# Penske Media Corporation (PMC) Builds the Infrastructure They Need to Pursue Machine Learning and Data Science

**"The great thing about Astronomer is they don't just manage Airflow. They share best practices, monitor DAGs and are always on deck to answer questions. Astronomer takes care of itself, more or less, and I'm free to focus more on the stuff I love, like data science and machine learning."**
—Andy Maguire, Data Scientist at PMC

**Introducing Penske Media Corporation (PMC)**

As a leading digital media and information services company, PMC's award-winning content attracts a global audience of more than 180 million through brands like Rolling Stones, Variety, IndieWire and many more. When Andy Maguire came to PMC from Google in 2015, he found himself tasked with first building the foundations for PMC's data infrastructure. A solid framework and approach here is crucial to making it easier to prove the importance of data science in understanding PMC's user base and content performance.

Andy had a deep understanding of the power of data and a plan to incorporate data and machine learning into the heart of every PMC brand to power things like recommendation engines, content pageview predictions, subscriber affinity modelling and much more.

Success, however, required a rich data infrastructure and ecosystem, from the breadth and depth of sources being used to the tools and technologies underlying it all.

**The Challenge**

When Andy joined PMC, the data infrastructure was still quite young. Clickstream data from

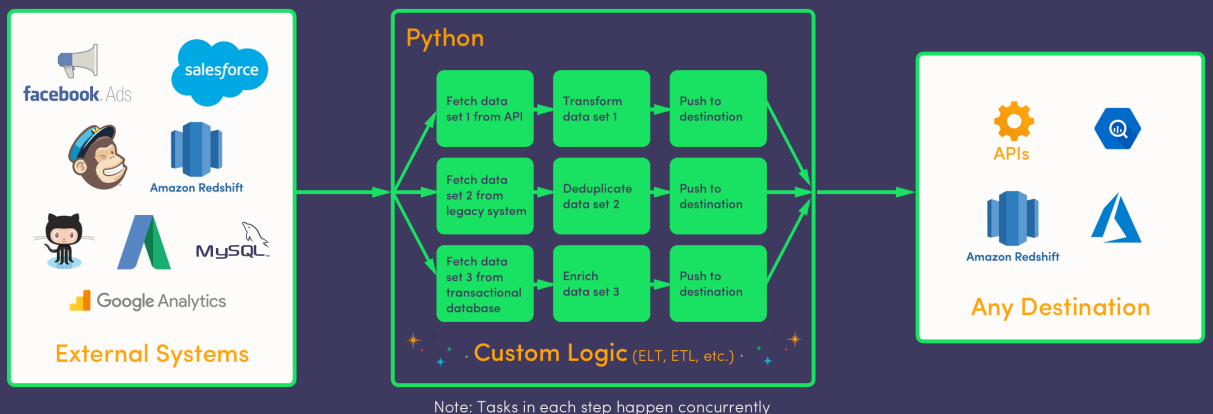
Google Analytics was flowing into Google BigQuery, but it was not being fully leveraged, enriched and made actionable to the business. Where possible, the decision was to leverage cloud tools and limit the "data ops" aspects of the infrastructure. Before long, a myriad of cron jobs, jobs servers and raw job log files had begun to eat away at the time Andy and his team had to extract insights from the data. "I was frustrated," says Andy, "that wasn't what I wanted to to be doing." A data scientist by trade, he wanted as little long-term overhead as possible when it came to data engineering.

When Andy discovered Apache Airflow, which programmatically authors, schedules and monitors workflows, he replaced his cron jobs and began to more efficiently engineer his data as directed acyclic workflows (or DAGs). But even Airflow required quite a bit of management. Andy found his engineering team was still stretched thin as they struggled to handle the robust data infrastructure required to build machine learning technology and run the in-depth analyses needed for insights. "I probably would have had to assign a full-time manager," Andy explains. "It was too easy to make a change to a helper file and kill all the DAGs. If I broke something, I broke everything. There was no testing framework. It simply wasn't efficient."

**Enter Astronomer**

As Andy and his team looked for alternatives that required less management to operate, they stumbled across Astronomer's managed Airflow option. It was exactly what they had been looking for—and more.

"The great thing about Astronomer is they don't just manage Airflow," says Andy. "They share best practices, monitor DAGs and are always on deck to answer questions." This dedicated support is offered through Intercom. Anytime Andy has a question, he hops onto a chat with the Astronomer




Note: Tasks in each step happen concurrently

Airflow

team. “It’s nice not to feel alone in this,” he says. “Astronomer was an easy sell because it took care of itself, more or less.”

PMC’s use of Astronomer has also solved a tricky monitoring issue with Airflow. Natively, Airflow tells you if something is failing, but as long as at it’s technically “green” and still processing, it’s impossible to tell if something isn’t quite right but still running successfully. Now, Andy feeds events from each task into BigQuery, where it is passed into an anomaly detection system that detects minute behavioural changes in Airflow.

“Now I’m free to focus on the interesting data science stuff,” Andy says. With increasingly little effort, he gathers the content analytics, pageviews, social media, comments and other stats that drive the recommendation technology and lay the foundation for his progressive social media analysis.

Andy also has time to pursue additional goals. “For one thing, we want to get smart with subscriptions by finding out how likely users are to subscribe in the first 60 days of engaging with our content,” he elaborates. This—and more—will be easy to do with the right data, including infrastructure improvements. For example, PMC’s data is currently delivered in batches, but Andy and his team will soon begin streaming data in real-time. This will improve the accuracy of their foundational machine learning and unlock new analytics opportunities.

“Basically, I want to learn everything I can about our users,” Andy explains. And now, that’s the bulk of his job.

**With Astronomer, PMC implemented the data ecosystem they need to pursue data science and machine learning to drive a truly personalized content experience for every user across their many brands. And since Astronomer handles the data engineering and ensures the ecosystem is healthy, PMC’s engineers can focus on analytics and data science.**