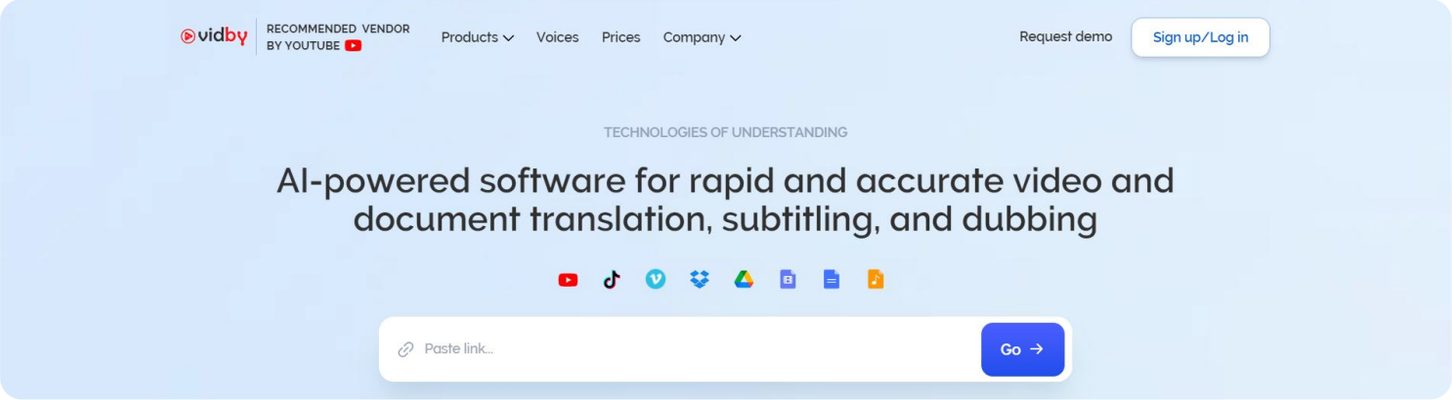
Secure Yet Agile: Scaling AI-Translation Startup Through Kubernetes and Google Cloud

Imagine being able to communicate easily and instantly in multiple languages. It's possible with Vidby, an AI-powered translation app. However, its multiple services require fast and dependable performance to enable seamless global interactions. Through a partnership with IT Outposts, the startup has established the technical foundations to rapidly scale real-time translations worldwide.

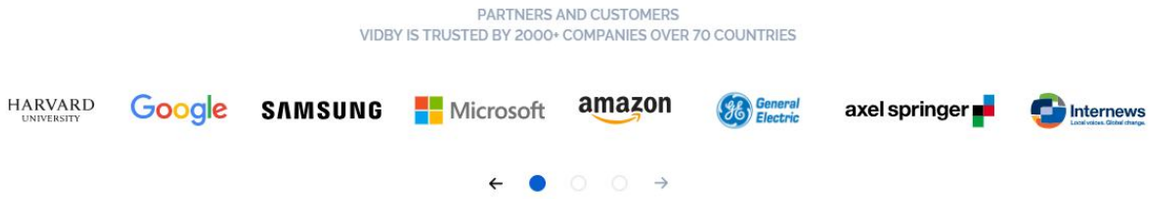


Project Description

Vidby utilizes artificial intelligence to provide fast and accurate translation, subtitling, and dubbing of videos, as well as document translation. The idea was inspired by the founder's firsthand difficulties conducting international calls with clients who spoke different languages.

Initially dependent on human linguists, Vidby recognized the immense potential of AI to interpret languages swiftly at scale. However, previously rigid infrastructure with manual code releases would slow their ability to continuously refine models, causing delays that risked the company's credibility.

With the vision relying on responsive translations, agile iteration was imperative to frequently enhance quality. IT Outposts has implemented and continues to refine a sturdy tech base to achieve project goals day-to-day.



Upload files for instant translation on our Telegram bot

Open Telegram

GET STARTED

Start sharing your words with the whole world

Start Now

Key DevOps Metrics

< 25	10 min	1%	1-2 min
Frequency of Deployments per month	Average Lead Time for Changes	Change Failure Rate	Time to Restore Service

Provided Services

DevOps Services

- Infrastructure and architecture design
- CI/CD automation
- Integration services

Kubernetes managed services

- Efficient scaling
- Improved reliability
- Faster deployments

SRE services

- Log management and monitoring
- Incident management
- Release management
- Performance optimization

Operations managed services

- Technical support
- Infrastructure maintenance
- Capacity planning
- Disaster recovery as a service

Cloud adoption

- Cloud infrastructure management
- Cloud assessment

Work agenda

Client

Technology for Understanding, AI-powered translation provider

vidby.com

Location

Switzerland

Technical team

Lead DevOps engineer

2 DevOps engineers

More DevOps engineers are added to the team as the project scales

Project timeframe

2021 - ongoing

Budget

200,000

Project goals

- Migrate infrastructure to Kubernetes on Google Cloud to enable faster, more agile development and deployments
- Implement CI/CD pipelines to automate builds, testing, and deployments
- Standardize processes and align disparate teams through documentation, knowledge sharing, and access controls
- Provide isolated, optimized infrastructure for resource-intensive video processing workloads
- Establish monitoring, alerting, and analytics to improve observability and reliability
- Scale up services easily without downtime to meet growing demand

Challenges

01 Manual infrastructure hampering agility

When Vidby first launched, its services were hosted on basic virtual machines through the Hetzner cloud platform. This initial setup required developers to manually clone code for each release.

Additionally, there was no CI/CD workflow established. As a result, the release process was inefficient. The manual efforts around routine tasks consumed an excessive amount of developers' time. It also led to considerable downtime for the services during updates.

02 Aligning disparate development teams

Vidby structures its development across multiple outsourced teams, with each team concentrating on distinct services such as text translation and lip-syncing. Although the technology stack had been modernized to leverage Google Cloud and Kubernetes, managing frequent release cycles with scattered teams remained difficult.

03 Managing security and access control

Vidby's reliance on multiple external developer teams accessing services on Kubernetes created potential security risks and service disruptions. Teams could unintentionally interfere with each other's work.

Moreover, Vidby leverages numerous Google Cloud Platform services, such as the Translation API and Cloud Storage Buckets. This required broadly granting service account credentials to enable access for the various development teams. However, widely sharing credentials didn't align with security best practices.

04 Providing resource-intensive video-processing

Many of Vidby's AI services involve video processing, necessitating compute nodes equipped with GPUs.

05 Performance monitoring and alerting

Microservices and distributed architecture introduce complexity when it comes to monitoring and managing application performance. Understanding how these services interact and where bottlenecks occur is difficult without effective application performance monitoring and alerting strategies in place.

Contacts

Expand your business horizons with cutting-edge DevOps and cloud solutions from IT Outposts. Offering comprehensive support from infrastructure design to cloud service management and performance optimization, we help companies like Vidby achieve global success. Join us for a reliable, scalable, and efficient IT solution.

Start your project

CEO Vidby

Alexander Konovalov

> The most impressive factor is their level of responsibility and commitment to the project goals

Read full review

Solutions

1. Migration to Kubernetes and Google Cloud

We chose Kubernetes on Google Cloud to reduce manual work and decrease downtime. Using Terraform scripts, we built the entire Google infrastructure, including Kubernetes clusters with optimized node pools.

Previously relying on GitHub Actions for CI, Vidby was migrated to Cloud Build. And with the CI/CD pipelines now in place, zero-downtime rolling updates can occur by leveraging health checks and traffic shifting. As new pods pass the ready state, Kubernetes seamlessly terminates old pods and routes traffic to the new versions.

2. Uniting teams through the kick-off file and knowledge-sharing

For access management, development teams are separated on GitHub with permissions granted only to their required repositories. In addition, we provide each new developer with a standardized kick-off file. The checklist covers key items like Docker usage and configurations, application ports, Git repositories, sensitive data handling, logging locations and formats, compute resource requirements (CPU/RAM), and permission policies.

We also hold on-demand knowledge-sharing sessions on Kubernetes and cloud-native best practices. These meetings help improve skills and align processes across teams.

3. Namespace-based access control and implicit authentication

We leverage Kubernetes namespaces to implement access controls and logical isolation of resources. Namespaces separate pods, containers, and services on a per-application basis. For instance, developers working on application A have access strictly to resources within namespace A.

To avoid unnecessarily exposing credentials, we leverage Google's Workload Identity. This allows us to associate each container with a Google service account for implicit authentication.

4. A dedicated node pool for video processing

To meet the demands of resource-heavy video processing, we enabled a separate node pool of similar virtual machine instances tailored for such workloads. This node pool runs VMs installed with the necessary video drivers and libraries. By isolating these video nodes into their own pool, we ensure AI services processing video can access the required resources and run effectively.

5. Observability stack with monitoring, alerting, and analytics

We've set up comprehensive monitoring across our systems to catch problems early. Custom dashboards visualize historical metric data to analyze trends. Alerts proactively notify teams of thresholds so they can address problems before services degrade.

Our monitoring stack consists of Prometheus for metrics storage, Grafana for visual analytics, and Prometheus Alert Manager for threshold-based notifications. Alerts are shared in a dedicated Slack channel for rapid response.

With proactive alerts and data-rich analytics, our engineers can continuously fine-tune performance across services.

Results

Enhanced scalability

Upgraded architecture provides flexibility to add services and scale capacity easily.

Improved productivity

Standardization and knowledge sharing unlocked developer time to focus on core product work instead of operational issues.

Reduced business risk

Hardened security policies and permissions management lowered the chances of data breaches or service disruptions.

Higher system reliability

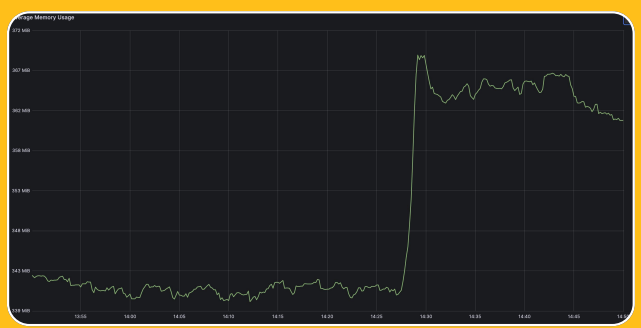
Monitoring and alerting tools reduced downtime and customer impact.

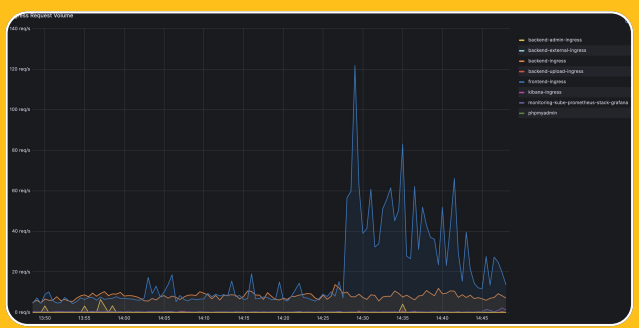
Faster time-to-market

Automation and improved infrastructure enabled Vidby to release new features and updates much quicker.

Improved customer satisfaction

Faster delivery of new capabilities combined with fewer service interruptions delighted end users.





DevOps tech stack

CI/CD

Github
Flux CD
GCP Cloud Build

Monitoring and logging

Prometheus
Grafana
GCP Alerting
GCP Cloud Build

Infrastructure component provisioning

GCP
Docker
Terraform
Kubernetes

Services & databases

Postgresql
MongoDB
RabbitMQ
Redis