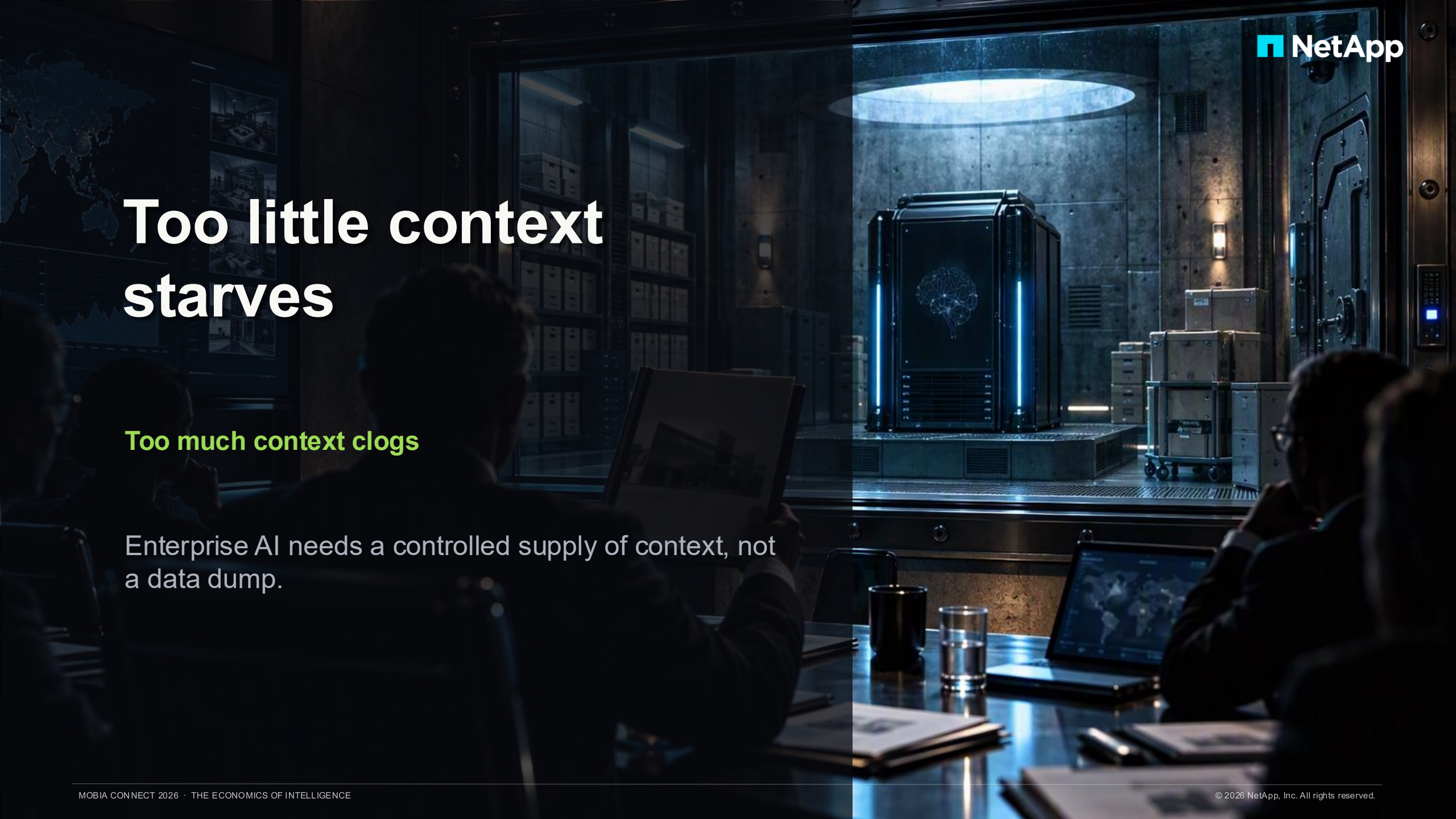
So we send the files

Every contract. Every policy. Every report.

© 2026 NetApp, Inc. All rights reserved.



© 2026 NetApp, Inc. All rights reserved.



Too little context starves

Too much context clogs

Enterprise AI needs a controlled supply of context, not a data dump.

© 2026 NetApp, Inc. All rights reserved.



NetApp

Wrong context creates defects

Old policy. Wrong region. Forbidden file.

An answer can be fluent and still be operationally wrong.

© 2026 NetApp, Inc. All rights reserved.

Context is inventory

Quality, freshness, access, and reuse matter

Treat context the way a manufacturer treats critical material.

© 2026 NetApp, Inc. All rights reserved.

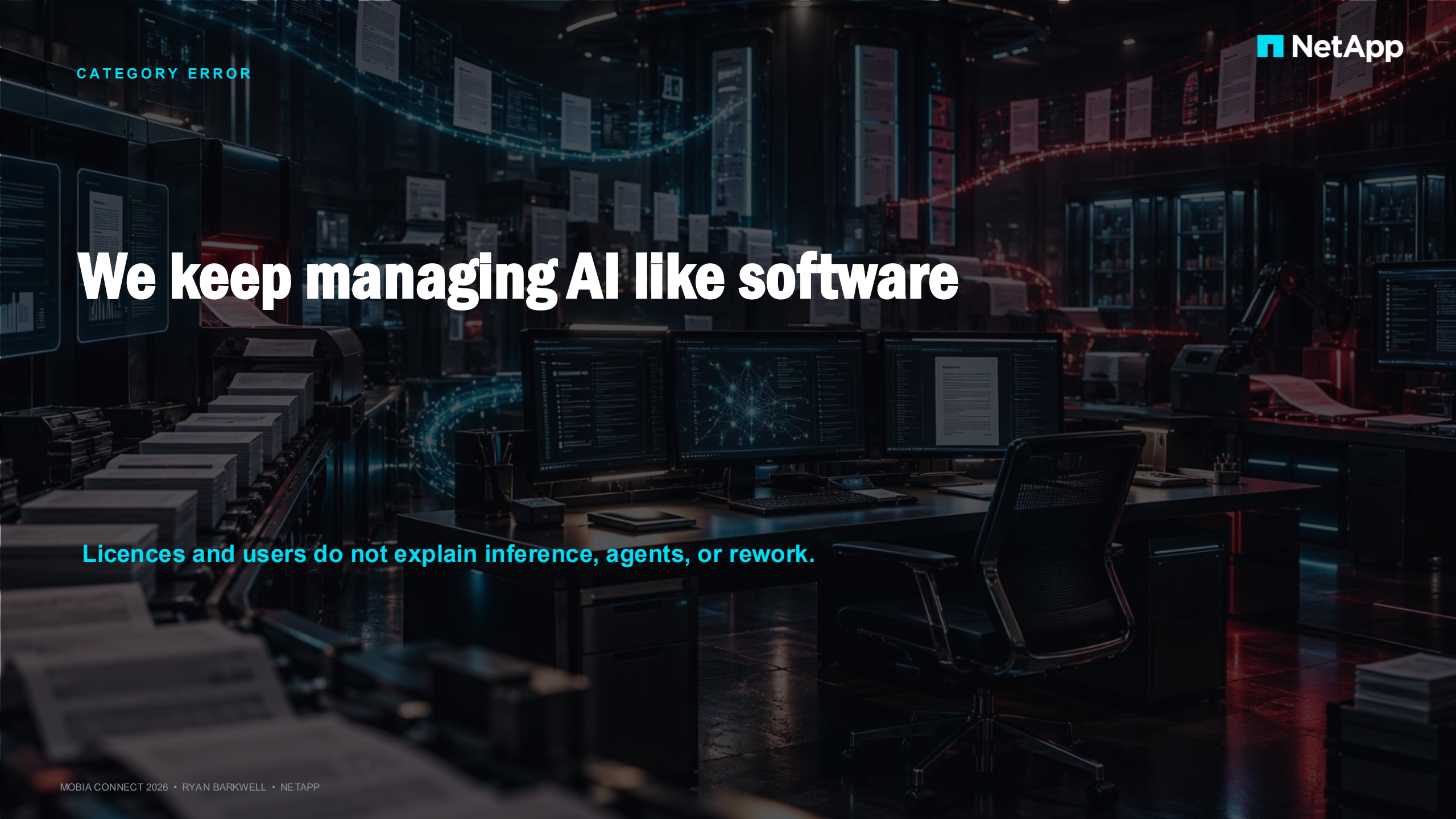
NetApp

Every answer has a bill of materials



Bill of materials

The economics of intelligence begin with seeing every input required to produce a usable outcome.



CATEGORY ERROR

We keep managing AI like software

Licences and users do not explain inference, agents, or rework.

METAPHOR

NetApp

It is closer to manufacturing



Every result consumes raw material, machinery, capacity, and inspection.



CONTRAST

NetApp

Intelligence is produced

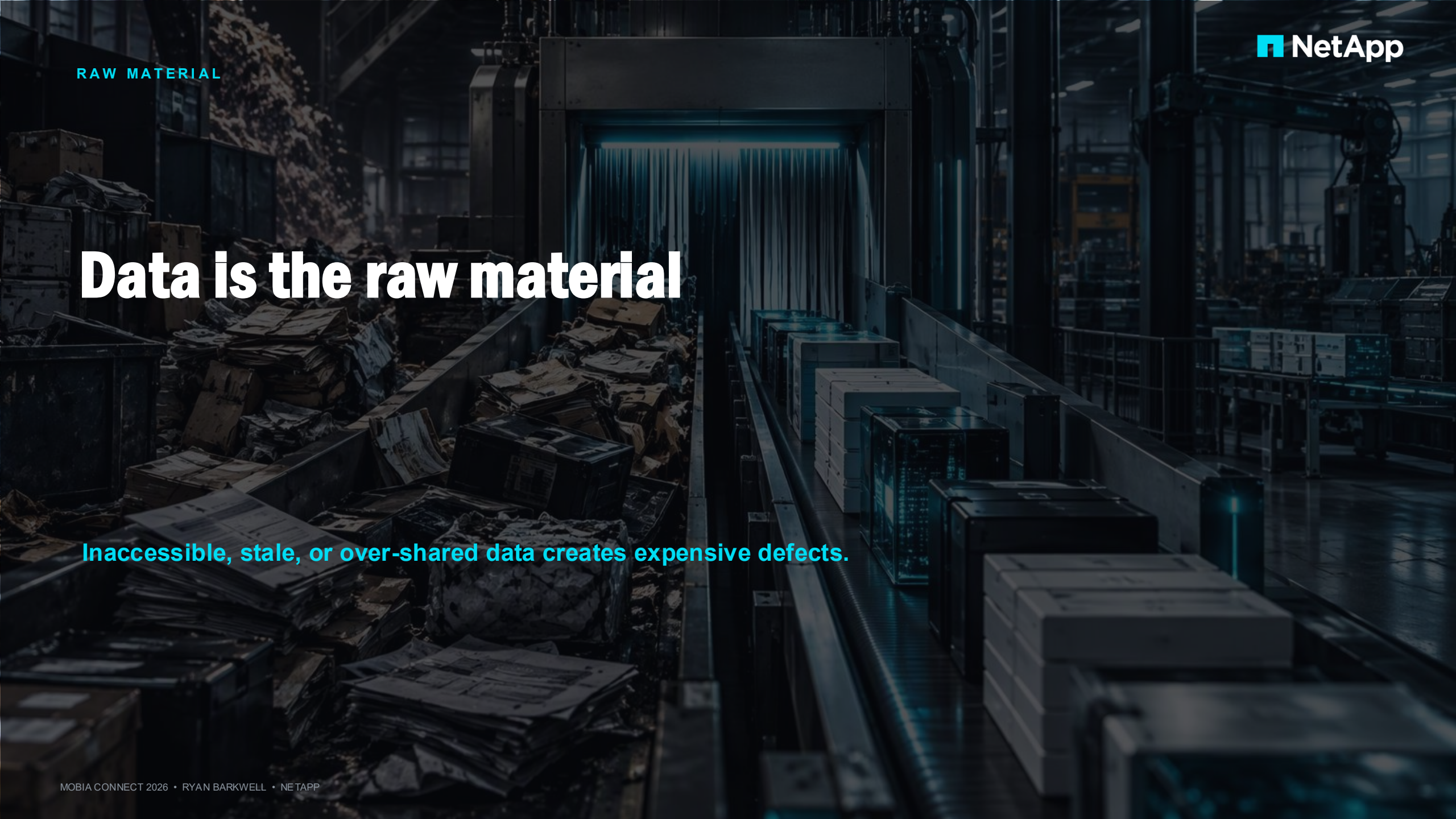
Every answer starts another production run.

NetApp

One answer crosses an entire production line



Factory



RAW MATERIAL

NetApp

Data is the raw material

Inaccessible, stale, or over-shared data creates expensive defects.

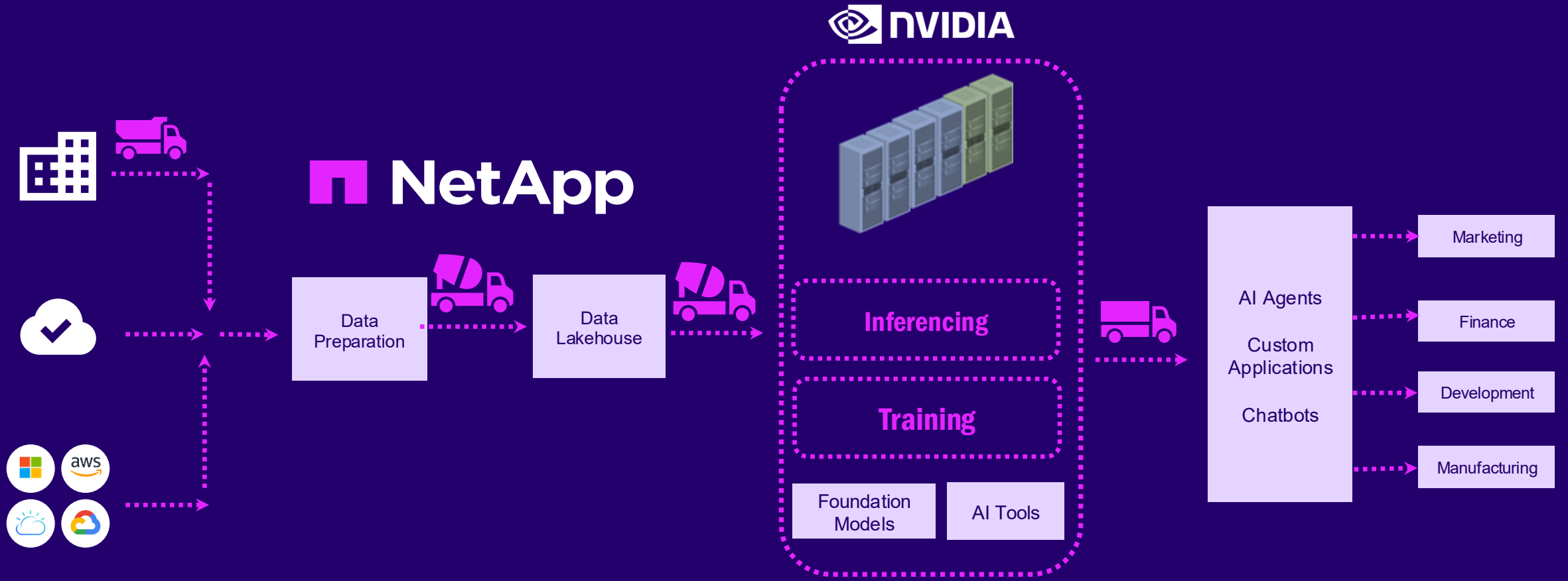
NetApp

“The most important thing is data. Data is the fuel of AI. That's what powers the agents.”

Jensen Huang



Data Infrastructure as the AI Supply Chain



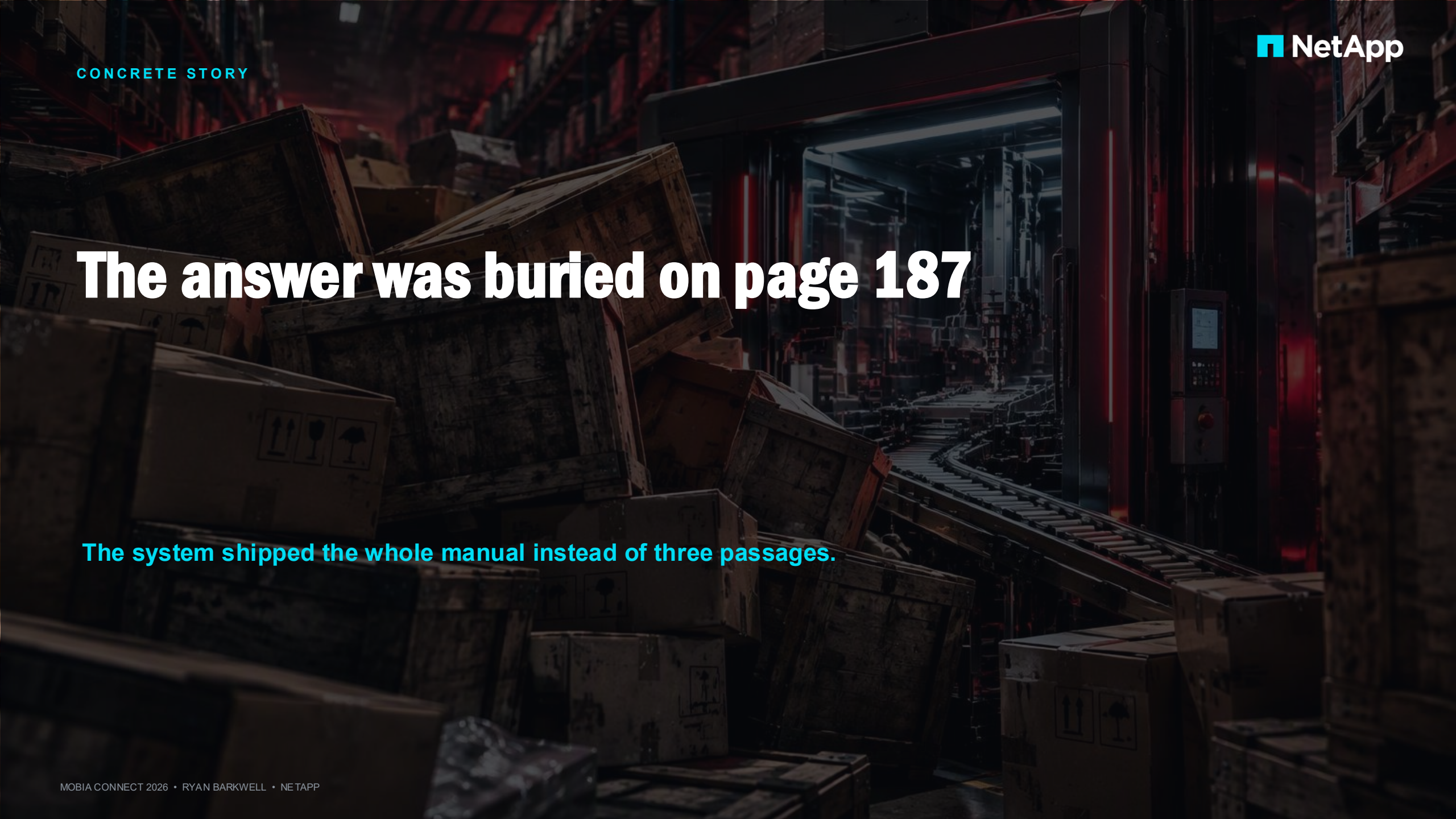
Raw Materials → Transportation & Orchestration → Assembly & Preparation → Quality Control → Manufacturing → Distribution → End User → Feedback Loop

Measure cost per outcome

Not merely cost per token

Executives fund resolved claims, approved decisions, and completed work. Tokens are an engineering measure.

© 2026 NetApp, Inc. All rights reserved.

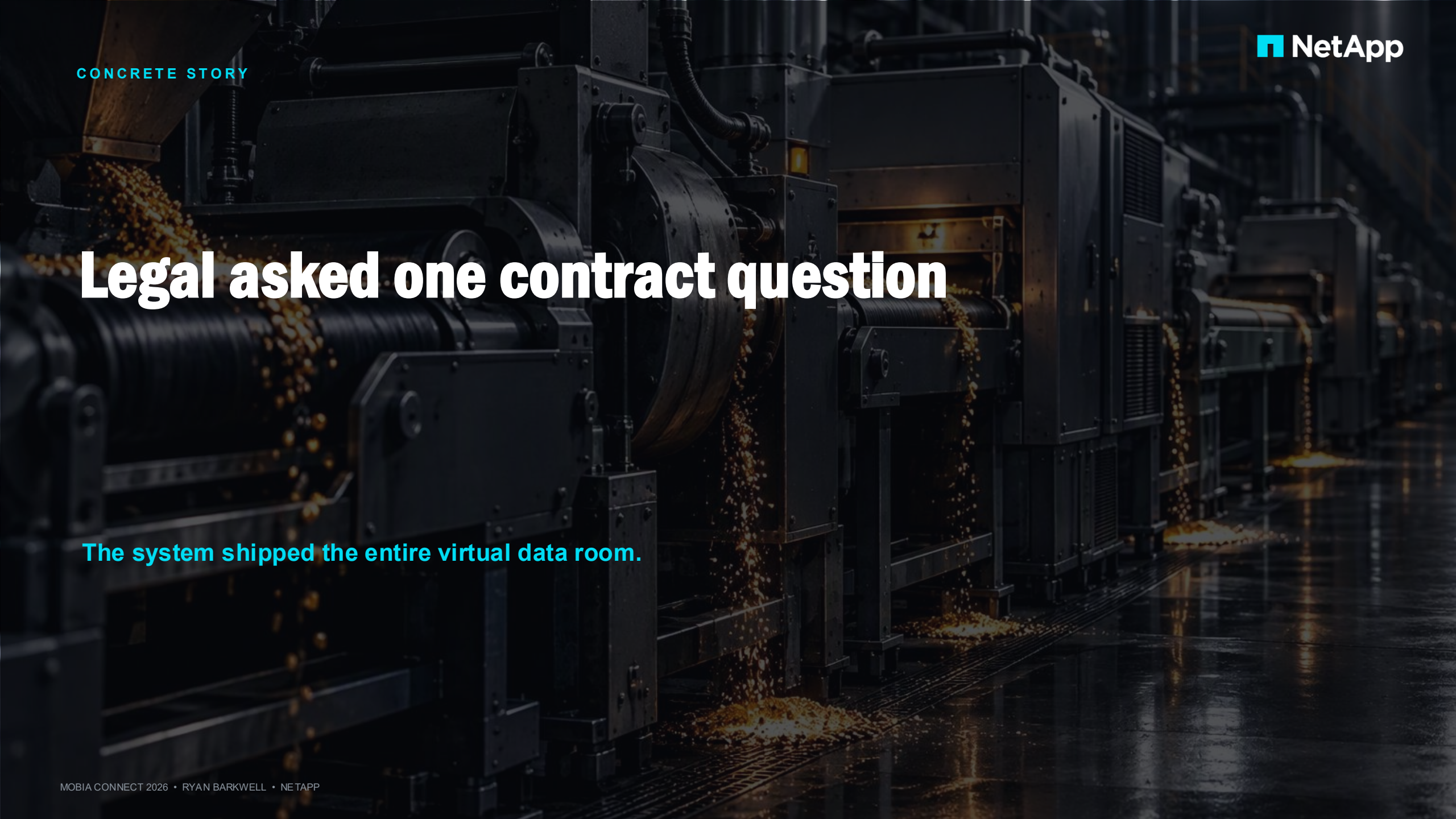


CONCRETE STORY

NetApp

The answer was buried on page 187

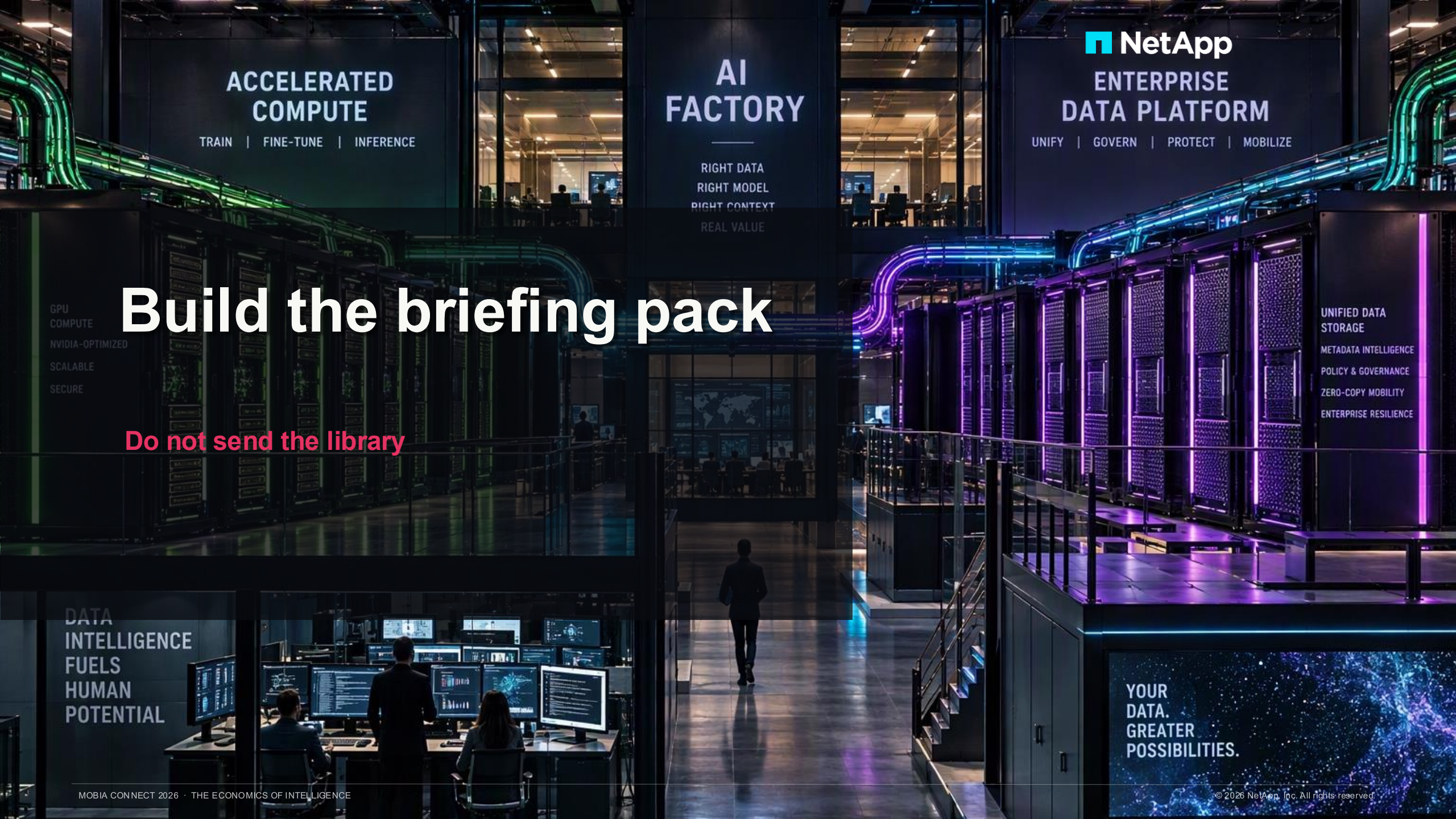
The system shipped the whole manual instead of three passages.



CONCRETE STORY

Legal asked one contract question

The system shipped the entire virtual data room.



Build the briefing pack

Do not send the library

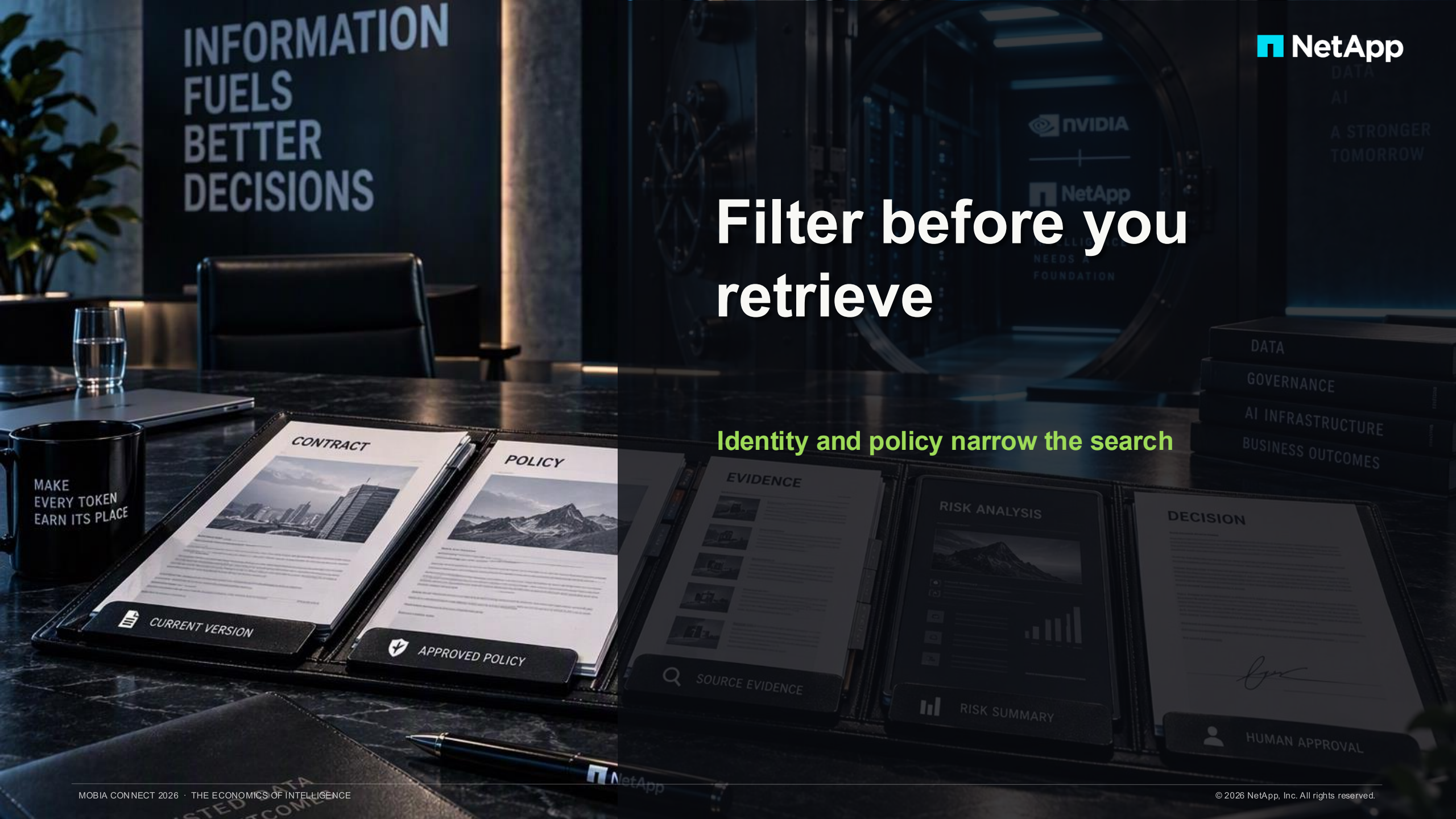
© 2026 NetApp, Inc. All rights reserved.

NetApp

Filter before you retrieve



Retrieval control



Filter before you retrieve

Identity and policy narrow the search

© 2026 NetApp, Inc. All rights reserved.

Retrieve the candidates

Find the small set that may matter

The goal is not maximum recall at any cost. It is a defensible candidate set.

© 2026 NetApp, Inc. All rights reserved.

Rerank the evidence

Put the best context first

The order and quality of context influence both answer quality and token consumption.

© 2026 NetApp, Inc. All rights reserved.

Prompt with purpose

Only the evidence needed now

A concise briefing pack reduces noise, latency, and repeated reasoning.

© 2026 NetApp, Inc. All rights reserved.



The policy changed Friday

Monday's answer is already obsolete

© 2026 NetApp, Inc. All rights reserved.



Update what changed

Do not rebuild everything

© 2026 NetApp, Inc. All rights reserved.

The answer was correct

The requester was not authorized



© 2026 NetApp, Inc. All rights reserved.



PERMISSIONS

The right answer reached the wrong person

The assistant surfaced a salary file the requester could not open.



Govern before retrieval

Permissions travel with the data

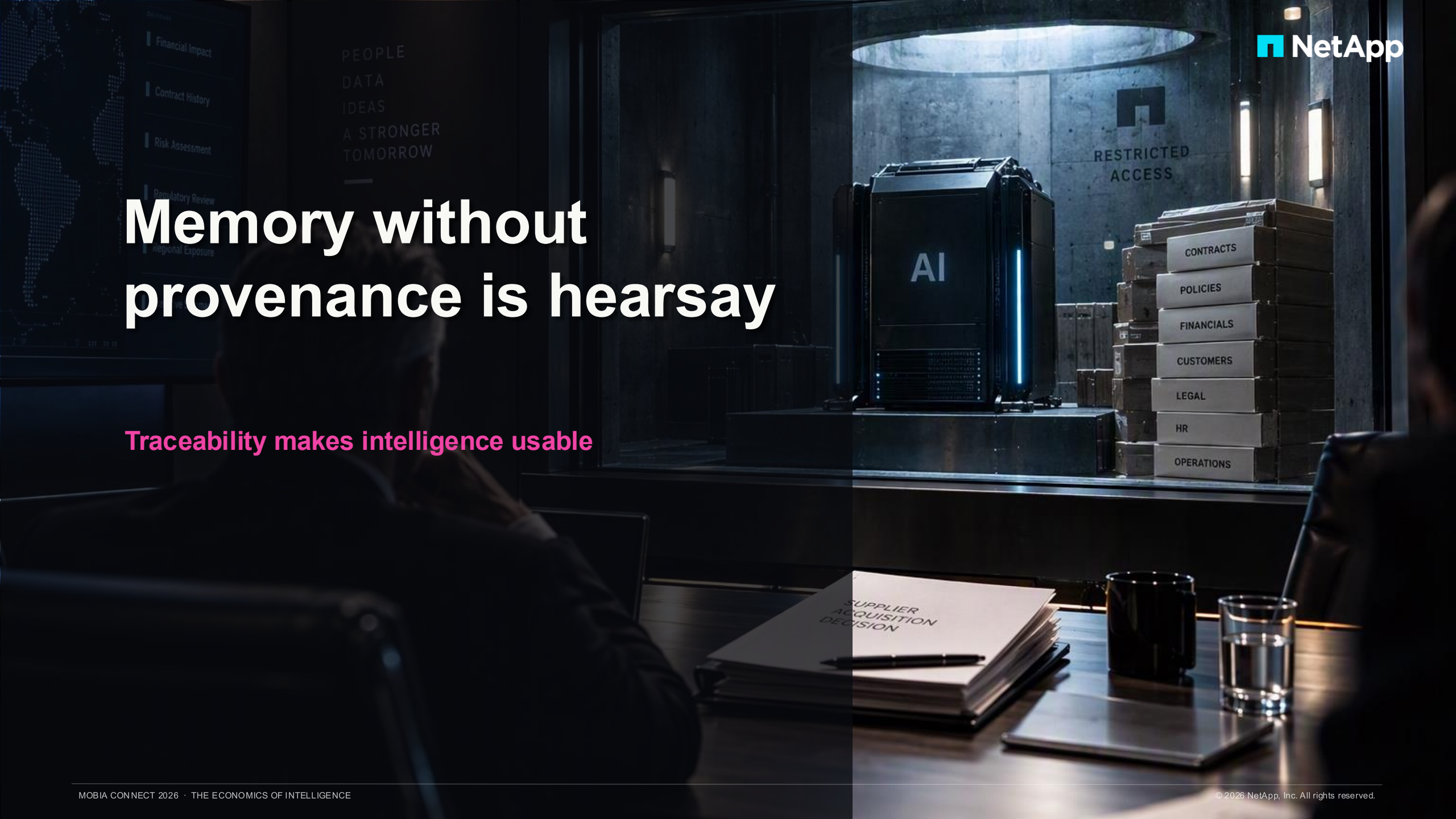
© 2026 NetApp, Inc. All rights reserved.



Every recommendation needs evidence

Show the source and version

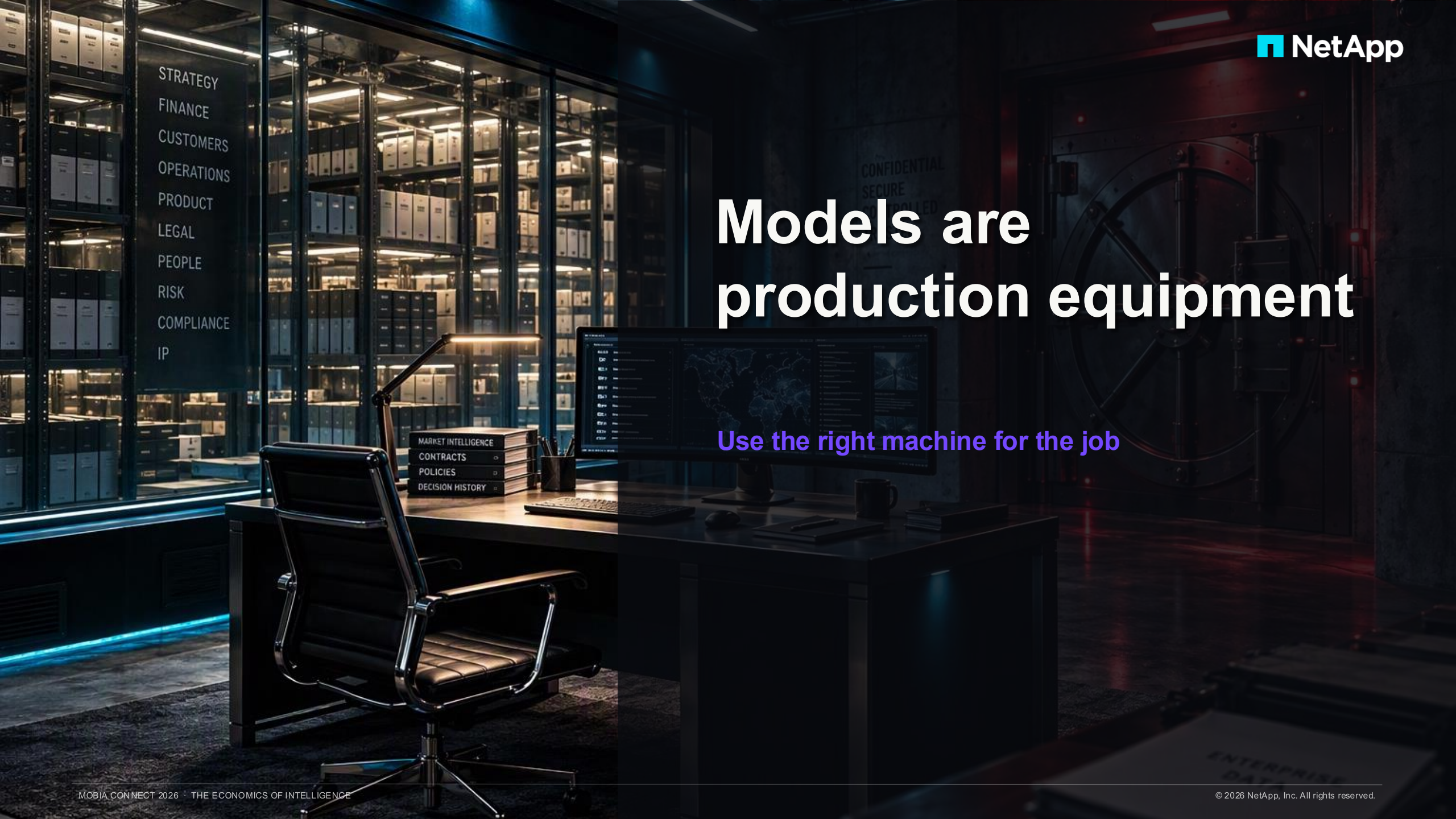
© 2026 NetApp, Inc. All rights reserved.



Memory without provenance is hearsay

Traceability makes intelligence usable

© 2026 NetApp, Inc. All rights reserved.



NetApp

Models are production equipment

Use the right machine for the job

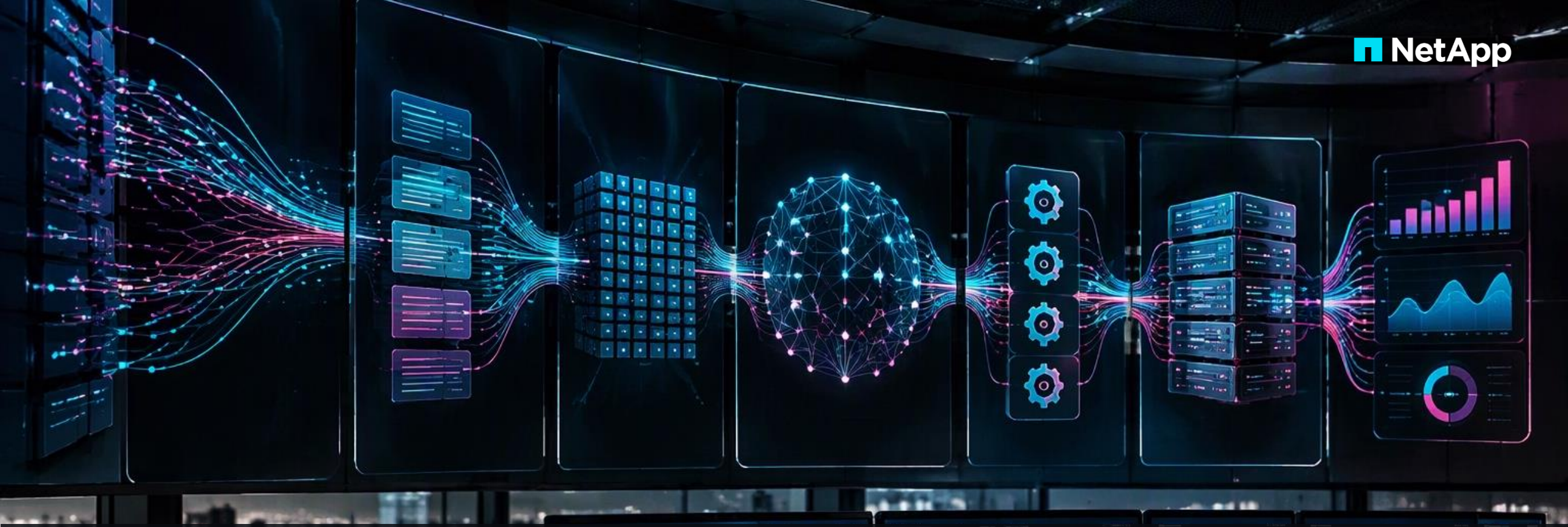
© 2026 NetApp, Inc. All rights reserved.

Don’t use a genius to alphabetize

Deterministic work needs deterministic tools



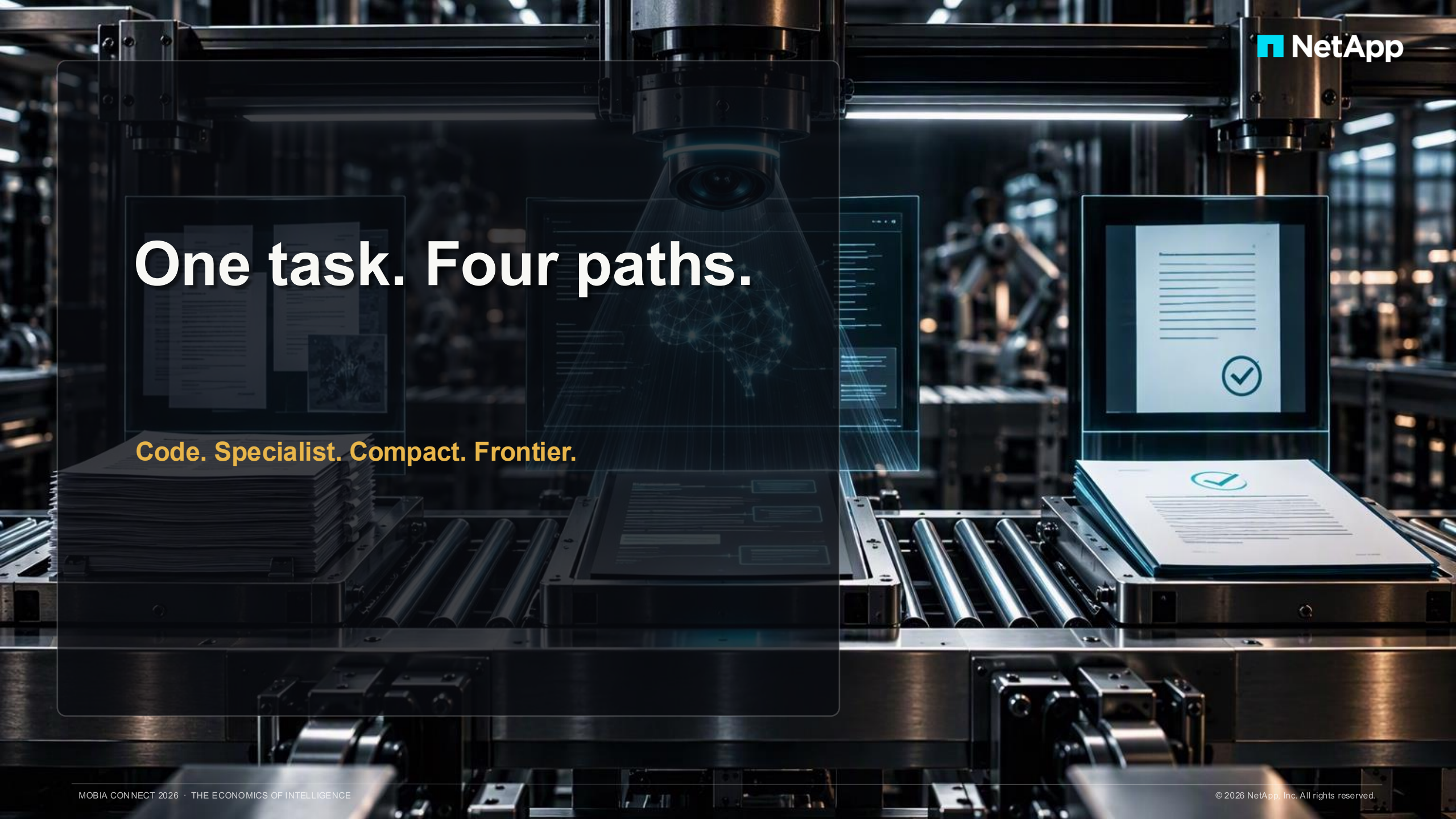
© 2026 NetApp, Inc. All rights reserved.



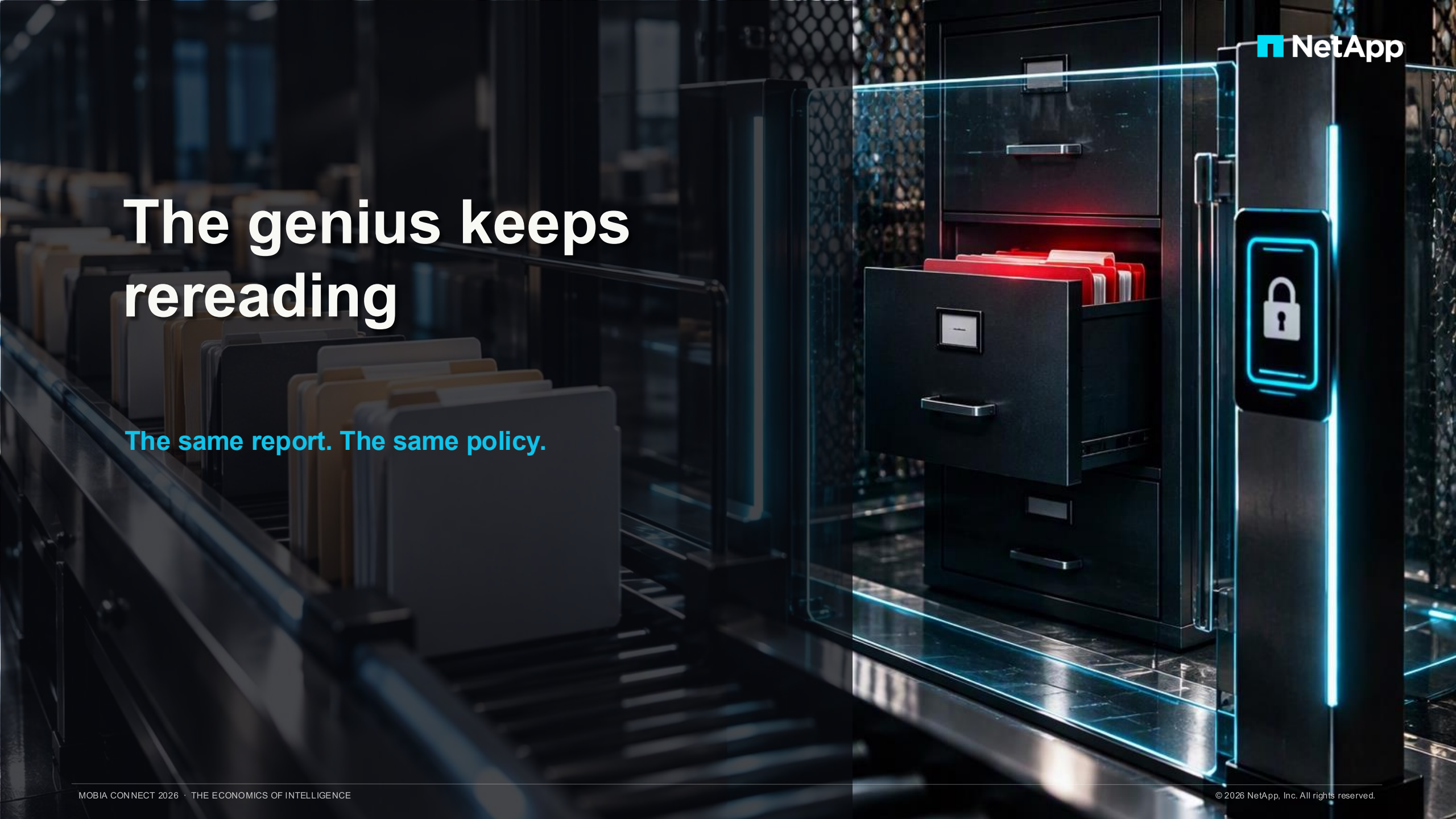
Route before you run

Match complexity to capability

© 2026 NetApp, Inc. All rights reserved.



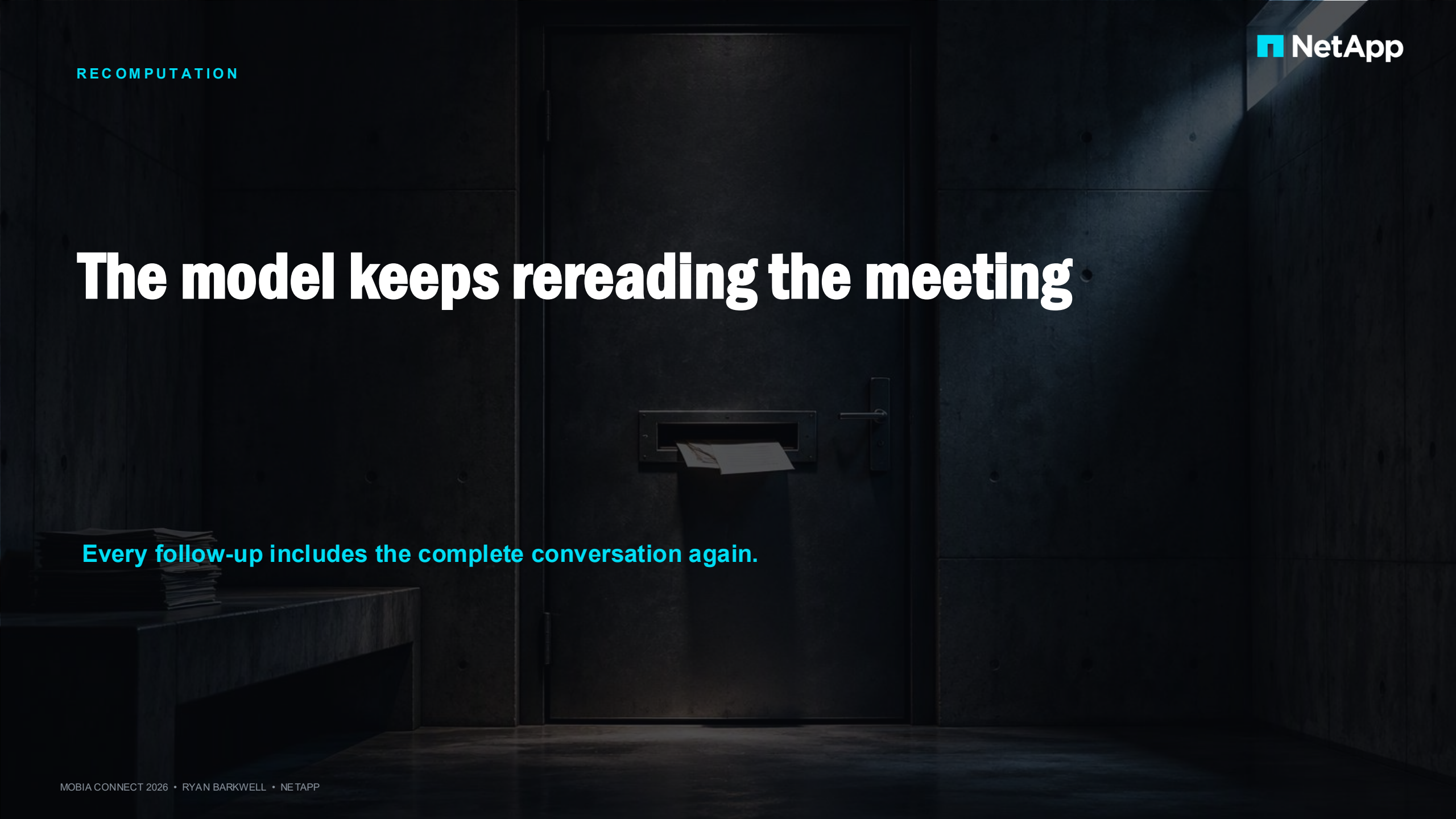
© 2026 NetApp, Inc. All rights reserved.



© 2026 NetApp, Inc. All rights reserved.

NetApp

The model keeps rereading the meeting



Every follow-up includes the complete conversation again.



Twenty analysts. One report.

Stable context should be reusable

© 2026 NetApp, Inc. All rights reserved.

Compute stable context once

Reuse it where the workload allows



© 2026 NetApp, Inc. All rights reserved.



Memory only pays when reused

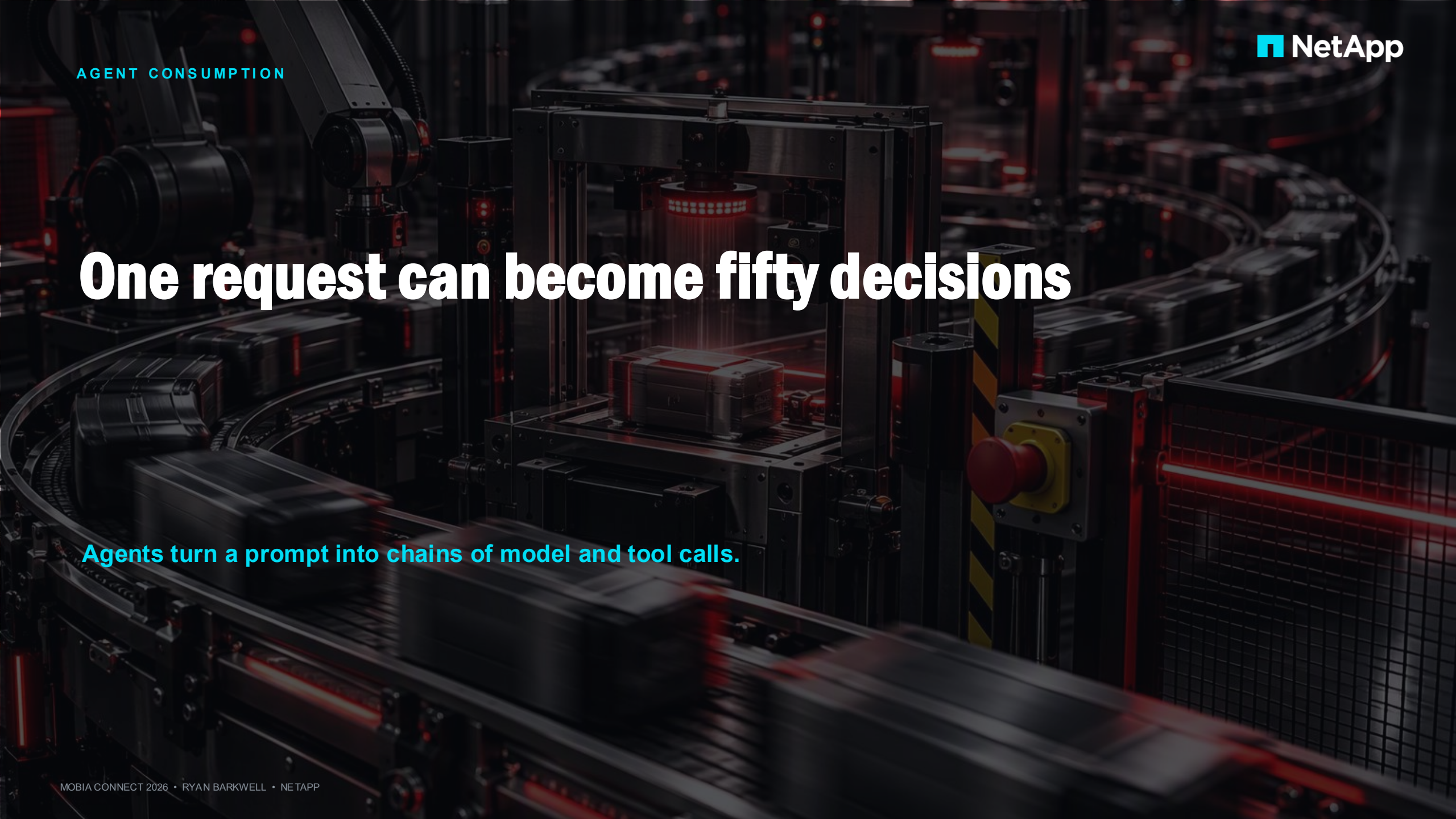
Measure hit rate, latency, and quality

© 2026 NetApp, Inc. All rights reserved.

The bunker door opens

The genius can now use tools

© 2026 NetApp, Inc. All rights reserved.

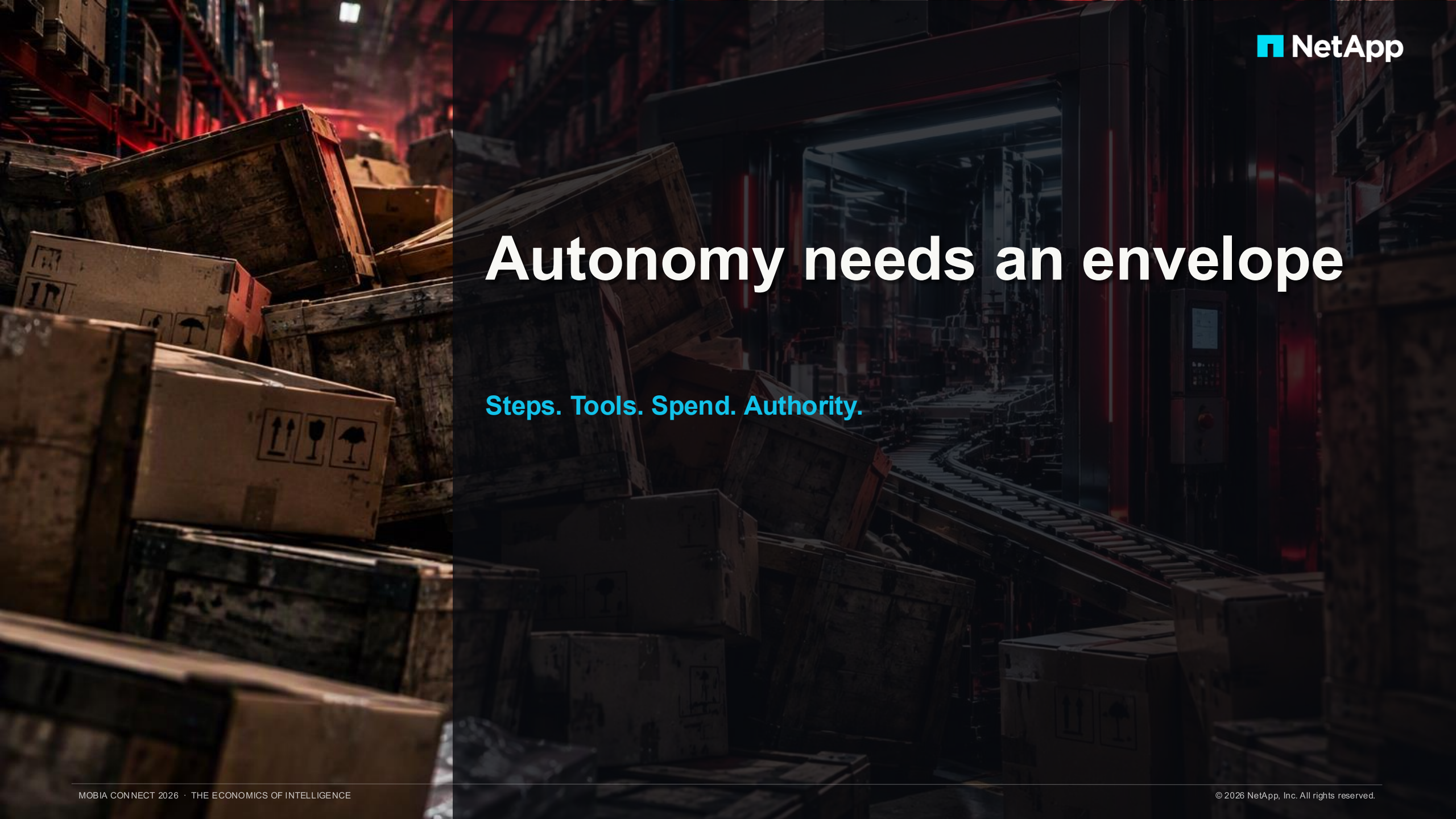


AGENT CONSUMPTION

NetApp

One request can become fifty decisions

Agents turn a prompt into chains of model and tool calls.



NetApp

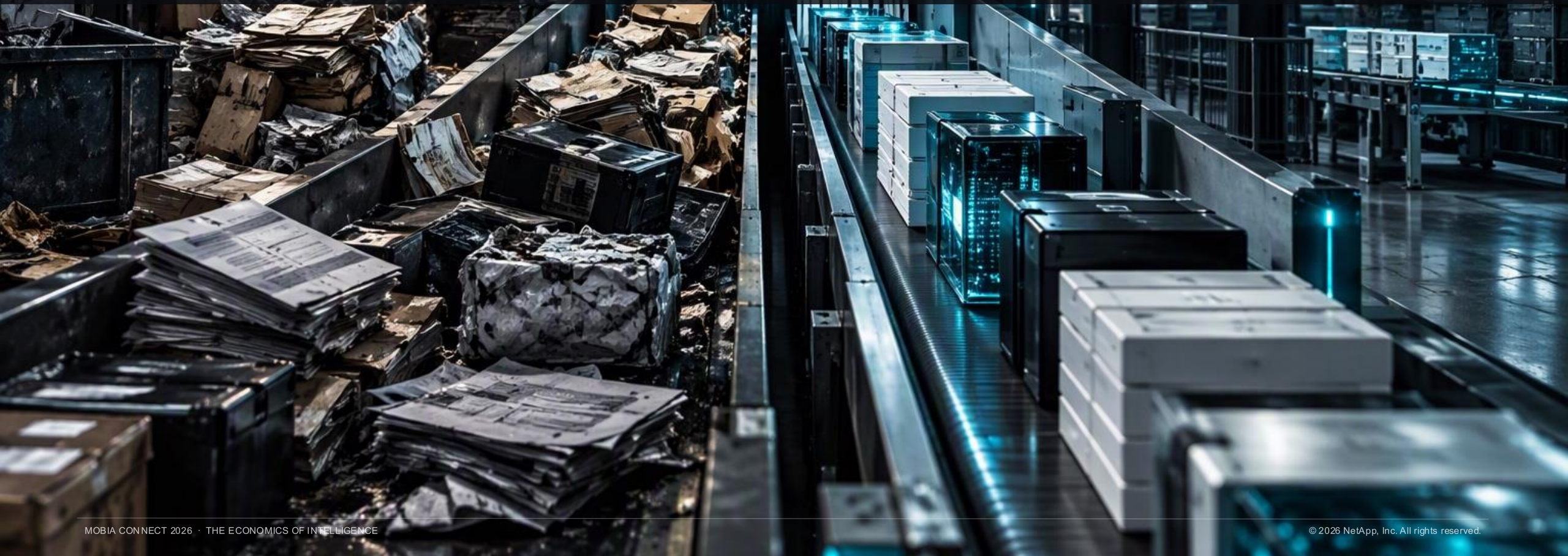
Autonomy needs an envelope

Steps. Tools. Spend. Authority.

© 2026 NetApp, Inc. All rights reserved.

Govern behavior and data

Two control planes. One outcome.



© 2026 NetApp, Inc. All rights reserved.



The CEO asks again

Should we modify the Q3 supply chain strategy?

© 2026 NetApp, Inc. All rights reserved.



© 2026 NetApp, Inc. All rights reserved.

Evidence → reasoning → rule → human

Accountability stays visible



© 2026 NetApp, Inc. All rights reserved.

NVIDIA builds the intelligence

Accelerated compute, models, and orchestration



© 2026 NetApp, Inc. All rights reserved.

NetApp gives it trusted memory

Discover, govern, protect, and mobilize data



© 2026 NetApp, Inc. All rights reserved.

NetApp



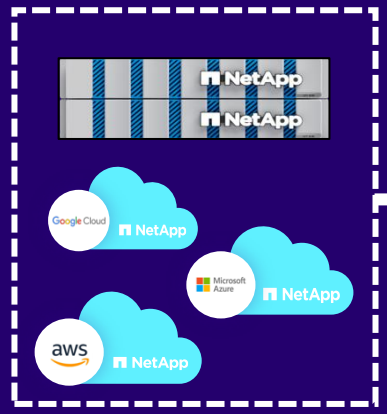
The Context Layer.

Nurturing the intelligence with your proprietary reality.

NetApp AI Data Engine: Accelerate enterprise AI with confidence

E2E AI data pipeline – seamless extension of ONTAP

ONTAP Storage



FIND & UNDERSTAND YOUR DATA

KEEP IT CURRENT

GOVERN & SECURE IT

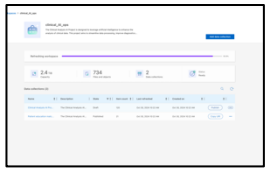
TRANSFORM IT

AI Data Engine (AIDE)

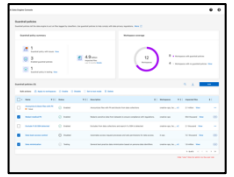
METADATA ENGINE



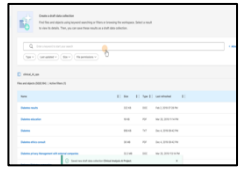
DATA SYNC



DATA GUARDRAILS



DATA CURATOR



across hybrid multiclouds

automatically, as source data changes

with guardrails throughout the data lifecycle

so it can be consumed by AI apps

TALK TO YOUR DATA

via semantic search

CHATBOT

AI AGENTS

LLaMA

Gemini

NetApp

© 2026 NetApp, Inc. All rights reserved.



Together: the intelligence factory

Right data. Right model. Right context.

© 2026 NetApp, Inc. All rights reserved.

THANK YOU

NetApp