The term “data quality” can be a little vague, and in discussions, seem more abstract and therefore less urgent than something like feature work.

But that’s not the experience of engineers, data scientists, executives, and product managers who are dealing with data quality issues.

Here are six specific stories from engineers and product managers at marquee companies about a moment when the importance of data quality really hit home.

**Shopify**

“In 2015, 2016, Shopify’s data collection system was completely un-schematized. We had a library called Trekkie where you could literally just say Trekkie.track() and send it any data. What we found was that there was this separation between engineering and data where the engineering team thought, ‘I am collecting this data for the data, and data scientists will tell me when things break and whether it’s useful.’ We found the iteration cycle and the data quality cycle here was in the order of months. We sometimes had outages where we discovered something was lost, three months out. It became obvious that this was just not sustainable.”

* [Zeeshan Qureshi](https://www.linkedin.com/in/zeeshanq/), formerly Tech Lead Data platform at Shopify

**Lyft**

“We brought on Mode, which is a SQL editor, we brought on Looker, we brought on Tableau, we started incubating Superset – these were all basically BI tools looking at different views of the same data. The challenge was that we started realizing that all these teams were looking at the same metrics in different ways. For example, we never got to a single source of truth on revenue or basic metrics like conversion, because at any given time there were several different versions of it floating around. And where this story ends up going is that you’re at a meeting with a CFO, they’re looking at two different numbers, and they don’t know which number to trust.”

* [George Xing](https://www.linkedin.com/in/georgexing/), ex-Head of Analytics at Lyft

**Dessa**

“We worked with a Fortune 100 bank that had built a really solid fraud detection model – highly accurate, etc, but over the span of 8 months, the model degraded so much in production. The team was busy dealing with other things, and they didn’t have the tools necessary to actually automatically monitor the model, and it led to a lot of revenue loss for the company.”

* [Mohammed Ridwanul](https://www.datacouncil.ai/speakers/mohammed-ridwanul), formerly Product Manager at Dessa

**ZipRecruiter**

“It was data that was duplicated in multiple places – when I joined a year and a half ago, we had six different tables in RedShift that were purportedly about applications. Just counting them and getting basic attribution was impossible. You would do queries and you would get weird results back; it was hard just to figure out which database you needed to go to, and very quickly, this became the constraining resource in product development - you just could not get the data you needed to make good decisions.”

For example, I wanted to ask the question of what percentage of our revenue comes from search traffic visitors. This is something that everyone who wants to make decisions about what products to build or not, should be able to figure out.

I opened up Periscope, which had a lot of complicated problems, but was one of our main workhorses, and started putting together a reasonable looking query from reasonable tables. It took me ten minutes to get my first answer, but it was obviously wrong. I pulled in another PM, who had been there longer, and we worked together for an hour, to figure out all the filtering conditions and joins we had to do, we had to reach out to another Slack channel. We finally got to an answer in the end, but when you added all that up, it was probably $1000 in time cost to get that answer, which is completely insane.”

* [Paula Griffin](https://www.linkedin.com/in/paulajgriffin/), Director of Product at ZipRecruiter

**LinkedIn**

“Back in October 2018, we had an instance at LinkedIn when data quality problems affected the job recommendations platform. Client job views and usage declined by 40 to 60% for a short period of time. Once this decline in views was detected, it took a total of 5 engineers 8 days to identify the root cause and 11 days to resolve the issue.”

* [Arun Swami](https://www.linkedin.com/in/arunswami/), Principal Software Engineer, LinkedIn

**AT&T**

“Some years ago I worked in AT&T's financial services business. At one point I served as an Analyst for the equipment leasing division. (We didn't call ourselves data scientists back then.)

Leadership wanted stats on conversion rates by vertical market. In the leasing business, conversion rate is the ratio of deals done to the deals you approve. Buyers shop around and get multiple approvals before they decide, so you want that rate to be close to 100%.

Naturally, we needed the numbers yesterday.

We didn't have a data warehouse, data mart, or data lake, because the CIO thought such things are foolish luxuries. I pulled data directly from the source system; we're talking EBCDIC files on an AS/400 LOL.

The conversion rates looked pretty good for all but one vertical market, where they were very low. Leadership looked at the numbers, spoke to sales reps, and they were all baffled.

The segment manager decided that the problem must be that our pricing was too high, so he quickly implemented price cuts.

Meanwhile, I started to drill down into the data.

When I pulled individual applications, I discovered the source of the problem. Some genius in IT had run a test of the lease approval system by planting fake applications in the production data. He assigned all of these applications to a single fake merchant, and for reasons known only to him, coded that merchant in the offending segment. The fake applications had no other distinguishing characteristics.

The lease approval system processed the applications and dutifully wrote them to the "Approvals" file. There were no real customers to match the fake applications. Thus, the conversion stats computed for that vertical market were wrong.”

* [Thomas Dinsmore](https://www.linkedin.com/in/thomaswdinsmore/), head of competitor intelligence, Domino Data Lab