On being a Data Detective

"I have a problem. I am an intellectual, but at the same time I am not very clever"

— Sue Townsend, The Secret Diary of Adrian Mole, Aged 13 ¾

Data is logical, organised – and a modern data platform will do exactly what you have told it to do, but it's doing what you told it to do on ALL the data, not only the edge cases. Unfortunately, data is also noisy and messy. Even though a lot of today's data is automatically generated – almost always are you adding it to human generated data. And humans invariably make mistakes.

Detective work is required to find the clues as to where there are problems, but as with all detective work, the devil is in the detail.

It's easy when looking into data inaccuracies, to look at all the problems and work on way too many things at once. When you are dealing with petabytes of data and millions of rows of data in thousands of tables, it's way too common to get caught up in the noise. In today's world you can no longer export your entire dataset into excel and scan for an anomaly.

Time and time again people end up looking in the wrong place for the problem. It's usually not something complicated, it's something simple. We need to stop trying to be too clever and stop overthinking it.

Below is the approach I regularly take is resolving issues to keep things simple.

Step 1: Identify a Single Problem to look for

> Start with your reporting layer – your business specialists know it's wrong there, trust their "feel" that the number is wrong and try to get them to explain why they think so? But get them to give you a single problem. Validate it, understand the problem – **Don't try to see what's causing it yet**

Step 2: Identify a record that has that problem

> Narrow to a single customer, a single account, a single date and ultimately a single record until you know you are looking at an incorrect record in a single place – **Don't try to see what's causing it yet**

Step 3: Follow that record

Follow the record down the transformation process – where does it change? No! This is still not the time to find the cause – you are following the data. Find WHERE the root cause is, not WHAT the root cause to the problem is.

Step 4: Recreate the issue in a small dataset

If you must load millions of rows through multiple layers of your data platform, you are going to be there forever waiting for things to load. Once you know a record or two that has the problem (and what you know that number should be), then and only then can you efficiently theorise and test.

Step 5: Theorise, test, retry keep looking – FIND root cause

This is where it gets fun. Try things out – guess, theorise. What happens IF...? Because you are no longer using the larger dataset, or the real code – you can't break anything.

Step 6: Fix it, then expand your test

Once you locate what you think will fix it, it's time to expand to other test cases, more data. Is it fixing the number? Great – try more records – does it also fix it?

Step 7: Regression Test

Test on larger and larger sets. Ask yourself what else *might* have broken with your fix – and test, test, test

Step 8: Next problem

Now that this is complete – start at Step 1, find the next issue and solve it!

Locating and resolving issues in this way seems methodical and a lot of things to think about, but I promise you unless you are truly gifted, your mind can only hold onto, and trace, a finite number of things. Jumping to solution mode too early, or trying to solve everything at once, only slows you down and causes mistakes. Methodically, calmly solving data issues one at a time will make your life much more pleasant, as well as ticking off regular wins as you go. Take a breath and dive in – remember, this is supposed to be fun!

To find out how we can help with your data challenges, please contact a Fujitsu Data & AI specialist now.

Contact
Fujitsu Data & AI
+61 3 9924 3000