



Concurrency Webinar · Microsoft Build 2026 Recap

Best of Build 2026: The Control Plane Ate the Keynote

What survived the week after the keynote confetti — three themes, one governed roadmap.

Brian Haydin — Solution Architect, Concurrency · Microsoft MVP candidate
Suneer Mehmood — Lead Data Architect, Concurrency

60 min · Two-voice · No Q&A

I Got Half of This Wrong

A confession to frame the hour, and the duality your developers are already living: move fast **AND** build right.

Pre-Build thesis

Microsoft is the marketplace, not the model competitor.

Post-Build

Microsoft is the marketplace **AND a model lab. The control plane is still the product.**

| *“AI models alone won’t change your business — the system running them will.” — Microsoft keynote, Build 2026*



One Map for the Next Fifty Minutes

Satya's frontier-intelligence stack — every Build announcement slots into one of these five layers.

Layer 5

Security & Governance — Entra Agent ID · Purview · Defender · ACS / ASSERT

Layer 4

Tooling — Foundry · GitHub Copilot · Microsoft Agent Framework

Layer 3

Agent Runtime — hosted agents · sandboxes · MXC containment

Layer 2

Models · Context · Tools — MAI models · Microsoft IQ · MCP

Layer 1

Compute Fabric — Azure · NVIDIA unified stack · MRC networking

The enterprise AI race is won on data context, not raw model power.



Agents Don't Have App-Shaped Traffic

Different call patterns mean different databases, a different warehouse, different context services.



∩ loops continuously — not request/response ∩

Hold that thought

That single fact explains every data announcement at Build. We'll cash it in around minute thirty.

Suneer's Five

Teaser speed — one line each. Payoff conversation on the themes slide.

Draft ranking · pending Suneer's final order

- 1 **Microsoft IQ — context as a product surface**
- 2 **Azure HorizonDB — Postgres rebuilt for agents**
- 3 **CoddSpeed — the Fabric warehouse got a GPU**
- 4 **Cosmos DB went AI-native**
- 5 **Fabric mirroring + medallion — the pattern we already ship**

Brian's Five

Each line is a promissory note redeemed in DD1. No explanation now — we'll get there.

- 1** Foundry's Operate layer goes GA — tracing, evals, one OTel pipeline, an ROI dashboard
- 2** ACS + ASSERT — an open, deterministic trust stack for agents
- 3** Agent 365 reaching local agents through MXC
- 4** Copilot Studio agents coming under the Foundry control plane
- 5** Microsoft Agent Framework v1.0 GA — multi-vendor, deterministic orchestration

Ten Picks. Three Stories.

Two scorecards collapse into three chapter titles — the rest of the hour walks them in order.

01 Theme — DD1 The control plane absorbed everything

Low-code, pro-code, foreign-cloud, local — one management plane, regardless of where the agent was born.

02 Theme — DD2 Context became infrastructure

The IQ layer surfaces. The data estate is rebuilt around how agents actually loop.

03 Themes — DD3 & DD4 Agents got identity and initiative

Entra Agent ID · Autopilots · trust perimeter reaches the laptop.



Worth Your Attention — Not Today's Hour

Seven real stories we're intentionally skipping today — the cut-line insurance policy.

Loudest community topic

GitHub Copilot app

Agent-native desktop development · Preview

Windows dev experience

WSL containers · Coreutils · Intelligent Terminal

Microsoft Discovery

Agentic R&D platform · GA

NVIDIA unified stack

+ MRC networking across the compute fabric

MDASH

Multi-agent security system for vulnerability hunting

Majorana 2

Quantum claims · peer review pending

If we do a Part Two, the GitHub Copilot app is slide one. The next forty minutes belong to the decisions that sit on **your** desk.

And yes — we're skipping quantum on purpose.





Theme 01 · Deep Dive 1 · 16:00 — 27:00

The Control Plane Absorbed Everything

"The governance layer is nowhere near the capability layer."

The sharpest community verdict, post-Build — in nine words

It no longer matters where your agent was born. Microsoft built one management plane — and is pulling all of it in.

You Can Finally See What the Agent Did

Foundry's Operate layer — AgentOps becomes a product, not a slide.

Azure AI Foundry — four layers

Build

Ground

Operate — tracing · evals · ROI

Reach

What shipped

GA Tracing & Evaluations

One OpenTelemetry pipeline · any framework, any deployment target.

GA Hosted agents

Per-session sandboxes · <100ms cold start · scale-to-zero.

Private Preview ROI for Agents

Task completion · time saved · cost efficiency — in-portal.

The analogy

An agent without tracing is a bird dog off in the woods — you **hope** it's working the right scent line, but you can't see it. Tracing is seeing what the dog is doing, every step.

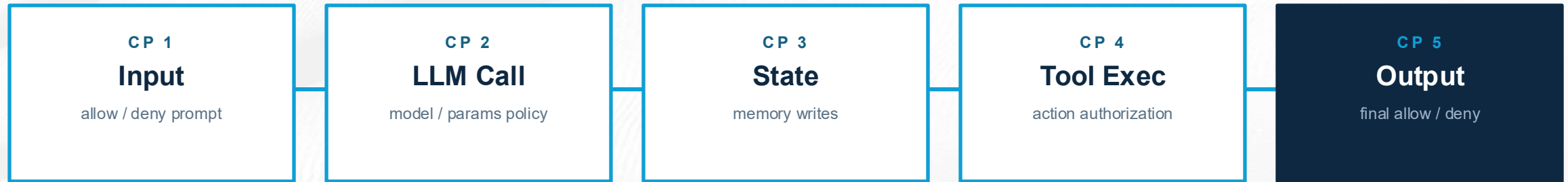
SUNEER COLOR → Every trace and eval is queryable telemetry. Your agent fleet just became a dataset.



Guardrails You Can Enforce, Not Just Request

The sleeper of the conference — probabilistic guardrails vs. deterministic enforcement.

Five deterministic checkpoints in the agent lifecycle



ACS — open spec

Agent Control Specification

Any runtime gets a **deterministic allow-or-deny** decision at each checkpoint. Structural enforcement — not “we instructed the model to behave.”

ASSERT — MIT-licensed

From Microsoft Research

Converts plain-text policy into **executable test suites**. Microsoft just open-sourced the enforcement half.

Works across Foundry · Microsoft Agent Framework · LangChain

“Innovation with seatbelts — finally bolted to the frame instead of duct-taped to the seat.”



One Control Plane, No Matter Where the Agent Was Born

Floor intel: Copilot Studio — pro-code — Google — AWS — local. All under one registry.



No slide exists for the best part — you had to be in the room

Confirmed on the floor with the Agent 365 + Foundry control-plane team: **Copilot Studio agents are being brought under the Foundry control plane.** Same registry, same policies, same observability as pro-code — and the discovery story extends to agents born in Google, AWS, anywhere.

GA · May 1

Agent 365 · \$15/user/mo · Entra Agent ID

Preview · ~July

Local-agent integration via MXC

The end of treating low-code and pro-code as separate governance tracks.



Build Multi-Vendor, Stay Deterministic

Microsoft Agent Framework — orchestrator deterministic; participants pluggable.

v1.0 · GA



For architecture decisions you make this quarter

Multi-vendor, modular, governed. That's not a demo pattern anymore — that's how you build now.



Theme 02 · Deep Dive 2 · 27:00 — 43:00 · Suneer leads

Context Became Infrastructure

I told you at minute four agents have different call patterns. Here's what Microsoft rebuilt because of it.

Callback → slide 04 · the store / retrieve / reason / act / learn loop

The Context Layer Goes Live Next Week

Microsoft IQ — the reusable context layer that grounds every agent.

Work IQ

M365 signals

Email, calendar, files, chats — permission-aware. Rides existing M365 signals; no new collector to deploy.

Fabric IQ

Business ontology

Semantic models + live telemetry — the “business graph” agents query for the real number, not a summary of one.

Foundry IQ

Knowledge & retrieval

Reusable knowledge sources, retrievers, and indexes — one definition, every agent in the workspace.

Web IQ

Live grounding

Live web grounding inside the same retrieval contract — same permissions, same audit trail.

Brightline grid-operator demo

Agent grounds in live web prices · queries the live grid ontology · pulls the SharePoint playbook — all permission-aware, in one call.

GA · Jun 2

Microsoft IQ

GA · Jun 16

Work IQ APIs

Limited Access

Web IQ



BRIAN COLOR → Work IQ APIs GA Tuesday. This isn't roadmap — it's next week.

Microsoft IQ — one context layer for every agent

GA · JUN 2026

WHAT

A context layer (no UI of its own) grounding agents in four sources: Web IQ (live web), Fabric IQ (business ontology over your Power BI semantic models + live telemetry), Work IQ (M365 signals), Foundry IQ (knowledge/retrieval).

WHY IT MATTERS — SCENARIO

Keynote demo (Brightline grid): the agent grounds in live electricity prices via Web IQ, queries the live grid ontology via Fabric IQ, then pulls the SharePoint response playbook via Work IQ — one reasoning pass, three layers of context.

ONBOARDING / MIGRATION

Existing estate is the on-ramp: Fabric IQ extends the Power BI semantic models you already run; enable via Azure Portal / M365 admin. Work IQ APIs GA June 16.

IMPLICATION

The bottleneck moves from model choice to data reach. Caveat: a lot of 'IQs' to track — coherent strategy, busy naming.



Keynote: Models + Context + Tools — the four IQs as the context layer.

Hi-res source: blogs.microsoft.com → 'Build 2026: Be yourself at work'

Azure HorizonDB — PostgreSQL rebuilt for agents

PUBLIC PREVIEW

WHAT

A fully-managed, PostgreSQL-compatible service built ground-up for scale: 128 TB per cluster, 15 read replicas, zone-redundant with automated failover, ~3× throughput vs self-managed Postgres. In-database model calls via the `azure_ai` extension, plus vector + semantic search.

WHY IT MATTERS — SCENARIO

A RAG app today juggles a transactional DB + a vector store + a search service. HorizonDB collapses that: retrieval and inference run next to the data, so a chatbot reads, embeds, searches and calls a model in one engine — fewer moving parts to secure and scale.

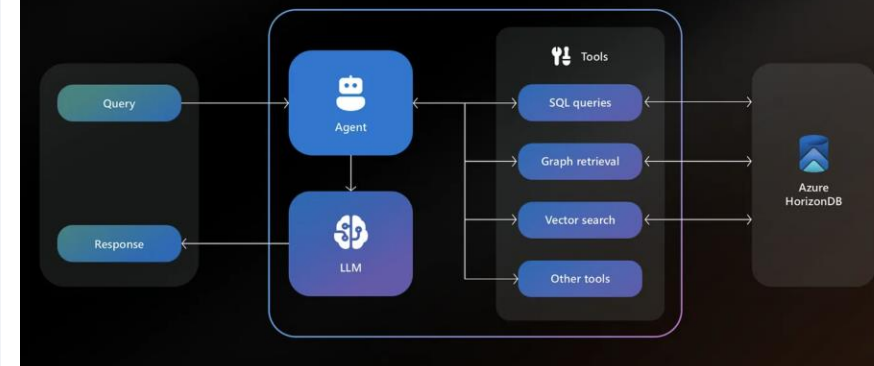
ONBOARDING / MIGRATION

New: spin up in one of 5 launch regions (more coming). Existing Azure DB for PostgreSQL clients have a familiar, Postgres-compatible migration path — but pilot only while in preview.

IMPLICATION

The data layer absorbs the AI retrieval stack. Preview + limited regions — pilot, don't roadmap yet.

Blueprint of an Agent on HorizonDB



Build expo: Azure relational databases — HorizonDB shown in-product.

Hi-res source: azure.microsoft.com → 'Building agentic apps with Fabric and Databases'

GPU-accelerated Fabric Data Warehouse (CoddSpeed)

EAP · JULY 2026

WHAT

Eligible warehouse queries run on NVIDIA GPUs inside the execution engine — no query rewrites, no cluster sizing. Microsoft reports up to 7× faster than three leading cloud DW vendors at 64-user concurrency (internal, May 2026). Research (CoddSpeed / TQP) won Best Industry Paper at SIGMOD 2026.

WHY IT MATTERS — SCENARIO

Heavy reporting and aggregation — month-end dashboards, agent-driven analytics — are where warehouses get slow and costly. Push that onto GPU compute transparently and the cost/performance math changes without touching your gold-layer models.

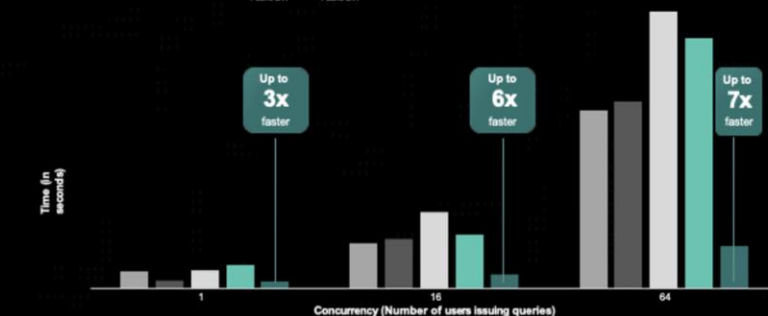
IMPLICATION

The warehouse query path is no longer CPU-only. Early Access Preview opens July 2026 — and I'd want independent benchmarks before quoting any number.

TPC-H 300GB Query Performance

GPU-powered Fabric DW

● Product A ● Product B ● Product C ● CPU-based Fabric DW ● GPU-accelerated Fabric DW



*Performance results are based on Microsoft internal testing conducted in May 2026 against three comparable data warehousing products from different companies in the same product space, using specific benchmark conditions. Actual results may vary depending on workload, configuration, and environment. Competitor identities and test configurations are not disclosed. These comparisons reflect selected commercially available products and may not be representative of all use cases.

Keynote: 'GPU-accelerated Fabric DW' — the first managed warehouse with GPU acceleration.

Hi-res source: [Fabric blog](#) → 'A new analytics frontier: GPU-accelerated Fabric DW'



Rayfin — the open-source SDK under Fabric Apps

PREVIEW

WHAT

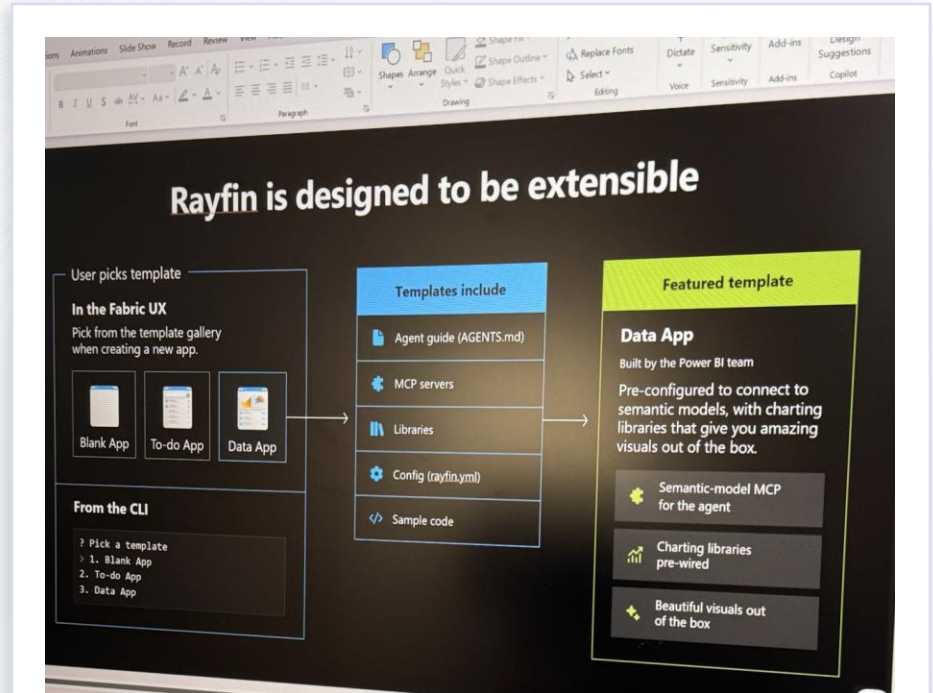
An open-source SDK + CLI (github.com/microsoft/rayfin). One command — `npx rayfin up` — provisions a full app backend on Fabric: SQL database, Microsoft Entra ID auth, a GraphQL API, and OneLake-connected hosting. Templates: Blank, To-do, Data App (wires straight to semantic models).

WHY IT MATTERS — SCENARIO

A BI team turns a governed Fabric semantic model into a working app or agent without hand-building the API and data plumbing. Because apps run in Fabric, their data lands in OneLake — instantly available to Power BI, notebooks and data agents. Replit partnership for the build experience.

IMPLICATION

Closes the old gap between governed analytics data and operational apps. Caveat: brand-new SDK — pilot it; ask how much you give up when your architecture doesn't fit the template's happy path.



Session: 'Rayfin is designed to be extensible' — templates, AGENTS.md, MCP servers, rayfin.yml.

Hi-res source: [Fabric blog](https://fabricblog.microsoft.com/introducing-rayfin) → 'Introducing Rayfin' · microsoft.com/microsoft-fabric/features/rayfin

Cosmos DB — GA today

Already shipped

Operational primitives that are safe to design against today.

🌐 **Global Secondary Index** — cross-partition queries.

🔧 **MCP Toolkit** for Cosmos DB.

🏠 **Agent Kit** — reference scaffolding.

🔄 **Per-partition auto-failover.**

📜 **All-versions & deletes change feed.**

🐧 **Linux emulator.**

Cosmos DB — Preview

AI-native primitives

Vector + memory + retrieval landed inside the operational DB.

🗂️ **Semantic Reranking.**

📄 **Integrated Embeddings.**

🧠 **Agent Memory Toolkit.**

🔍 **Agentic Retrieval.**

📁 **Distributed Transactions.**

Dating discipline

Cosmos→Fabric mirroring went **GA in November 2025** — it's not new at Build 2026. The AI-native primitives are what's new.



Fabric mirroring + medallion — the pattern we already build

WHAT

Operational store (e.g. Cosmos DB) mirrored into OneLake with zero ETL, then refined bronze → silver → gold so each audience — engineers, data scientists, business users — gets the right cut. Workload-isolated, full source-DB SLAs.

WHY IT MATTERS — SCENARIO

The ingestion layer is the most fragile, expensive thing we harden. Mirroring the operational store straight into OneLake in near-real-time takes a whole class of pipeline failure off the table — so the team spends its time on silver/gold modelling, not babysitting pipes.

MY TAKE

This is GA and shipping since Nov 2025 — that's the point. It's why I'm comfortable putting it at the centre of a client roadmap, not treating it as a bet. Trade-off: you give up some control over how data lands; the real modelling still lives in silver and gold.

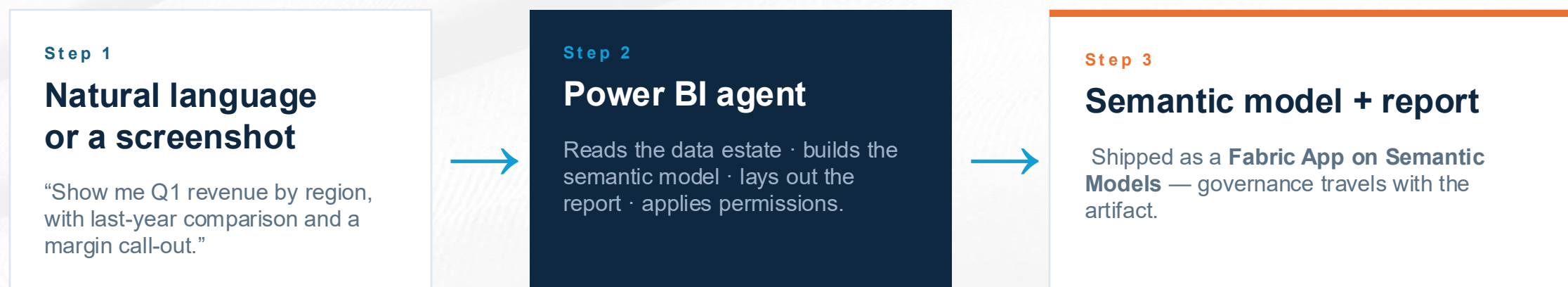


My floor photos: Cosmos→Fabric zero-ETL mirroring (top) feeding a bronze/silver/gold medallion (bottom).

Describe the Report. The Agent Builds It.

Preview

Power BI Agent Skills — from prompt or screenshot to a semantic model and a finished report.



Fabric Apps on Semantic Models — permissions, lineage, and audit travel with the agent-built artifact.

Say it plainly

"An agent builds your model" is only as good as the data estate underneath it.

Which is why the previous four slides matter before this one does.

Optimization for People Who Aren't Optimization Experts

Preview

Floor find

OptiGuide — Microsoft Research agent bringing supply-chain & scheduling optimization to non-OR experts.

What it does

Translates “reschedule the line, minimize overtime, hit the customer ship date” into a constrained-optimization run — without the OR PhD.

Who it's for

Plant operators, supply-chain managers, scheduling teams who own the constraints but not the solver.

If you run a plant

Go sign up.

Preview signup is open; the bar to enter is low and the upside is real.

Preview signup

aka.ms/optiguide-preview

Bonus slide · cuts first within DD2 if behind pace



Build on · safe to design against

GA pillars

Production-grade today. Build your 2026 estate around these.

■ **Fabric mirroring + medallion** — the proven pattern.

■ **Cosmos DB GA features** — GSI, MCP Toolkit, Agent Kit.

■ **Microsoft IQ** — Work IQ APIs GA next Tuesday.

Pilot · scope — not bet

Previews

One team, one workload, a real path to production.

■ **Azure HorizonDB** — Public Preview.

■ **CoddSpeed** — Early Access in July.

■ **Power BI Agent Skills** — Preview.

Operating rule

Context is the new bottleneck. The data estate is the moat.

of concept is a concept car; don't drive it on the highway.



Context. Interaction. Initiative.

One mental model for the brand sprawl — three stacked planes.

Plane 3 · new

Autopilots · Initiative

A copilot granted **identity, persistence, and initiative.**

Plane 2

Copilots · Interaction

Humans ask, AI answers. Episodic. Reactive.

Plane 1

The IQs · Context

Work IQ · Fabric IQ · Foundry IQ · Web IQ — what the agents **know.**

Scout — the first Autopilot. Own Entra Agent ID, not a shared service account. The attributability is what gets it past your CISO.



The First Autopilot — and Why You Pilot It Carefully

Frontier Enrollment

Scout — Microsoft's first Autopilot. Lives in Teams, Outlook, OneDrive, SharePoint. Acts in the background.

What it does

Acts on your behalf in the background — reads inbox, drafts replies, organizes files, prepares meetings. Own Entra Agent ID; every action attributable in the audit trail.

The gates that are real — use them

Frontier enrollment · Intune policy · opt-in attestation · Copilot license required.

One honest note

Scout's launch came with a leaked-memo controversy about engagement-driven design. Microsoft denied the framing.

I'm not litigating it. **Watch how preview terms and telemetry disclosures evolve** before you put it in front of your workforce.

90-day posture

Watch list, not buy list.

The community verdict is mixed, not settled. Stay close; defer the deployment decision.



Theme 03 · Deep Dive 4 · 49:00 — 57:00

The Trust Perimeter Reaches the Laptop

“Windows is the best place to run agents.”

— Satya Nadella, Build 2026 keynote

Routine · sensitive · latency

Local — near-zero tokens



Heavy reasoning

Cloud — escalate

The strategy isn't all-cloud or all-local. It's **routing**.

The Whole Hybrid Thesis in One Screenshot

MXC · OS-enforced

OpenClaw on Windows + Microsoft Execution Containers — receipt from the floor.

●●● OpenClaw — Windows ARM · localhost gateway

```
$ claw-gateway --bind localhost:6789
# Windows ARM node · OS containment active
> session: claude-opus-4.7
> permissions: fs.read /workspace · net: off
> tokens: 14,302 in · 2,118 out
# ← local inference: unmetered
```

OS-enforced
containment

Permissions
panel

Token meter
(unmetered local)

All in one frame

Open-source runtime · on a
Windows ARM machine ·
running an Anthropic model ·
permissions + token meter
visible.

Local-first · multi-vendor · governed ·
observable. Every box ticked.

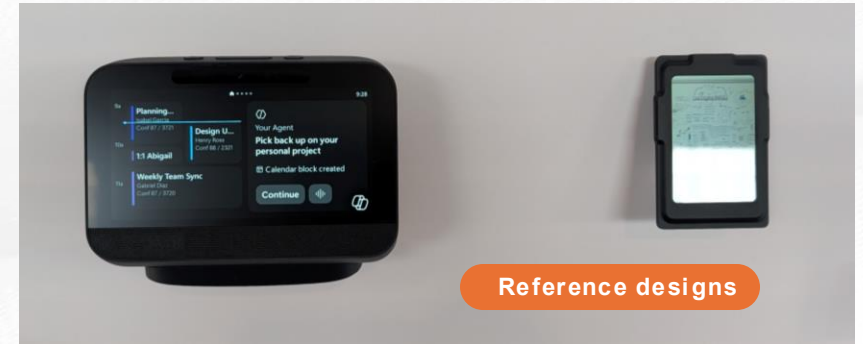
Preview · July Agent 365 local integration

Defender, Entra, Intune, and Purview reach local agents — including GitHub Copilot CLI and Claude Code on the roadmap.
The local-agent blind spot just got closed.

SUNEER COLOR → Local inference = unmetered tokens. Aion 1.0 Plan ships in-box — 14B reasoning · tool-calling · data never leaves the device.

Right Question. Wrong Device.

Project Solara — respectful disagreement. The race isn't the device; it's the local inference layer underneath.



Microsoft asked

“What is the next form factor for an agent-first world?”

Exactly the right question to be asking.

They answered

Project Solara — desk display + wearable badge.

Reference designs, not products. Retail pilots: CVS, Best Buy, Target, Levi's.

Brian's take

Right question. Wrong device.

For deskless retail — maybe. For everyone else, a desk device is a phone on a stand.

Steel-man

For deskless retail workers, a badge might genuinely beat a phone. Grant the strongest version of their case.

Where the race actually is

The body real estate is already taken — glasses, watch, lapel, phone. **Invest attention in the local inference layer underneath.**

A Petaflop on Your Desk — Eventually

Surface RTX Spark Dev Box — today it's a story, not a thing. The direction is unambiguous.



FCC pending



Local AI compute

~1
petaflop

verify spec morning-of

Unified memory

128
GB

shared CPU + GPU

Models that fit on the desk

120B
parameters

runs locally

Honest stance

Watch this space; don't budget for it yet. FCC authorization pending · ships later this year. Frankly, Microsoft catching up to where Apple has been on local inference.

DD4 closes on direction, not product.

Two Things We Couldn't Stop Talking About

Lightning round · below cut line

Brian · ~0:45

Data Formulator

Microsoft Research · open-source · agent-driven data viz.

Connects to

OneLake · Kusto · Postgres

On GitHub today. Go play with it.

Suneer · ~0:45

TRELLIS.2

Image-to-3D · research preview.

Why it matters

A glimpse of where multimodal is heading.

Thirty seconds of pure fun.



Concurrency

Thank you!

The models made the headlines. **The plumbing made the future.**

Three stories

- 01 The control plane absorbed everything
- 02 Context became infrastructure
- 03 Agents got identity and initiative

Follow us on  for the latest IT insights and industry updates! @Concurrency, Inc

