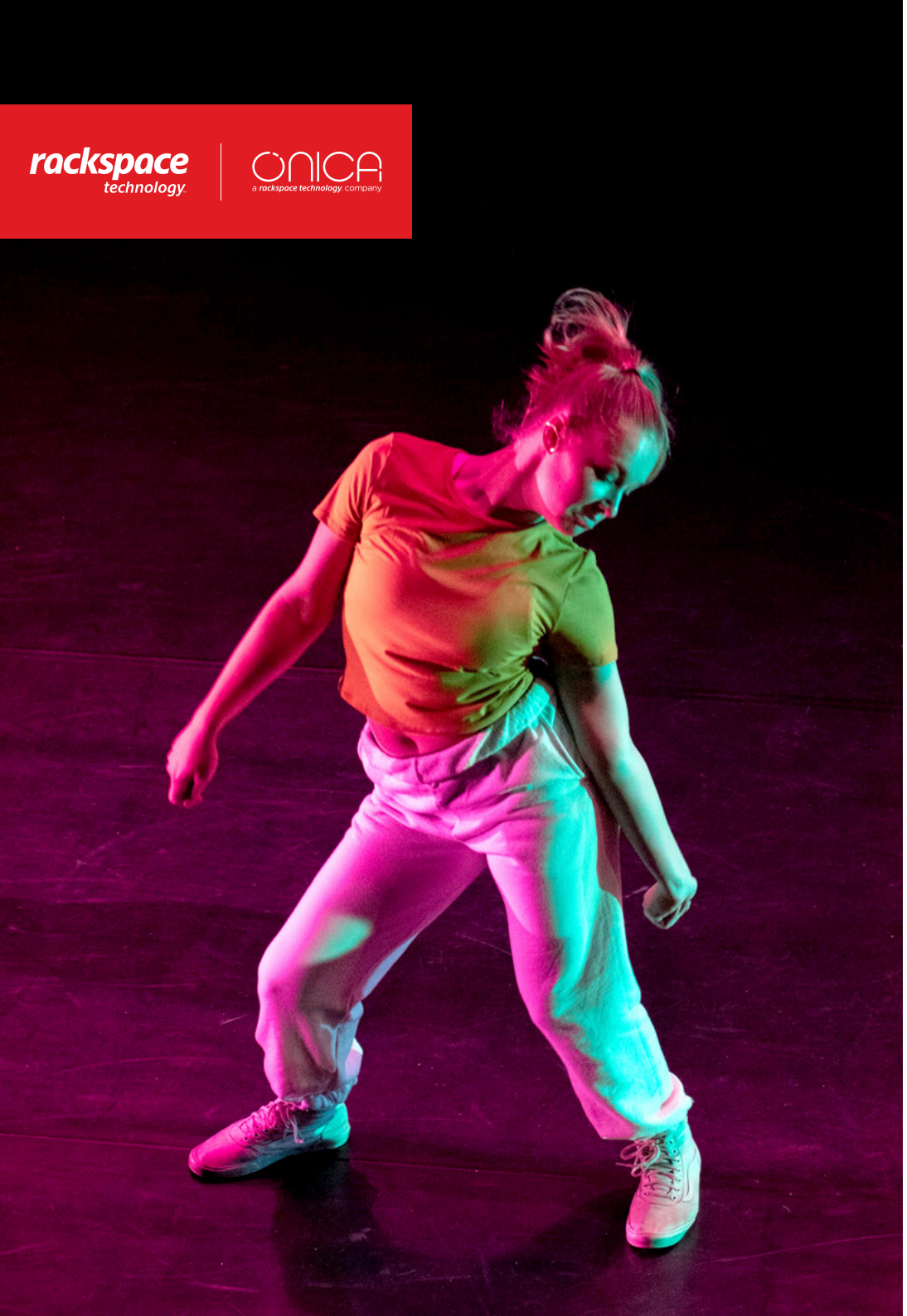

Customer Case Study – Application Development

Humen.Ai dances its way to 70% savings.

Through the extreme integration of serverless and machine learning technologies, the Sway: Magic Dance app makes anyone a great dancer, with video to prove it.

Our customer

Developers of the popular iOS app, Sway: Magic Dance, Humen.Ai is an AI-focused video synthesis and content creation company that uses deep learning and AI to create personalized, interactive experiences.

The obstacles they faced

The infrastructure running Humen.Ai's popular iOS app became burdensome to run and threatened its financial stability. The company needed to lay a foundation for rapid releases and quickly reduce infrastructure costs without impacting response times.

How we helped

Amazon ECS, Amazon ECR, Spot ASG (containers on ECS), AWS Step Functions, AWS Lambda (serverless).

What we achieved together

In just six weeks, Onica (a Rackspace Technology company) replaced Humen.Ai's hand-built environment with Amazon ECS with Spot Instances, potentially reducing infrastructure costs by 70%. Employing a unique, Onica-built technique for scheduling jobs, the application runs faster and more efficiently. The new lightweight, containerized infrastructure enables AI and production teams to get app enhancements to market faster.

HUMEN

> *"We were definitely on the path of switching to a containerized pipeline. Onica put in substantial effort to make that happen for us quickly."*
>
> Tinghui Zhou
> Co-founder and CEO, Humen.Ai

Application innovation using AI

Humen.Ai is an AI-driven, synthetic media company. Its popular iOS app, Sway: Magic Dance, is based on AI modeling. The fun-to-use app generates videos of users based on a source video uploaded by the user doing basic motions, like moving around, kicking their legs or waving their arms, for a few seconds. Its PyTorch-backed proprietary machine learning (ML) model creates a digital skeleton of the user. Using that skeleton, users can generate a new, photo-realistic stunt double of themselves dancing like Michael Jackson, twirling like a ballet dancer, or making karate moves like a Black Belt in 30 seconds.

The big dance reveals big problems

To operate, Sway: Magic Dance relied on 400 AWS G4 on-demand instances controlled by a complex and expensive-to-maintain system that was built internally. During the 2020 Superbowl, the app was launched in partnership with Doritos. It hit number two in the Apple App Store and, with help from AWS, Humen.Ai was able to scale up to serve the traffic spike.

The successful launch uncovered scalability and efficiency issues in its backend. The AI app's concept was built on its ability to quickly train models for each customer and required an immense amount of on-demand compute power, as is common in any AI-driven application. "One thing that we were pretty concerned about was the cost of compute in the backend," said Tinghui Zhou, Co-founder and CEO of Humen.Ai. "Our infrastructure is driven by machine learning, which requires the usage of graphics processing units (GPUs) for model training and inference on the cloud."

In addition to the infrastructure issues, Humen.Ai also faced another common challenge in AI operations. Much like the age-old friction between software developers and engineers, data scientists often face the same friction in moving AI projects from concept through production. AI teams operate much like an R&D team. They are abstracted from engineering once the model is ready to be operationalized. This puts the burden on engineering to figure out architecture design, resource management and model monitoring to run it efficiently, securely and reliably. This friction can slow down releases and hinder the pace of innovation. Humen.Ai wanted to remove engineering barriers to shorten the time between building products and

enhancements and releasing them to users while keeping its small, startup team agile.

Addressing containerization obstacles

The Humen.Ai team was extremely proficient in AI/machine learning operations and in AWS. It had tried a few pathways to optimize infrastructure, like containerization, but those didn't work out. Amazon SageMaker would have been a great option for model training and inference, as it supports running on Spot Instances, a less expensive route compared to on-demand instances. The tradeoff is needing to wait for AWS to allocate capacity to run spare training, which would hinder the end user experience. Users needed to quickly upload video, have it immediately processed and be able to generate Instagram filters in minutes.

As part of the Jumpstart program, AWS referred Humen.Ai to Onica for containerization support. The Jumpstart program provides organizations with low-cost infrastructure, credits and training to support growth. Working with Onica, they were able to get to the bottom of the containerization issues. The Onica team helped Humen.Ai containerize its entire AI app to effortlessly deploy on Amazon ECS with Spot Instances.

The Onica team then took it a step further by delivering a custom solution built on AWS. Devising a unique method to schedule Amazon ECS containers, Onica empowered Humen.Ai's machines to do more AI tasks at the same time. This method, combined with managed services for ECS, enabled it to complete near-real-time training, inference, pre-processing and post-processing using higher-density machines. "We were definitely on the path of switching to a containerized pipeline. Onica put in substantial effort to make that happen for us quickly," explained Zhou.

Unique solution exceeds expectations

Combining this extreme software and hardware integration with the use of Spot Instances has resulted in a potential 70% cost reduction — well beyond the expected goal of a 30% reduction. The Humen.Ai team was so enthusiastic about the new infrastructure design that they began building around the project before it was complete. And the Onica team was able to keep up with the speed of the small, agile team moving from proof of concept to production in just six weeks. Humen.Ai is now leveraging the new lightweight, efficient infrastructure and newfound agility to create more products and reinvest in technology.

Previously, Humen.Ai managed hundreds of instances. By moving to Amazon ECS with Spot Instances, the infrastructure is now easier to manage and operates at lower cost. The AI startup has been able to replace a whole host of infrastructure with managed services. Onica created a lightweight AI infrastructure that allowed Humen.Ai to manage everything using only serverless technologies. This takes a huge burden off of the engineering team and allows the AI team to quickly innovate and efficiently deploy to production.

Expertise provides the foundation for long-term growth

According to Zhou, "We've doubled, maybe even tripled, our throughput of processing demands from our users." Instances are better utilized, allowing the machines to run more efficiently. This allows the app to handle more concurrent user requests. "I think that definitely had an impact on the user experience," Zhou continued.

With a very small team, Humen.Ai is now able to build instead of just managing what they've already built. It has been able to reduce its technical debt to a level where it's able to move forward. "We were able to bring down costs significantly and it really improved our back-end efficiency," Zhou notes as the biggest outcome of his engagement with Onica. The cost reductions were critical in ensuring the early stage startup's long-term viability by conserving cash flow.

Reflecting on the experience with Onica, Zhou said, "Working with a professional team of engineers from Onica on optimizing the back-end was overall a very positive experience and really helped us scale our infrastructure during the critical, early stages when we don't have a big, back-end engineering team."

With the new infrastructure design in place, Zhou plans to expand Humen.Ai's dance videos to sports, TV, gaming and movies. Additionally, the team wants to provide a way for users to share their created content within the app for collaboration. Humen.Ai is also trying to improve the photorealism of its output with 3D-based perception for its skeleton and scene-analysis ML models.

About Rackspace Technology

Rackspace Technology is the multicloud solutions expert. We combine our expertise with the world's leading technologies — across applications, data and security — to deliver end-to-end solutions. We have a proven record of advising customers based on their business challenges, designing solutions that scale, building and managing those solutions, and optimizing returns into the future.

As a global, multicloud technology services pioneer, we deliver innovative capabilities of the cloud to help customers build new revenue streams, increase efficiency and create incredible experiences. Named a best place to work, year after year according to Fortune, Forbes, and Glassdoor, we attract and develop world-class talent to deliver the best expertise to our customers. Everything we do is wrapped in our obsession with our customers' success — our Fanatical Experience™ — so they can work faster, smarter and stay ahead of what's next.

Learn more at **www.rackspace.com** or call **1-800-961-2888**.

November 19, 2020 / Rackspace-Case-Study-HumenAi-AWS-TSK-2848