FRONTIER.

FRONTIER / NEW ANALYSIS



AWS Just Filled In the Agent Stack. Now It Wants the Road, the Guardrails, and the Tollbooth.

A two-week AWS release wave spanning frontier models, managed agents, runtime, observability, data access and monetization is more than feature volume. Collectively, it turns AWS into a credible operating environment for enterprise agents.

FRONTIER / Frontier Research / October 1, 2026

NEW ANALYSIS

Built to change a decision.

A two-week AWS release wave spanning frontier models, managed agents, runtime, observability, data access and monetization is more than feature volume. Collectively, it turns AWS into a credible operating environment for enterprise agents.

EDITION	**New Analysis**
AUTHOR	**Frontier Research**
PUBLISHED	**October 1, 2026**

This PDF is a publication rendering of the canonical Frontier News article. The web article remains the live source of truth; analysis, figures, and source citations travel with this edition.

What changed

AWS has spent the last two weeks shipping what looks, at first glance, like the usual hyperscaler spray of product updates. Taken together, it is something much more coherent.

The company has now filled in most of the missing layers required to make AWS a credible operating environment for enterprise agents: frontier-model choice, a managed agent harness, long-running execution, machine identity, human approvals, observability, retrieval, operational-data access, migration tooling and even the commercial plumbing needed to meter agent-enabled software.

The clearest signal is Amazon Bedrock Managed Agents, powered by OpenAI. AWS and OpenAI first disclosed the service in limited preview in April; on September 29 AWS announced broader preview availability. The service uses a customized version of the OpenAI Agents API engineered for AWS, with durable sessions, reusable skills, MCP-connected tools, an IAM role for each agent, human approval before consequential actions and CloudTrail recording of supported API activity.[1] [2]

That landed amid a rapid expansion of the intelligence layer. Bedrock added GPT-6.1 Sol on September 29, positioning it for agentic coding, computer use and professional work, including explicit prompt caching for repeated context.[3] A day later AWS added GPT-6 Astra UltraFast mode for latency-sensitive interactive agents.[4] Claude Sonnet 5.5 arrived on AWS with two access paths — Bedrock and Anthropic's native Claude Platform on AWS — while Grok 4.7 also joined Bedrock.[5] [6] AWS's own September 28 roundup framed the week's theme as model choice and highlighted Strands harness as a portable, Apache-licensed agent harness that can run across Bedrock, Anthropic, OpenAI, Google or local models.[7]

Below the model layer, AWS also reworked AgentCore Runtime for production-scale, long-running and bursty agents. AWS says the new runtime reclaims memory as sessions release it and uses snapshot-based startup to keep cold-start behavior more consistent across image sizes and concurrency. In AWS's own test, P75 platform cold starts stayed around two seconds across 200 MB to 2 GB images, compared with materially higher and size-sensitive startup times in the prior runtime.[8]

Then comes operations. CloudWatch Omni became generally available in late September as an OpenTelemetry-based observability and evaluation layer for applications and agents, available through a standalone web experience and IDE tooling rather than only the AWS console. AWS says Omni can span AWS accounts and Regions and ingest workloads from other clouds, including Azure.[9] Independent analysis from SiliconANGLE correctly identifies why this matters: traditional observability can tell you that an agent is healthy while missing that it chose the wrong tool, used stale context or produced the wrong business result.[10]

The data plane moved at the same time. Aurora PostgreSQL can now query Apache Iceberg and Parquet data directly using DuckDB embedded inside Aurora, allowing applications and agents to combine live operational records with historical data without first building and maintaining an ETL copy into the operational database.[11] SiliconANGLE independently confirmed the architecture and the elimination of a common reverse-ETL step for these workloads.[12] S3 Vectors, meanwhile, added metadata pre-filtering that evaluates tenant, user, category, path and other scope constraints before vector similarity search; AWS says selective filtered searches can return up to five times more matching vectors.[13]

AWS is also agentizing its own cloud machinery. A new AWS Marketplace skill can guide an AI coding assistant through building, deploying and validating usage-based SaaS metering via the AWS MCP Server, including code generation, CloudFormation, guardrails and an end-to-end test.[14] AWS Transform now uses agentic assessment to evaluate Kafka-to-MSK migrations, including compatibility, sizing and TCO scenarios.[15] And one layer lower, Amazon and Synopsys announced a multi-year agreement exceeding $1 billion aimed at accelerating Amazon custom silicon and AWS infrastructure, while extending engineering collaboration around AI-powered design and Trainium/Graviton optimization.[16]

Individually, several of these are incremental. Collectively, they amount to a platform event.

Why it matters

The strategic change is not that AWS suddenly has “agents.” It has had agent services, Bedrock, AgentCore and an increasingly large AI portfolio for some time. The change is that the stack is becoming continuous enough that a CIO can now plausibly treat an agent as a first-class workload on AWS rather than as an application assembled from a dozen disconnected AI-specific services.

That distinction matters because enterprise agents are beginning to look less like chatbots and more like persistent machine principals. They preserve state. They invoke tools. They act across multiple steps. They may carry credentials. They need to be restarted, observed, audited, constrained, upgraded and eventually retired. The winning enterprise platform therefore needs more than a strong model. It needs an operating environment around the model.

AWS is now unusually credible across five of those layers.

First, model neutrality is becoming a practical advantage rather than a brochure claim. OpenAI, Anthropic and SpaceXAI frontier models are now available inside the same broad Bedrock environment, with AWS controls around access, networking and audit. Claude Sonnet 5.5 can even be consumed either through Bedrock or Anthropic's native platform on AWS.[5] That gives AWS an increasingly strong answer to the question that has haunted hyperscalers since the generative-AI boom began: what happens if today's preferred model is not tomorrow's?

The answer AWS is trying to make obvious is: change the intelligence, not the estate.

Second, the runtime and governance layers are converging. Bedrock Managed Agents gives OpenAI-powered agents durable state, tools, skills, a dedicated IAM role, approval gates and CloudTrail integration.[2] AgentCore supplies the managed execution substrate beneath long-running and bursty agents.[8] This is a meaningful architectural shift because identity and runtime used to be things an enterprise AI team had to assemble around the agent. AWS is trying to make them native properties of the workload.

That is exactly where the enterprise-agent market is heading. Microsoft Agent 365 is explicitly positioning itself as a control plane for inventory, identity, governance, security and lifecycle management across Microsoft, third-party and custom agents.[17] Google's Gemini Enterprise Agent Platform similarly provides Agent Registry, Agent Identity and Agent Gateway to govern agents, tools, MCP servers and egress.[18] AWS is not alone in seeing the control plane as strategic. The entire hyperscaler market is converging on the same insight.

Third, AWS is connecting agent observability to conventional operations. Omni is important because the failure mode of an agent is not merely downtime. The agent can be up, fast and technically error-free while making a poor decision. By combining application telemetry with agent traces and evaluation, AWS is trying to let operations teams move from “is the service healthy?” to “why did the machine decide to do that?”[9] [10]

That is a much bigger deal than another dashboard. Persistent agents will create far more decision traces than humans can manually inspect. Evaluation, provenance and investigation history therefore start to become operational controls.

Fourth, the data announcements reduce one of the largest hidden taxes in enterprise agents: getting the right context into the right execution path. Aurora's direct Iceberg/Parquet querying means an operational application or agent can reach historical lake data without maintaining a separate replication path.[11] S3 Vectors pre-filtering makes scoped retrieval — this tenant, this employee, this case, this jurisdiction — more precise before similarity search even begins.[13] Both are mundane-looking data features that become much more consequential in an agentic architecture, where the cost and risk of copying every potentially useful dataset into a separate AI store are difficult to justify.

Fifth, AWS is extending agents into economics and migration, not just development. Marketplace metering delivered as an agent skill is a small release with a large tell: AWS wants its APIs, commercial systems and cloud-management workflows to become agent-consumable primitives.[14] AWS Transform makes the same point from another direction: an agent can now do the compatibility, sizing and business-case work that precedes a cloud migration.[15]

This is AWS behaving like AWS again.

The company does not need to own the dominant assistant, the dominant foundation model or the dominant enterprise productivity surface if it can own the substrate beneath heterogeneous agents. The late-September releases make that strategy much easier to see.

Frontier take

Our read is that AWS's emerging advantage is not “better agents.” It is becoming the least-regret place to operate everyone else's agents.

That is a materially stronger strategic position than AWS had even a few months ago.

Microsoft still has an extraordinary advantage at the point where employees actually work: Microsoft 365, Teams, Entra, Purview, Defender and Agent 365 create a direct bridge from human identity and productivity context into machine work.[17] Google has a similarly explicit control-plane architecture in Gemini Enterprise Agent Platform, including registry, agent identity and gateway enforcement.[18] Neither should be treated as behind.

AWS's edge is different. It is the possibility of combining heterogeneous intelligence with the cloud's existing operating disciplines: IAM, CloudTrail, networking, serverless execution, databases, object storage, vector retrieval, observability, marketplace economics and increasingly custom silicon.

That gives AWS a credible way to win without declaring a single agent or model the center of the universe.

There is also an important limit to the thesis. Bedrock Managed Agents is still in preview, and some of the most strategically interesting elements of this stack are newly launched enough that production economics, operational rough edges and governance depth still need to be proven at scale.[2] CloudWatch Omni is generally available, but the observability market is crowded and OpenTelemetry lowers the switching barrier even as AWS tries to make its intelligence layer sticky.[9] [10] Model neutrality can also become table stakes rather than differentiation as Microsoft, Google and independent platforms expand their own cross-model and cross-agent support.

But the direction is now hard to miss.

The agent-era cloud will not be won solely by whoever has the smartest model. It will be won partly by whoever makes autonomous software easiest to run, constrain, observe, feed with trusted context, meter and replace.

AWS has spent late September filling those boxes in.

That is why the announcement wave should be read as one strategic event rather than a collection of unrelated product notices. AWS is not merely adding AI features to the cloud. It is turning the cloud itself into the operating environment for machine labor.

The savage part of the strategy is that AWS does not have to win the agent.

It can let OpenAI, Anthropic, SpaceXAI, open-source projects and enterprise developers fight that war — while AWS sells the road, the guardrails and, increasingly, the tollbooth.

Three moves for CIOs

1. — Separate the agent brain from the operating substrate Architect new enterprise agents so model selection is replaceable while identity, policy, tool contracts, telemetry and data-access boundaries remain stable. Treat Bedrock's widening model catalog as evidence that intelligence should be a swappable dependency, not the root of your agent architecture.
   - Decision trigger: Apply this pattern whenever an agent is expected to survive more than one model generation, use more than one model class, or become embedded in a business process with a multi-year life.
   - Why now: Frontier models are converging and leapfrogging too quickly to justify rebuilding the surrounding controls every time the preferred model changes. AWS's current wave makes model portability operationally realistic enough that CIOs should demand it as an architecture property.
2. — Promote agent identity and decision traces into the control plane For every persistent agent with write authority, require a dedicated machine identity, scoped credentials, explicit approval boundaries for consequential actions, and end-to-end decision telemetry that joins model, tool, application and infrastructure traces. On AWS, test the IAM-per-agent, CloudTrail and Omni pattern as one integrated control design rather than three separate products.
   - Decision trigger: Make this mandatory the moment an agent can change production data, communicate externally, spend money, alter infrastructure or invoke another agent with greater authority.
   - Why now: The industry is moving from agents that recommend to agents that act. Once authority becomes persistent, traditional uptime monitoring and shared service credentials are insufficient; the operational question becomes who authorized the action, what context drove it and how the organization can reconstruct the decision.
3. — Benchmark the whole agent economics stack, not token price Run one production-shaped agent workload across your leading cloud options and measure total cost per completed business outcome: model inference, runtime idle/peak behavior, retrieval, data movement, observability, evaluation, approvals, storage and human exception handling. Include the cost of duplicating or synchronizing enterprise data when the agent cannot reach it in place.
   - Decision trigger: Do this before committing a material agent portfolio to one cloud, and repeat when a platform introduces a major runtime, data-access or observability change such as the current AWS wave.
   - Why now: AWS is attacking agent economics above and below inference: AgentCore changes runtime utilization, Aurora and S3 Vectors change context-access costs, Marketplace skills automate monetization, and custom silicon continues to pressure the infrastructure layer. Token price alone is becoming an increasingly misleading proxy for the cost of autonomous work.

Sources

- AWS — Amazon Bedrock Managed Agents, powered by OpenAI, preview [2]
- AWS — OpenAI models, Codex and Managed Agents limited preview [1]
- AWS — GPT-6.1 Sol on Amazon Bedrock [3]
- AWS — GPT-6 Astra UltraFast mode on Amazon Bedrock [4]
- AWS — Claude Sonnet 5.5 on AWS [5]
- AWS — Grok 4.7 on Amazon Bedrock [6]
- AWS — The new AgentCore runtime [8]
- AWS — Amazon CloudWatch Omni general availability [9]
- SiliconANGLE — CloudWatch Omni and agent observability [10]
- AWS — Aurora PostgreSQL direct Iceberg and Parquet queries [11]
- SiliconANGLE — AWS embeds DuckDB in Aurora PostgreSQL [12]
- AWS — S3 Vectors metadata pre-filtering [13]
- AWS — Marketplace AI agent metering skill [14]
- AWS — Transform agentic Kafka migration assessments [15]
- AWS — Weekly Roundup, September 28, 2026 [7]
- Microsoft — Agent 365 general availability [17]
- Google Cloud — Gemini Enterprise Agent Registry [18]
- Synopsys — Amazon custom silicon agreement [16]

SOURCES / CITATIONS

Trace the argument.

Every source below is linked. Source numbering follows first appearance in the report.

01 **AWS — OpenAI models, Codex and Managed Agents limited preview**
https://aws.amazon.com/about-aws/whats-new/2026/04/bedrock-openai-models-codex-managed-agents/

02 **AWS — Amazon Bedrock Managed Agents, powered by OpenAI, preview**
https://aws.amazon.com/about-aws/whats-new/2026/09/bedrock-managed-agents-preview/

03 **AWS — GPT-6.1 Sol on Amazon Bedrock**
https://aws.amazon.com/about-aws/whats-new/2026/09/openai-gpt-6-1-sol-on-amazon-bedrock/

04 **AWS — GPT-6 Astra UltraFast mode on Amazon Bedrock**
https://aws.amazon.com/about-aws/whats-new/2026/09/openai-gpt-6-astra-ultrafast-on-amazon-bedrock/

05 **AWS — Claude Sonnet 5.5 on AWS**
https://aws.amazon.com/about-aws/whats-new/2026/09/claude-sonnet-5-5-aws/

06 **AWS — Grok 4.7 on Amazon Bedrock**
https://aws.amazon.com/about-aws/whats-new/2026/09/amazon-bedrock-grok-4-7/

07 **AWS — Weekly Roundup, September 28, 2026**
https://aws.amazon.com/blogs/aws/aws-weekly-roundup-gpt-6-sol-and-luna-claude-opus-5-5-on-amazon-bedrock-strands-harness-and-more-september-28-2026/

08 **AWS — The new AgentCore runtime**
https://aws.amazon.com/blogs/machine-learning/the-new-agentcore-runtime-elastic-optimized-and-consistently-fast-starts/

09 **AWS — Amazon CloudWatch Omni general availability**
https://aws.amazon.com/about-aws/whats-new/2026/09/amazon-cloudwatch-omni-ai/

10 **SiliconANGLE — CloudWatch Omni and agent observability**
https://siliconangle.com/2026/09/27/aws-cloudwatch-omni-goes-after-the-hardest-question-in-agentic-ai-why-did-the-agent-do-that/

11 **AWS — Aurora PostgreSQL direct Iceberg and Parquet queries**
https://aws.amazon.com/blogs/aws/amazon-aurora-postgresql-now-supports-direct-querying-of-apache-iceberg-and-parquet-data-in-your-data-lake/

12 **SiliconANGLE — AWS embeds DuckDB in Aurora PostgreSQL**
https://siliconangle.com/2026/09/30/aws-embeds-duckdb-in-its-postgresql-dbms-to-speed-queries-of-live-and-historical-data/

13 **AWS — S3 Vectors metadata pre-filtering**
https://aws.amazon.com/about-aws/whats-new/2026/09/s3-vectors-introduces-metadata-pre-filtering/

14 **AWS — Marketplace AI agent metering skill**
https://aws.amazon.com/about-aws/whats-new/2026/09/marketplace-ai-agent-metering/

15 **AWS — Transform agentic Kafka migration assessments**
https://aws.amazon.com/about-aws/whats-new/2026/09/amazon-msk-migration-assessment-with-aws-transform/

16 **Synopsys — Amazon custom silicon agreement**
https://news.synopsys.com/2026-09-30-Synopsys-and-Amazon-Announce-Strategic%2C-Multi-year-IP-Agreement-for-Custom-Silicon-Collaboration-Also-Extends-to-Cloud-and-AI-Powered-Engineering

17 **Microsoft — Agent 365 general availability**

https://www.microsoft.com/en-us/security/blog/2026/05/01/microsoft-agent-365-now-generally-available-expands-capabilities-and-integrations/

18 **Google Cloud — Gemini Enterprise Agent Registry**

https://docs.cloud.google.com/gemini-enterprise-agent-platform/govern/agent-registry

FRONTIER.

TAKE THE NEXT STEP

Push this further.

Research is useful when it changes a decision. Bring us the question, the constraint, or the argument you want tested.

RESEARCH / BRIEFINGS

inquiries@thisisfrontier.ai

Briefings, advisory work, executive discussions, and research inquiries.

REPRINTS / LICENSING

reprints@thisisfrontier.ai

Reprints, enterprise distribution, excerpts, and licensing.

FRONTIER

www.thisisfrontier.ai

Research, Data Insights, News, and the live Frontier corpus.

All rights reserved. Contact Frontier for redistribution or reprint rights.

SEE FURTHER. MOVE SOONER.