Pachyderm

CASE STUDY

RTL Nederlands Relies on Pachyderm's Scalable, Data-Driven Machine Learning Pipeline to Make Broadcast Video Content More Discoverable

RTL Nederlands, part of Europe's largest broadcast group, wanted to use artificial intelligence (AI) to make video content more valuable and discoverable for millions of subscribers. Pachyderm delivered the data-driven machine learning (ML) automation, scale and reproducibility the team needed to work with massive amounts of unstructured video data.

A Modular Approach to Complex AI

RTL Nederlands broadcasts to millions of daily TV viewers, along with delivering streaming content that garners hundreds of millions of monthly views online. Its parent, RTL Group, is Europe's largest broadcaster and part of Bertelsmann, one of the world's largest media conglomerates.

One of the key growth metrics for RTL Nederlands is viewership, but optimizing the value and discoverability of video assets is an extremely labor-intensive endeavor. That makes it ripe for automation, and the team applied machine learning to optimize key aspects of its video platform, like creating thumbnails and trailers, picking the right thumbnail for those trailers, and inserting ad content into video streams. The right thumbnail might not seem crucial but it makes all the difference in the world to whether someone clicks or passes a video by forever.

"We want to use artificial intelligence to make sure we optimally apply our human intelligence, so our teams can be more creative and connected," says Vincent Koops, senior data scientist at RTL Nederlands. "The problem is that it's not easy to apply AI to managing unstructured content; these content operations are computationally complex, making video AI challenging from a data science perspective."

The team solved this problem by breaking complicated tasks into simpler subtasks, eliminating larger task-specific models in favor of an assembly of reusable modules. This modular approach to machine learning allowed the company to train the AI on various elements of the video stream across visual (frame extraction, shot segmentation, facial recognition), audio (tagging, speech identification, musical genre) and text (language detection, key phrases) subtasks.

However, this led to the next big challenge:

> *How to weave these individual ML pipelines together to solve more complex tasks?*

That's when the team turned to Pachyderm.

rtl

Key Benefits

- Automates and orchestrates multiple ML pipelines to solve complex challenges
- Speeds debugging and output analysis through data versioning and immutable lineage
- Scales ML pipelines through parallel processing to handle large volumes of files simultaneously
- Lowers compute and storage costs by only processing new data or incremental changes
- Saves 1,000s of work hours per year through a combination of pipelines

> “We've tried solutions such as Argo Workflows or DVC that delivered some of the same capabilities, but it's cumbersome to juggle different tools. The benefit of Pachyderm is that all of these features are highly coupled in one coherent platform, which works really well for us.”

VINCENT KOOPS
SENIOR DATA SCIENTIST
RTL NEDERLANDS

Pachyderm: the Data Foundation for Machine Learning

Pachyderm provides the data layer that allows machine learning teams to productionize and scale their machine learning lifecycle. With Pachyderm's industry leading data versioning, pipelines and lineage, teams gain data-driven automation, petabyte scalability and end-to-end reproducibility. For RTL Nederlands, Pachyderm was the key to combining and orchestrating the various subtasks into a unified way to process videos at scale. Not only that, but video processing is resource intensive. Pachyderm's incrementality allowed the team to only process new videos as they arrive or change, rather than reprocessing everything from scratch. This delivered tremendous speed to their approach, saving time and money.

"Pachyderm's pipelines are the building blocks that enable us to solve these complex problems," notes Koops. "Effectively applying AI to just one of these tasks can save over a thousand hours a year," he explains.

RTL Nederlands also needed to track metadata and content extracted from video streams to assess and improve the effectiveness of its AI models. Pachyderm's immutable lineage ensured this end-to-end reproducibility. Even more importantly, Pachyderm allows teams to rewind to older versions of the code, data or models, so if a new thumbnail or clip wasn't performing as well as a previous version they could roll back the change to the more high-performance versions.

Lastly, video data is petabyte-scale data, which made Pachyderm's ability to scale to multiple petabytes crucial to meeting RTL Nederlands' goals. Pachyderm's inherent parallel processing and code agnosticism allowed the team to handle a huge volume of video without code changes, while its scalable data versioning optimized storage and computational costs.

> “One of the powerful features of Pachyderm is that data is treated as a first-class citizen. Automation is as simple as adding data to the input repository to trigger pipelines downstream, which creates a nice, traceable way of computing output.”
>
> VINCENT KOOPS
> SENIOR DATA SCIENTIST
> RTL NEDERLANDS

"We've tried solutions such as Argo Workflows or DVC that delivered some of the same capabilities, but it's cumbersome to juggle different tools," Koops says. "The benefit of Pachyderm is that all of these features are highly coupled in one coherent platform, which works really well for us."

Orchestrating ML to Solve Complex Challenges with Pachyderm

With Pachyderm, RTL Nederlands can easily combine various AI models to solve much more complex video challenges, such as identifying the optimum point for ad insertion, or selecting clips for a compelling thumbnail.

Determining ad insertion can be tricky. Business rules dictate ad frequency, but just playing an ad at predetermined intervals will disrupt dialogue or ruin a key scene, substantially degrading the viewing experience. Ideally, the ad should appear between scene and dialogue transitions – something the team at RTL Nederlands has trained its AI to recognize.

In this instance, videos are housed in the cloud on Azure, S3 or a similar service, and imported into a video repository in Pachyderm, where pipelines extract the information necessary to detect an ideal ad insertion point, staying in line with business rules about ad frequency, content restrictions, and more. Boundary detection determines shot transitions with high probability based on frame-to-frame changes in color histograms. Finally, speech detection ensures a break in the dialogue so that speakers aren't cut off mid-sentence by an unexpected ad. Combined together, all these models allow automated ad insertion without degrading the viewing experience.

"One of the powerful features of Pachyderm is that data is treated as a first-class citizen," says Koops. "Automation is as simple as adding data to the input repository to trigger pipelines downstream, which creates a nice, traceable way of computing output."

Koops notes that this approach – creating simple subtasks that are orchestrated through Pachyderm to solve complex challenges – can apply across domains. It allows the company to use or adapt publicly available AI models to its unique needs. "This sort of orchestration works for any task that can be distilled down into smaller components; it's much more efficient and flexible than creating monolithic models, and with Pachyderm it's easy to recombine the components to solve other interesting challenges."

About RTL

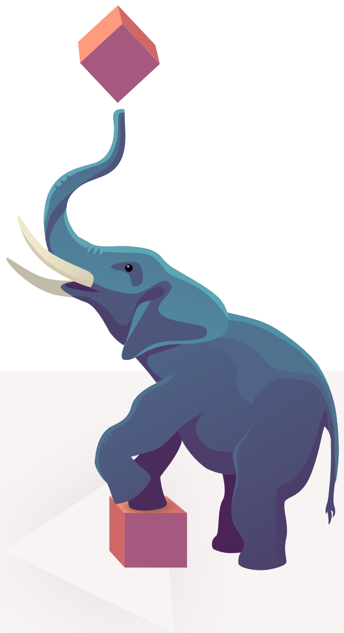
RTL Nederland is a 100% subsidiary of RTL Group, Europe's largest TV, radio and production company. RTL Group is 75.1 percent owned by Bertelsmann, a large international media group. Sister companies of RTL Group are the publishers Penguin Random House and Gruner + Jahr, music company BMG and customer service provider Arvato.

About Pachyderm

Pachyderm is the data foundation for machine learning. Pachyderm provides industry leading data versioning, pipelines and lineage that allow data science teams to automate the machine learning lifecycle and optimize their machine learning operations (MLOps).

With investment from Benchmark, Microsoft M12, and others, Pachyderm, Inc. offers a user-deployed Pachyderm Enterprise Edition, a hosted SaaS Pachyderm Hub and an open source Pachyderm Community Edition.

Pachyderm helps customers get their ML and AI projects to market faster, lower data processing and storage costs, and supports strict data governance requirements through data driven automation, petabyte scalability and end-to-end reproducibility.



Contact Pachyderm

To learn more about Pachyderm's machine learning data foundation, contact us:

info@pachyderm.com • 888-338-9597 • www.pachyderm.com