Secure & Intelligent Meetings

AI Benefits Without the Cloud

AI-Powered Intelligence Without the Cloud

Tools like Microsoft Copilot, Zoom AI Companion, and Google Gemini make meeting intelligence feel automatic, but nearly all of them depend on the public cloud.

For some organizations, that's a dealbreaker.

While the cloud has long been hailed as the answer to everything — flexibility, cost-efficiency, scale — it's becoming clear that it's not a universal solution. Public sector institutions, defense organizations, financial services, and healthcare providers increasingly find themselves at odds with the very model that enabled today's SaaS boom. Their problem isn't with AI. It's with where AI lives.

The Rise and Limits of the Cloud

The cloud story was supposed to be simple. Offload infrastructure. Get elastic compute power. Pay only for what you use. But in practice, cloud usage introduced new kinds of complexity:

- **Loss of control** — Cloud software updates are pushed by vendors without notice. Features can disappear or morph overnight, without input or consent from IT leaders. Worse, data often flows outside the organization's firewall — not just into external storage, but into opaque, shared-model AI engines that may use that data to train algorithms. For regulated sectors, this isn't just a compliance concern — it's an operational risk. When meeting data is processed off-site and outside of a customer's control, confidentiality, sovereignty, and even ownership become difficult to guarantee.
- **Expense creep** — AI workloads, especially voice transcription and meeting video analysis, aren't light. Organizations often find themselves surprised by costs that scale with hours of meetings, not headcount.
- **Vendor lock-in** — Each platform builds its own AI tools — useful, but siloed. Microsoft's features don't transfer to Zoom, and vice versa. Organizations working across multiple platforms are forced to duplicate their efforts.

These include not only the risk of losing control over sensitive data, but also the growing threat of systemic failure.

When the Cloud Fails: The CrowdStrike Outage and Delta's $500 Million Meltdown

CrowdStrike's July 2024 outage wasn't just a high-profile event; it was a wake-up call. A routine update from the cybersecurity firm's Falcon sensor triggered a configuration error that brought down roughly 8.5 million Windows systems worldwide. The scale of disruption was staggering: Hospitals rerouted patients, banks suspended operations, emergency services reverted to manual protocols, and airport terminals went dark. But what made the outage especially unsettling was that it wasn't caused by a malicious attack or catastrophic weather event — it was the unintended consequence of a single background update.

The incident laid bare the hidden risk of modern cloud-reliant infrastructure: loss of autonomy. Organizations that pride themselves on building resilient, well-architected systems often find themselves compelled to integrate third-party services to meet compliance mandates or fulfill functional expectations, such as endpoint protection, AI transcription, or cloud video analysis. And yet, a failure in any one of these dependencies can ripple through the enterprise with devastating speed. The interconnectedness that enables agility also creates blind spots, where updates are pushed without warning, services depend on other services, and a single point of failure can paralyze operations.

Even organizations with strong on-prem strategies often find themselves tethered to the cloud in ways they can't fully control. The push toward hybrid architectures blurs the lines between local and external systems. One cloud API dependency, one required license check, or one centralized update server can undo years of planning and segmentation. The more tightly integrated the cloud becomes, the more critical it is to ask: Who really holds the keys?

Delta Air Lines became the most visible casualty of the CrowdStrike incident. Over a five-day period, the airline canceled more than 7,000 flights, affecting approximately 1.3 million passengers. The financial impact was estimated at $550 million in lost revenue and compensation. The U.S. Department of Transportation launched an investigation into the airline's response, but the root of the problem wasn't Delta's infrastructure. It was the dependency chain itself. The airline's operations were crippled by a third-party software update, on a system outside its direct control.

This case underscores the risks associated with over-reliance on cloud-based services and third-party vendors for mission-critical operations. It serves as a stark reminder that while cloud-based solutions offer scalability and convenience, they can also introduce significant risks, especially when organizations cede control of critical systems to external providers.

The Repatriation Reckoning: Why CIOs Are Rethinking the Cloud

A growing number of CIOs are no longer treating cloud-first as a foregone conclusion. Instead, they're reassessing their infrastructure choices with a new level of scrutiny. The movement — often referred to as cloud repatriation — involves pulling select workloads back from public cloud providers and redeploying them within private or on-premises infrastructure. What once felt like a retreat is now seen as a strategic recalibration.

This shift isn't about nostalgia for racks and server rooms. It's about minimizing risks and reasserting strategic priorities across cost, performance, compliance, and control. Organizations are increasingly reevaluating their reliance on the cloud in response to escalating costs, unpredictable billing, latency concerns, regulatory scrutiny, and data sovereignty challenges. In sectors like healthcare, financial services, government, and defense, the risks of handing over critical systems to third-party providers are beginning to outweigh the benefits of elasticity and rapid provisioning. For these organizations, the choice isn't between cloud and on-prem — it's between control and compromise.

In 2025, cloud repatriation is expected to continue trending upward. A survey conducted in late 2024 found that 86% of CIOs planned to move some public cloud workloads back to private cloud or on-premises infrastructure — the highest percentage recorded to date.[1] While most organizations aren't abandoning the cloud entirely, many are selectively pulling back critical components such as production data, backups, and compute resources. This more nuanced approach reflects a growing awareness of the long-term costs, risks, and limitations of relying on the cloud, especially in light of unpredictable billing models, complex vendor relationships, and increased regulatory scrutiny.

On-Prem AI: Innovation Without Exposure

What's surprising is that this shift is taking place during the height of the AI boom — an era largely defined by cloud-native generative models. But repatriation doesn't mean abandoning innovation. It means implementing innovation differently. Enterprises are discovering that the benefits of modern AI — including automation, transcription, summarization, and analytics — can still be realized without shipping data across the internet or into third-party large language models (LLMs).

Take transcription, for example. What began as a simple convenience has quickly become foundational to AI-powered collaboration. Transcripts create a persistent record of what was said, enabling everything from compliance and audit trails to smarter search, real-time translation, accessibility support, and even automatic action item extraction. They also form the basis for more advanced capabilities, such as meeting summarization

1. Barclays CIO Survey, Q4 2024. https://tinyurl.com/Barclays2024Survey

and intent recognition, providing AI with a common language to work with — a kind of lingua franca for downstream intelligence. The key insight is that these tools don't inherently require the public cloud. With the right architecture, transcription and its powerful downstream applications can be handled entirely on-premises, preserving both privacy and control while still delivering real innovation.

Deploying AI in on-prem environments allows organizations to preserve the value of features like transcription, summarization, and analytics, while retaining full control over their data. It enables governance of the entire data lifecycle, from capture to retention, ensures that model updates happen on the organization's terms, and allows training on domain-specific language without breaching confidentiality policies. This is especially critical in areas built on unstructured human communication — such as meetings, support calls, and incident reviews — where the insights are rich, but the compliance stakes are high. In these cases, on-prem AI isn't a compromise or a concession. It's a requirement.

Meeting Insights — And Recognizing the Need for On-Prem AI

Meeting Insights is a brand of AI created by the company AudioCodes. Founded over three decades ago, AudioCodes has deep roots in enterprise voice and collaboration infrastructure, with solutions trusted by IT teams around the world.

Meeting Insights began as a cloud-based platform designed to transform how enterprises manage their spoken communications. By capturing, transcribing, summarizing, and indexing meetings from a wide range of sources, Meeting Insights creates a centralized, searchable repository of institutional knowledge. The platform was built to serve not just individual productivity but enterprise-wide intelligence, offering dashboards, speaker analytics, and integrations with existing IT workflows.

As adoption accelerated, AudioCodes quickly recognized a major limitation. While the cloud model suited many businesses, a significant number of global organizations — particularly in regulated sectors — could not legally or operationally allow their meeting content to be processed outside their own infrastructure.

The company's response was **Meeting Insights On-Prem:** an enterprise-grade, AI-powered meeting capture and intelligence solution that runs entirely inside the customer's firewall.

Unlike consumer-grade AI bots, Meeting Insights isn't tied to a platform or a vendor ecosystem. It works across Teams, Zoom, SIP-based PBXs, and even legacy telephony. It captures meetings in real time or processes recordings afterward.

More importantly, it provides structured summaries, indexed search, keyword spotting, action item extraction, and compliance-ready output — all without ever sending data to the cloud.

Meeting Insights On-Prem: A Secure AI Framework Built for Compliance

AudioCodes Meeting Insights On-Prem is designed for sectors where data security and compliance come first: defense, government, finance, and healthcare. It brings the power of AI to environments that can't risk cloud exposure.

Key capabilities include:

- **Zero Internet Requirement:** The system operates entirely on-premises with no external connectivity. All data — from audio files to summaries — is stored, processed, and encrypted within the organizational infrastructure.
- **Trainable, Customizable SLMs:** Unlike generic LLMs, Meeting Insights On-Prem uses pre-trained Small Language Models that are refined over time using internal data. Corrections made by users help the model adapt to organizational jargon and terminology.
- **UC and Legacy System Integration:** The solution supports audio capture via traditional PBX, SIP, analog systems, or modern unified communications (UC) platforms, such as Microsoft Teams and Zoom. Bots can dial into meetings or capture room audio directly.
- **Compliance-Ready Reporting:** Output templates are configurable to match regulatory requirements. Summaries can include participant names, timestamps, action items, and rubrics formatted to meet documentation standards.
- **Telephony Audio Support:** Unlike cloud tools that rely on VoIP capture or video feeds, this system works with telephony audio, meaning it's compatible with virtually any meeting format.
- **Solves Interop Challenges:** Many platforms still struggle to bridge UC and legacy environments. Meeting Insights is equipped to interface with both.
- **Transcript Universality:** Once meetings are transcribed, the resulting text becomes a flexible asset — usable by any AI tool for search, translation, action item extraction, or summarization. It creates a persistent, structured record that unlocks downstream insights while keeping data securely within the organization's control.

A Neutral Platform in a Locked-In World

Meeting Insights On-Prem offers significant advantages over the fragmented, limited, and sometimes risky alternatives that organizations often rely on to capture meeting intelligence. Compared to embedded features in UC platforms, third-party

transcription tools, or manual note-taking, the On-Prem approach addresses critical gaps in control, coverage, and consistency.

Busting Siloed UC Platform Features. Most video meeting platforms, such as Microsoft Teams, Zoom, and Google Meet, offer built-in transcription and AI-powered summaries. But these features are typically available only to meeting organizers or attendees, and often only within a single platform's interface. There's no unified repository of insights, no multi-platform support, and no enterprise-level governance. Meeting Insights On-Prem transcends these limitations by ingesting audio and video from multiple sources — SIP, PSTN, UC platforms — and creating consistent outputs in a single, secure environment.

Cloud-Based Overlays: Flexible but Risky. Some services promise platform-agnostic transcription via cloud upload or browser overlays. While these tools may improve coverage, they introduce significant concerns around compliance and data handling. Every audio file must leave the corporate perimeter, creating potential exposure risks and jurisdictional complexity. Meeting Insights On-Prem avoids this risk by keeping data fully inside the enterprise network, eliminating external dependencies while maintaining full AI functionality.

Manual Notes and Recordings: Incomplete and Inconsistent. In the absence of automation, many organizations rely on meeting recordings and human note-taking. These approaches are time-consuming, error-prone, and lacking in scalability. Delays in transcription or interpretation often result in key details being lost or misunderstood. Meeting Insights On-Prem automates capture, indexing, and summarization — ensuring that no critical point is overlooked, and context is preserved in real-time.

No Capture: A Missed Opportunity. Meeting Insights On-Prem doesn't just replace one tool — it replaces a patchwork of mismatched approaches with a secure, consistent, enterprise-grade platform.

About AudioCodes

AudioCodes isn't always the loudest voice in the collaboration market, but it's one of the most experienced, capable, and consistent, operating since 1993. With a decades-long history in voice communications infrastructure, the company built its reputation delivering enterprise-grade Session Border Controllers (SBCs), voice gateways, and Microsoft Teams integrations — earning the trust of IT leaders across regulated industries.

WHAT'S AN SLM?

Small Language Models (SLMs) are compact AI engines trained on specific, curated datasets. Unlike Large Language Models (LLMs), which are broad, cloud-heavy, and general-purpose, SLMs can run on-prem and adapt to the organization's own language, policies, and workflows. That makes them ideal for secure, local applications such as meeting transcription and compliance reporting, without the cloud overhead or risk.

That track record matters. In a world where many AI features are being rushed to market, AudioCodes has taken a deliberate, standards-aligned approach, focusing on compatibility, security, and long-term viability. Its Meeting Insights platform reflects that ethos. Initially developed in response to the needs of high-security environments, the solution has matured into a tried-and-tested, enterprise-ready product that addresses real-world requirements, including on-premises deployment, multi-platform capture, and compliance-grade outputs.

AudioCodes also brings an unusually broad perspective to the meeting space. It's one of the few companies delivering a full stack of solutions — phones, video room systems, Session Border Controllers (SBCs), and management tools — all designed to work in heterogeneous environments. That means Meeting Insights isn't an isolated tool; it's part of a larger vision for integrated, secure communications.

Just as importantly, AudioCodes has shown a willingness to adapt ahead of the market. The company was one of the first to align its Android-based room systems with Microsoft's new Device Ecosystem Platform (MDEP), even contributing code to accelerate the shift. That collaboration-first posture, combined with its deep infrastructure knowledge, positions AudioCodes especially well to lead in emerging areas like secure, on-premises AI.

In short, while others focus on flash, AudioCodes is focused on fundamentals: getting it right, keeping it secure, and helping organizations unlock real value from their communications investments.

The Right Tool for the Real World

From an analyst's perspective, Meeting Insights On-Prem is more than a compliance play. It signals where secure enterprise AI is actually going.

We've spent the last decade talking about agility, flexibility, and scale, all framed through the lens of public cloud infrastructure. But that narrative is shifting. The AI era intensifies the need for secure and sovereign data handling. As organizations layer AI on top of core communications, every spoken word becomes a data asset. In this environment, control becomes a strategic imperative.

Meeting Insights On-Prem offers something the collaboration market has largely overlooked: a comprehensive solution for transforming meetings into actionable insights without compromising compliance, confidentiality, or context. It's not just AI that lives behind the firewall. It's insight that stays connected to business outcomes, and disconnected from public LLMs.

What makes this particularly important is that the collaboration and communications stack has never been more fragmented. Most vendors still

treat intelligence as a feature embedded in their own platform, not an asset that needs to span all systems. AudioCodes has stepped outside that frame to offer a neutral solution that brings AI to every conversation, regardless of vendor, network, or environment.

This is about enabling real transformation in places that can't afford to compromise. In a world where AI is often seen as inherently cloud-based, Meeting Insights On-Prem proves that intelligent systems can be just as effective — and far more secure — when deployed at the edge.

Government agencies, financial institutions, healthcare providers, and global firms managing sensitive IP need trusted solutions that fit the real-world constraints of their business, and AudioCodes is delivering exactly that.

In an era where "secure AI" often feels like an oxymoron, this product reminds us that it doesn't have to be.

David J. Danto has had over four decades of developing and delivering successful business analyses, strategies and outcomes, serving in technology leadership roles with multiple firms. He has been honored by many industry organizations and publications that range from general technology to traditional AV to enterprise communications and collaboration. David also has a significant industry following that read his articles and posts and watch and listen to his multiple podcasts. Today, he is the principal analyst with TalkingPointz and is also the non-profit IMCCA's director of emerging technology.

Dave Michels is founder and principal analyst at TalkingPointz. TalkingPointz offers research and analysis on enterprise communications. Dave has over 30 years of experience in telecommunications and unified communications, and is regular contributor to industry sites and conferences. Dave holds graduate degree in Telecommunications.