

Andrew Jones

andrew@andrew-jones.com

<https://andrew-jones.com>

Principal Engineer and leader with a background in software and platform engineering, and more recently specialising in data platforms and data engineering. Author of *Driving Data Quality with Data Contracts*. Regular speaker at conferences, meetups and on industry podcasts.

OVERVIEW

I'm a Principal Engineer with over 15 years experience, with the first half in software engineering and platform engineering and more recently in data platforms and data engineering.

I joined GoCardless as its first data platform/engineering hire in 2017, and led a great team as we built out our data platform from scratch. After initially following a typical data architecture and getting frustrated with facing the same old challenges we've faced for years, I started thinking *there must be a better way*, which led to me creating and defining [data contracts](#) in 2021.

Since then, data contracts have really taken off and become a well-adopted architecture pattern in the data engineering/infrastructure space. I've continued to write about it, as well as presenting regularly at meetups and conferences and appearing on numerous podcasts and interviews.

At the start of 2023 I was presented an opportunity to write a book about data contracts. It was released in July 2023 as [Driving Data Quality with Data Contracts](#) and has been very well received.

I've recently moved up to a Principal Engineering role - one of the 8 most senior software engineers at GoCardless - and am working with many teams across the organisation on some of our largest projects. I'm self-directed, and happy to work wherever I can add the most value. I make my projects successful in any way necessary and wear multiple hats. I avoid attempting to deliver projects or change myself, and instead I guide, mentor, and influence others, to ensure the best outcome and the greatest impact for the organisation.

Outside of work, I'm a father to 2 young children who keep me busy! Beside that, I enjoy running and brewing my own beers.

PROFESSIONAL EXPERIENCE

GoCardless — *Principal Engineer*

SEPTEMBER 2017 - PRESENT

I joined GoCardless in September 2017 as the first hire on our new Data Platform team, tasked with **building out our data platform from scratch**. As tech lead, I played a key role in designing our architecture, setting our direction, prioritising our work and managing the team.

We spent the first couple of years building out the primitives of our platform, ensuring all our data was available and accessible, and putting in place the initial BI and Data Science tooling. As part of that, we built a change data capture (CDC) service, that captured changes from our upstream Postgres databases and, via Google Cloud Pub/Sub, wrote those changes to BigQuery, from which we could recreate the database with the same structure. We also deployed Airflow as our orchestration solution, and implemented a feature store to support our data scientists.

This was successful, and by then the **data platform was driving key data-driven applications**, including our analytics and ML-based products, such as our internal fraud detection and customer facing applications such as our Success+ and Protect+ offerings.

Despite that success, I felt **things could and should have been better**. I spent a lot of time with the data consumers - particularly our BI and Data Science teams, though also our Product Engineering teams - and I kept hearing the

<https://andrew-jones.com/cv.pdf>

same problems about how the poor quality of the data, and its unreliability, affected their day-to-day work and the quality of their applications.

I was also concerned at the complete separation between those data consumers, and the data generators. They never spoke to each other. The data consumers felt they could only have the data they were given, and were not asking for any improvements. The data generators didn't even know they *were* data generators, and certainly did not know how the impact of their changes to the data affected key business processes and applications built on that data.

As well as speaking to people across GoCardless, I also spoke to people at other organisations, and found they had the same problems. So, I started to explore why this was the case, and if we could do things a different way.

Eventually, I came up with **data contracts** as the solution to these problems.

The idea is simple really. If data is important to us - and it is, as without quality data we can't build the data/ML applications that are part of our company's strategy - then let's start treating it with the same discipline as we do our software. We'll stop building on the internal models of upstream databases, and instead have a well defined interface for our data - just like we do for our APIs. That interface will be owned by the data generators, and through that interface they'll provide data products to their data consumers. They'll do so directly and with autonomy, which will naturally bring them much closer to their data consumers.

From the start, we knew that achieving this would require two things. The first was the **data tooling** and architecture to support this model. The second was to change the **data culture** to one where data is treated as a first class citizen, and data generators are happy to produce quality data because they understand the value it creates for the business.

That's what we've been working on over the last couple of years. We've had a lot of success, as well as some challenges. But overall, we've seen enough success to know we're on the right track. We have over 300 data contracts in production, and have seen them adopted for some of our most critical data.

It's been a great **learning opportunity** for me. I've had to speak to many different teams, and people in many different roles, and often ask them to do *more*. I've learned a lot about influencing and persuading and about aligning incentives in order to push things forward across multiple teams. It's my growth in this area in particular that was key to my promotion to Principal Engineer.

Arm — Staff Engineer

SEPTEMBER 2008 - SEPTEMBER 2017

I first joined Arm in 2006 for a one year placement position as part of my University course and accepted the offer to come back once I graduated in 2008.

The first few years were spent mainly on building internal tools and solutions that helped Arm's verification engineers run tests on their processor designs. These were run on internal high performance compute clusters, and my team's mission was to ensure they could do that as efficiently and as effectively as possible.

As the amount of tests they could run increased, the amount of data generated from those tests also increased. That led us to needing a data platform to store that data and make it available and accessible to the verification engineers, who used the test results to debug their processor designs.

I became the tech lead of the team responsible for building that data platform, and from then on have been specialising on data platforms and data engineering. We built this using the *Lambda Architecture*. We used Apache Flume to ingest and collate data from our data centres around the world, stored that data in HDFS, and processed it using initially Apache Hadoop and then later Apache Spark. These were deployed on an Apache Mesos cluster, which we also managed.

