AI on Set: Understanding the tech stack and infrastructure requirements of the entertainment industry

Altman Solon is the largest global strategy consulting firm exclusively working in the TMT sectors. In this insight, we examine findings from our survey of 335 production executives at Media & Entertainment companies worldwide to better understand their AI (Artificial Intelligence) tech stack and infrastructure requirements. Results show that AI tooling is heavily used in video and audio editing, and investments are expected to increase. While infrastructure is primarily addressed on-premises or through colocation, hyperscale data centers can position themselves as an essential component of the AI tech stack by providing secure, scalable solutions.

This summer, while Lucasfilm touted the AI software that de-aged Harrison Ford's face by 40 years for Indiana Jones and the Dial of Destiny, the Writers Guild of America and the Screen Actors Guild joined forces in a historic strike, aimed in part at curbing the use of artificial intelligence in screenwriting and digital performances.

At a time when streaming giants are tightening their belts, AI can create cost savings and efficiencies by automating labor-intensive tasks and optimizing workflows. While AI's place in the entertainment sector remains subject to fierce debate, the global market for AI in the industry is expected to grow from $10.88 billion in 2021 to $99.48 billion by 2030, with a compound annual growth rate (CAGR) of 26.9%.

Altman Solon surveyed 335 production executives from the U.S., Canada, the E.U., and the U.K. to better understand how and why Media & Entertainment companies use AI and Machine Learning (ML) software today. Findings show that video and audio productions primarily use AI tools for caption and transcription generation, editing, and even narration, lowering costs and increasing productivity. Production and post-production companies have indicated that their AI tech stack—data, models, and deployment capabilities—is housed mostly on-premises or in colocation facilities. However, hyperscale data centers have an opportunity to capture market share as entertainment companies plan on further investing in AI technology, given the data storage requirements, compute capacity for processing needs, and technical requirements to secure the environment. With major players like Microsoft Azure signing long-term partnerships with OpenAI, and AWS adding Anthropic, Stable Diffusion, and Cohere to their libraries, it seems inevitable that entertainment companies will see data centers as a critical component for their AI needs.

altman solon

AI tools are in heavy use in video and audio editing, as respondents turn to AI for cost savings

Post-production activities—including editing, creating visual effects, sound design and mixing, and color grading—historically required the work of highly paid experts who worked on short deadlines and produced low margins for production houses. AI tools can make these processes more efficient while maintaining high-quality production standards.

Our survey data found that over 7 out of 10 (71%) respondents have leveraged various AI capabilities for video editing. Some AI-enhanced video editing tools use computer vision to track objects, analyze footage to find the best shots, add visual effects, and automate color correction, among other use cases.
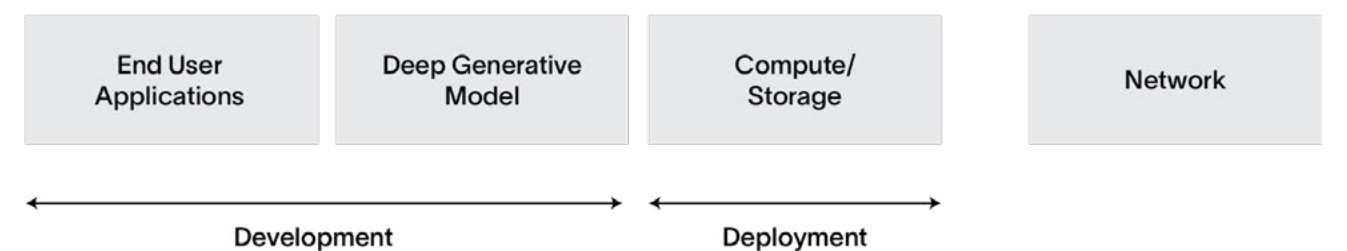
Audio editing is the second most popular artificial intelligence production use case, with 58% of respondents reporting implementation. AI-powered audio editing tools can automatically remove background noise, enhance voice clarity, or analyze audio to remove unwanted sounds.

Tied for third place are visual digital asset creation and content localization, with 39% of respondents reporting AI and ML assistance for these two use cases. Close behind, 32% of respondents say they use AI to assist in scriptwriting.

The downstream effects of studio budget cuts are being felt across the content supply chain, with even smaller service providers and vendors trying to find ways to be more cost-efficient. Indeed, most respondents (60%) cite "cost savings" as a motivator for adopting AI tools, and nearly as many are influenced by "operational efficiencies" provided by these tools.

The AI Technology Stack

| End User Applications | Deep Generative Model | Compute/Storage | Network |

Development — Deployment

altman solon

On-premises and colocation are favored for infrastructure hosting, but future growth will likely be in hyperscale data centers

AI tools often require a more complex tech stack than typical Software-as-a-Service (SaaS) business tools. IT decision-makers have to choose a development type—including the AI application itself and the model that powers it—as well as how they want to run and store (or deploy) the application. Deployment types can include:

- **On-premises facilities**: Computing infrastructure and resources that are owned, operated, and maintained by an organization within its physical premises.

- **Colocation facilities**: physical space, power, cooling, and networking resources for businesses to host their own servers and IT equipment within a shared data center environment.

- **Public cloud – hyperscale data centers**: distributed data centers maintained by cloud providers to deliver scalable compute, storage, and networking resources over the internet.

- **Edge data centers**: smaller facilities located closer to end-users and devices, designed to reduce latency and support real-time data processing.

- **Hybrid private/public**: combined private on-premises infrastructure and public cloud services to optimize performance, scalability, and flexibility.

Our findings show that most respondents use more than one infrastructure environment. Investments in infrastructure depend on the level of investment a specific company makes in developing or deploying AI solutions. It should be noted that in-house solutions remain the norm in the entertainment industry. Sixty-nine percent of survey respondents report using enterprise, on-premises infrastructure to host their AI tools, followed by colocation facilities (61%). Hyperscale data centers rank third, with 52% of respondents using these to support AI tooling.

Given the challenges with moving large data sets that are already hosted in a colocation facility, most media companies are not opting to move these, but rather maintain on existing infrastructure. Oftentimes, media companies seek control over the tech stack, establishing a clear understanding for process structures, transparency into capabilities, and limitations of the solutions. The industry is much more accustomed to managing the adoption of these new and emerging technology capabilities given their heightened focus on operational performance, scalability, and data security.

Data security and control were often cited as the main reasons for managing capabilities through on-premises or colocation. Contrary to data centers, on-premises solutions provide total control over hardware, security, and set-up. These are especially valuable to media companies as they leverage IP to experiment with new AI processes. According to a Virtual Screen Producer at a video production studio, "If you are a large tech company and IT security is of utmost concern, then it is important to bring everything in-house to retain everything on-premises."

On-premises solutions may also be cheaper and easier to set up in the short term. The same can be true for colocation. When examining decision criteria for AI infrastructure, 80% of respondents look to "cost savings," followed by "scalability" (69%) and "speed" (53%).

altman solon

Even though on-premises and colocation facilities are in higher use among respondents, the future of AI infrastructure management may include more High-Performance Computing (HPC) environments. HPC, which is capable of processing large amounts of data and performing complex calculations at high speed, is well suited for AI which requires large data sets and complex models.

According to our survey, 95% of respondents strongly or somewhat agree with the statement, "The demand for specialized computing environments called High-Performance Computing clusters will be high and continue to grow as more media companies train and deploy their AI solutions."

Respondents who rely on data centers for their AI infrastructure cite reduced energy costs (typically power consumption and activating cooling systems), greater storage capacity, and heightened incident response capabilities as their top priorities.

As AI investments grow, data centers have a strategic role to play

As use cases for AI in entertainment production grow, so does investment in the technology. Among our respondents, AI investments varied widely, spanning from $1 million to $50 million. However, most expect to increase spending on AI in the near future. A substantial 41% of respondents anticipate growth rates ranging from 11% to 20% in AI investments over the next five years. Data center providers should expect to see an uptick in demand powered by AI investments. To address this demand, data centers should not only invest in supporting higher internet exchange (IX) densities, but also forge partnerships with new, strategic customers, including media companies. **In developing a go-to-market strategy for the media and entertainment sectors, data centers should lead with cost savings and scalability of services. Colocation facilities can also position themselves as a compromise between the benefits and risks of hyperscalers and on-premises infrastructure.**

In the near term, data center providers should prioritize:

1. **Scalability**: Getting more efficient and having the capacity to accommodate the demand. Data centers should focus on the sourcing and procurement of critical hardware such as TPUs and GPUs in anticipation of increasing demand for compute services.

2. **Security and compliance**: Given the large data models used to support the learning process of these solutions, clear data security protocols and compliance measures are essential. Stemming from security breaches at a few large studios, media companies prioritize data security and compliance when evaluating and deploying new technologies, and threats across the media landscape continue to evolve.

3. **Optimizing data operations**: Managing latency, moving data, and maintaining the infrastructure performance (energy consumption and cooling management).

The use cases for AI in the entertainment industry are plentiful, spanning primary research, content creation, post-production, marketing, and more. While the future of AI is still uncertain, companies in the entertainment industry are leveraging this technology to improve operations and lower costs. As IT decision-makers re-evaluate their tech stack, data centers have opportunities to provide low-latency networks and high-capacity capabilities. To best anticipate the needs of the media and entertainment industry, data center providers should focus on developing a secure, scalable solution to an industry that will most likely be ramping up their AI investments.

Leadership
& Oversight

Derek Powell

Director, Los Angeles
derek.powell@altmansolon.com

altman solon