SUCCESS STORY

Ava's innovative live captioning software, powered by Elastic Observability makes work, school, and daily life accessible for deaf and hard-of-hearing people

Ava uses Elastic Observability to maintain real-time, uninterrupted performance of its live captioning software used by 150,000 people to participate in meetings, video calls, and everyday conversations.

Region
United States

Industry
Telecommunications

Solution
Elastic Observability

Reduces errors in new software features

With Elastic Observability, Ava engineers can test and evaluate performance early in the development process, reducing errors when the software goes live.

Accelerates the launch of innovative features

Engineers spend less time analyzing and fixing performance issues and can concentrate on the development of innovative live captioning features with Elastic.

Reduces cloud computing costs

With Elastic Observability, Ava can quickly identify and address performance bottlenecks and achieve cost-effective scalability in its cloud computing platform.

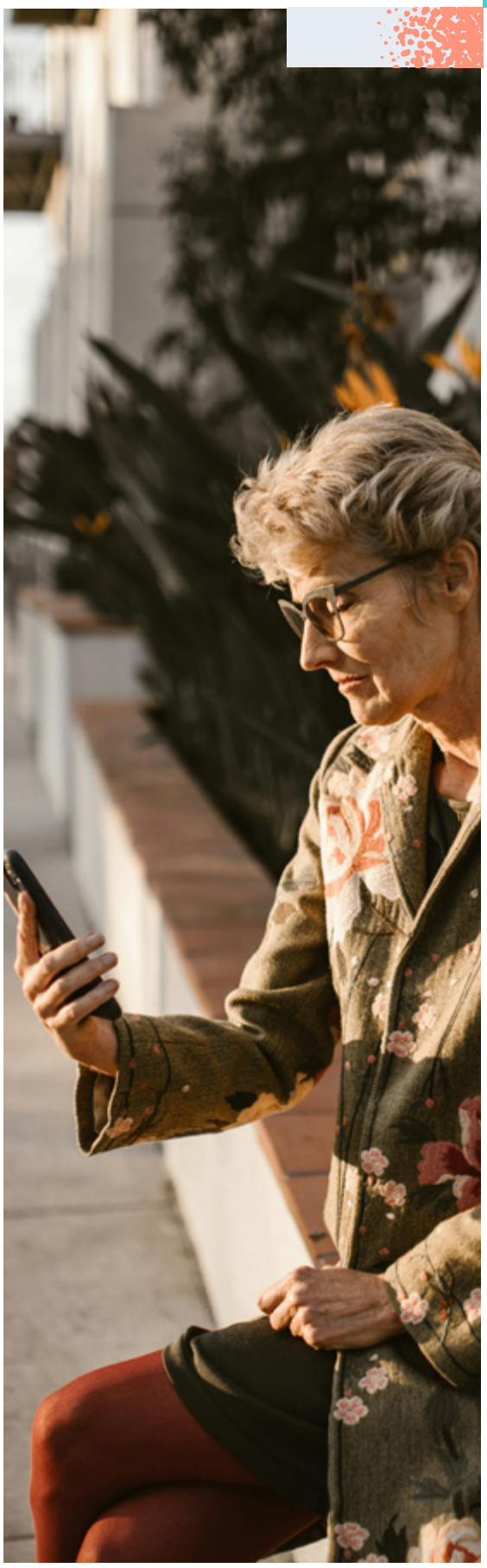
 © 2024 Elasticsearch B.V. All rights reserved.

Every day, countless people are excluded from conversations because they cannot hear what others are saying clearly. Ava was founded to enhance accessibility for deaf and hard-of-hearing people. Its products include a web and desktop app that enables users to read real-time closed captions in online meetings or videos, a mobile app that captions live conversations, and a real-time captioning service that combines machine and human intelligence.

Like any startup, Ava must innovate fast to stay ahead of larger competitors, especially in recent years, as the number of transcription services has increased sharply. Providing a stable, enjoyable user experience is also a priority. Juan Pablo Ospina, Platform Engineer at Ava, says, “Our biggest challenge is providing the best possible experience to our users. But we lacked a comprehensive observability solution to track application performance, monitor issues proactively, and rectify problems quickly.”

This is a weighty task for an application that must transcribe speech to captions in real time. Any deceleration in performance, even a few milliseconds, diminishes the end-user experience. When this happened, the Ava team faced a difficult choice: add CPU capacity or try to identify the source of the performance issue as quickly as possible.

“Neither situation was satisfactory,” says Ospina. “Adding CPUs was expensive while persisting with sub-optimal performance while diagnosing the issue risked alienating our customers. This led us to select Elastic Observability and replace our previous solution that couldn’t deliver the visibility and accuracy we required.”



 | © 2024 Elasticsearch B.V. All rights reserved.

Cutting the cost of cloud computing

Ava now runs Elastic Observability on Elastic Cloud with AWS as its primary cloud computing provider. This configuration monitors core activities, including a service dedicated to AI tasks and Ava's ASR automated speech recognition (ASR) system. Elastic ingests 50 million events per day. "Currently, 80% of our internal services are covered by Elastic Observability, and we hope we can move to 100% in the near future," says Ospina.

Ospina says that Elastic makes it simpler to verify different types of information, including bottlenecks that slow down application performance. "We collect trace data from our applications using an Elastic APM agent and create powerful visualizations, enabling us to identify the root cause of any holdups and optimize the code base to solve the issue."

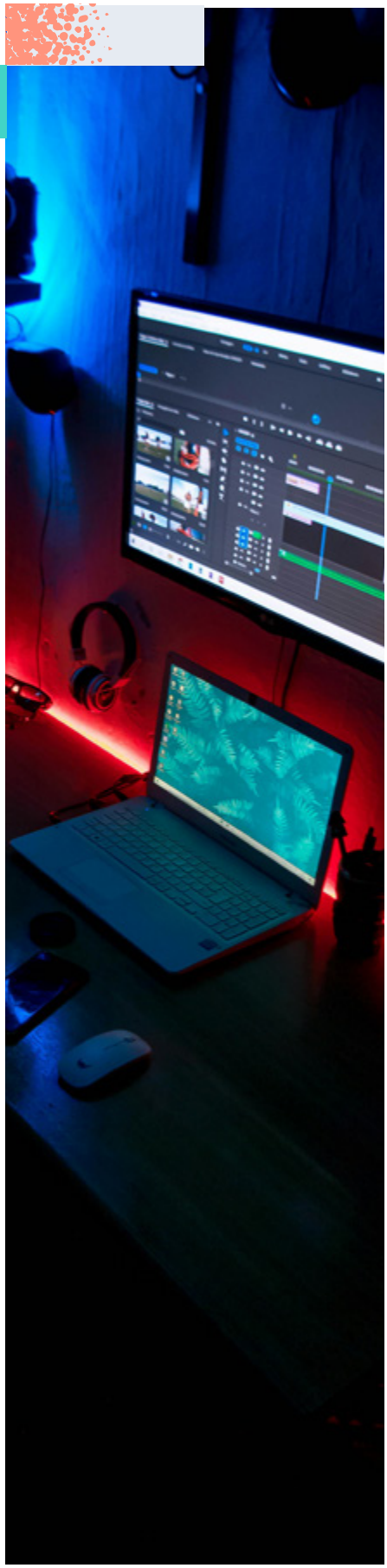
Most recently, Ava has adopted Elastic Universal Profiling, a recent addition to the Elastic Observability solution that collects application, infrastructure and services performance data without needing code instrumentation or recompilation. Universal Profiling measures code efficiency in three dimensions: CPU utilization, CO2, and Cloud Cost.

> By correlating data from different sources, such as traces, logs, profiles, and metrics, we gain a holistic view of our applications and can identify the specific components contributing to performance issues.

Juan Pablo Ospina
Platform Engineer
Engineering, Ava

 | © 2024 Elasticsearch B.V. All rights reserved.

elastic



For Ava, this means being able to streamline code efficiently, optimize hardware and infrastructure, and fine-tune services. It also improves financial management and capacity planning. “In the past, we would scale up to hundreds of instances, which had a significant impact on cloud computing budgets,” says Ospina. “With Elastic, we can quickly identify and fix the issue and achieve cost-effective scalability.”

> Elastic Universal Profiling provides us with clear visibility into our applications and infrastructure. We can not only troubleshoot performance issues in production, but also understand the impact of new features and deliver an optimal user experience.

Juan Pablo Ospina
Platform Engineer Engineering, Ava

Becoming an early adopter of Elastic Universal Profiling also led to a shift in the relationship between Ava and Elastic. “When we subscribed to the beta program, there was a lot of support and communication from the Elastic Profiling team. They did a great job taking our feedback and adding features that we needed, including adding support for V8, NodeJS/V8 symbols, and automatic symbolisation of Amazon Linux kernel frames that gave us a full picture of what’s happening in our applications.” Daan Wynen, AI Engineering Lead at Ava, was also impressed with Elastic Universal Profiling’s ability to reduce costs for Ava.

> Within days of rolling out Elastic Universal Profiling, we identified a new hot code path in a feature before release and could fix it before it made our cloud cost skyrocket.

Daan Wynen
AI Engineer Engineering, Ava

 | © 2024 Elasticsearch B.V. All rights reserved.

Faster time-to-market with innovative features

Greater stability also keeps Ava on the fast track when adding new features. Ospina says, “Instead of firefighting in our production environment, we can dedicate resources to improving the user experience.” In addition to improved stability, Wynen says, “We know exactly which parts of the code will yield the biggest improvements.”

Ospina gives the example of a feature called ‘diarization,’ which enables the software to detect the number of speakers in a conversation. If you are using the app on your phone, it will detect how many people are in the room and identify each speaker by highlighting text in different colors.

Ospina and his team are also exploring new Elastic AI features. “We’re keen to see how vector search can be used with LLMs to provide transcription summaries to clients.”

Meanwhile, he remains focused on helping the business concentrate on its primary mission. “Our goal is to provide the best possible live-captioning and transcription experience to our customers. With Elastic Observability, we can continue to deliver first-class performance, add innovative features, and stay ahead of the competition.”

Get started with Elastic Universal Profiling today and unlock deep, whole-system visibility into your code’s behavior without instrumentation, code changes, or service restarts. Just deploy the agent and begin optimizing.

See for yourself how your business can benefit from Elastic in the Cloud, with a free 14 day trial.

Get started

 © 2024 Elasticsearch B.V. All rights reserved.