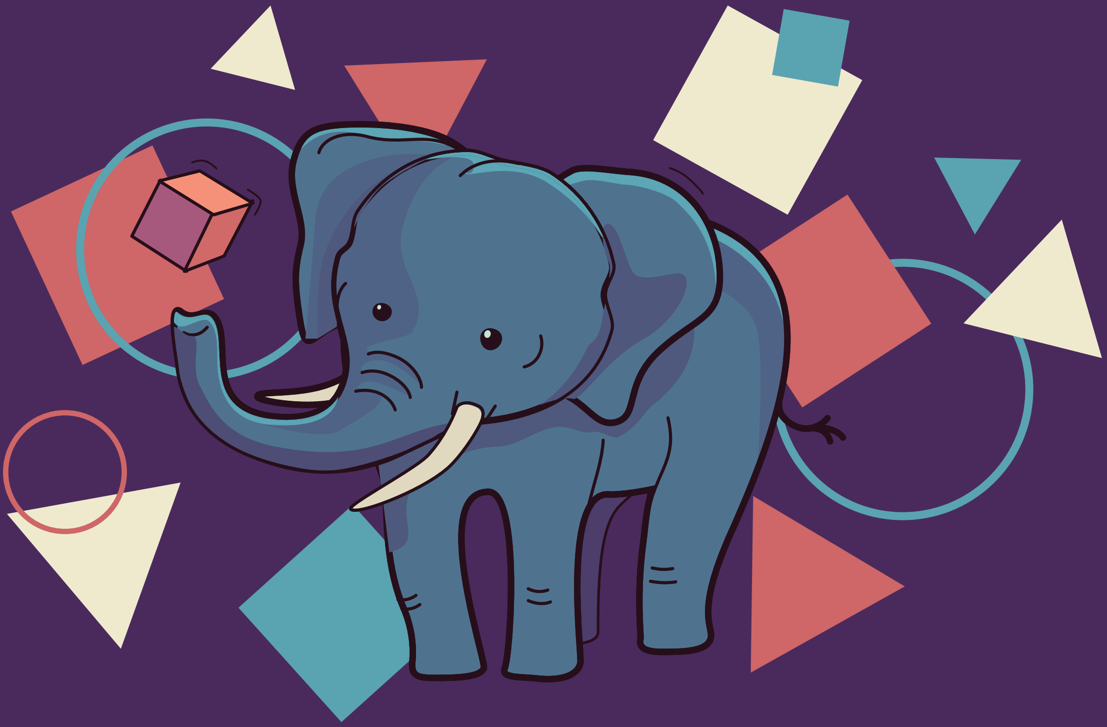
Pachyderm Customer Use Case:

Epona Science

How one company uses Pachyderm to bring it all together.



Epona Science specializes in buying, breeding and identifying the best racehorses in the world. With every thoroughbred a multi-million dollar investment the stakes are high. Buyers and breeders want the best information possible to give them any edge in picking the next legendary champion.

"Pachyderm has enabled us to rapidly build and maintain a robust and automated data science pipeline that is scalable and completely reproducible. We can now confidently make high-stakes purchases in a shorter time-frame because Pachyderm rapidly delivers our results and allows us to verify their accuracy and integrity."

– Ryan Smith, Head Of Data Science at Epona Science

https://www.pachyderm.com/case-studies/epona/

Overview

The racehorse business is one with a long and traditional history. In the past many buyers simply bought horses with a pedigree or from a trusted breeder who seemed to know the best horse by instinct. But are those really the best predictors of success? In other sports, like baseball, we've already seen deep statistical analysis beat the gut instincts of famous managers. Sabermetrics, immortalized in the movie Moneyball, helped sweep the Boston Red Socks to victory after nearly a century of championship drought. Now every major sports team relies on the data as much as instinct. There are millions of factors that go into making a winner and too often people have a tendency to focus on the wrong ones.

Epona set out to revolutionize this historic industry with machine learning, statistical analysis and science. Along the way, they've discovered that everything from the horse's entire genetic profile and lineage, to the animal's height and gait, to the size of its heart can all make the difference between a winning horse and one that never really makes it out of the gate.

The Challenge

With the stakes so high, horse breeders are an insular and close-knit group. They've quickly realized that data can tell its own tale and upend their traditional sales patterns so they're protective. That means Epona has to pull from a lot of sources all over the world, whether its x-rays or genetic profiles or track records from previous races. Gathering all that data, cleaning it, standardizing it and getting it into a consistent format that their machine learning models can train on is a lot of work.

Too often people don't realize that 80% of data science is finding the right data, pulling it down, extracting it, transforming it and loading it. Every type of data presents its own challenge. It only takes a day and a few hundred dollars to sequence an entire genome but often these rapid sequencing machines make little errors.

"Genetic data is always imperfect," says Ryan Smith, Head of Data Science at Epona. "There are missing genotypes, wrongly labeled bits. You can mitigate some of those problems [with different algorithms that fill in the blanks] but if you change the method you use to mitigate it, you need to know exactly why it changed." A sudden change in mitigation solutions can easily throw your models into disarray.

Compiling all this data is a bit like "fiscal modeling," says Smith. "How could the horse's share price go up or down?" But processing it all was taking weeks or months for Epona's team. They had too many manual steps and lots of little glue scripts to pull the data and transform it. They needed to go a lot faster. That's where Pachyderm came into the picture.

Why Epona Chose Pachyderm

Pachyderm immediately stood out to the team for dealing with everything from data lineage, to data transformation and versioning, to containerization. "Without containerization," says Smith, "dealing with the setup is tough and if you can do it in Docker you can save a lot of pain.

The Pachyderm platform's versioning and provenance tools deliver the key ability to roll backwards and forward. They can look at what changed, when and why. The team's models are subtle and sensitive and sometimes their engineers need to do a detailed forensic analysis to figure out just where the model went wrong so they can fix it fast.

They also found the platform much easier than alternatives like Airflow, which are more rigid and not designed for Kubernetes first. Like anyone in data science, they do a lot in Python but they need the flexibility to work in other languages.

"A lot of the software tools come out of academia," says Smith. "They're developed by researchers." That means many of the tools they need are not enterprise ready with a slick and easy to setup feature set and installer. The tools are cutting edge but they're rough around the edges. Pachyderm lets them easily string together a series of independent and isolated tools into a smooth pipeline. That's changed the way they do business because they had to run everything in isolation in the past.

Their model development throughput is now effectively continuous. Every single model they have and every sample, especially genetics samples, runs through the pipeline, gets tested and uploaded to the website in minutes. That took days or weeks in the past, with lots of manual steps along the way. Before Pachyderm they didn't have the ability to totally automate every part of the process, as the model got rebuilt again and again with new information. With a small team they needed to concentrate on model building not manual steps.

Autoscaling made a massive difference for them as well. In 2020, they processed 10,000 new photos in a month and the year before it took a year to do that many pictures. With autoscaling, they can leave a cluster on stand up and build it up as needed. When they ran an animal model in the past they used a massive machine with a terabyte of ram and as many cores as they could get but it was cost prohibitive. Now kubernetes can spin up training and distribute it out into pods so they don't have to launch a mega-instance, run the job and debug it and remember to turn it down.

Epona is changing the way the high stakes horse racing business works but it's Pachyderm that gives them the engine they need to bring new life to a business steeped in tradition.