



BEYOND THE HIRE

# You Hired the Data Scientists. Who Is Building the Pipes?

Why analytics ambition keeps outrunning the infrastructure that has to feed it

Giovanni Leonardi

January 2014

# You Hired the Data Scientists. Who Is Building the Pipes?

Transformation · No. 182

Originally written January 2014 | Re-edited and re-rendered for the WhereBeyond Edition, 2025

---

*"We have spent two years now hiring the priesthood  
and forgetting the plumbing."*

---

## 1. The Question That Empties a Room

---

There is a particular silence that falls in the second week after a data scientist joins. The appointment itself went well — a genuine specialist, courted for months and hired at a premium, someone who can talk fluently about regression and classification and the models that were going to change how the business understood its customers. Then, somewhere in that second week, they ask the question that empties the room: where is the data? Not the monthly dashboard, not the extract that someone emails round on the first working day, but the clean, current, connected data on which the entire appointment was predicated. The answer, when it comes, arrives with an apologetic shrug: it is in nineteen different places, a good deal of it in spreadsheets, almost none of it reconciled, and nobody is entirely sure which version is the true one.

We have spent two years now hiring the priesthood and forgetting the plumbing.

I want to argue something slightly heretical in the middle of the present enthusiasm: that the data scientist is the most visible and, hired first and alone, the least useful thing an organisation can buy. The scientist is not the binding constraint. The binding constraint is the thing nobody photographs — the pipes that move data, reliably and repeatably, from the places it is created to the places it could be used. Hire brilliance and starve it of that, and you have bought an ornament.

## 2. Why We Reach for the Scientist First

---

The reason is not stupidity. It is a visibility asymmetry, and once you have seen it you see it everywhere.

Hiring a data scientist is a board-legible act. "We have appointed our first Head of Data Science" is a sentence a chief executive can say on an earnings call and write into the annual report. "We have built a nightly job that reconciles the customer table across three source systems" is not a sentence anyone says out loud, even though it may create far more value. One is a headline. The other is drainage. Boards fund headlines.

The prevailing enthusiasm sharpens the effect. Since a well-read business magazine christened it the sexiest job of the century, the data scientist has become the emblem of an organisation that takes data seriously, and the technology vendors have been only too happy to sell that emblem back to us. The pitch always has the same shape: buy the cluster, hire the wizards, and insight will follow as night follows day. So the commodity-server cluster is duly bought, and sits half-idle; the wizards are duly hired, and sit frustrated. Everybody is selling the model. Nobody is selling the tarmac.

And there is, as yet, genuinely no home on the org chart for the person who lays the tarmac. We have a settled title and a clear career for the modeller. We do not have one for the person who builds and maintains the pipelines — the name for that role is still being argued over in conference corridors. So the work that matters most has no obvious owner, no career path, and no dedicated line in the budget, while the work that photographs well has all three. A newly minted Chief Data Officer is far more likely to be judged on the calibre of the people hired than on whether a given number means the same thing on Tuesday as it did on Monday.

*The uncomfortable arithmetic of the last two years is this: we have hired hard for the ten per cent of the problem that is glamorous, and looked away from the ninety per cent that is plumbing.*

### 3. The Mandate That Made It Worse

What has widened the gap rather than closed it is the very mandate meant to close it. Told from the top to "do something with big data", organisations have responded by acquiring more of it — standing up clusters, retaining log data they previously discarded, tapping transaction and web and sensor streams they had never thought to store. All of which would be admirable if any of it were connected to a pipe. Instead the volume of data an organisation possesses has raced away from the volume it can actually use, and the ratio between the two — call it the yield — has quietly collapsed.

We have mistaken accumulation for capability. A larger reservoir behind the same broken tap does not get you more water; it gets you a larger reservoir and the same trickle, plus a maintenance bill. The scientists were hired to drink from a firehose. What they were handed was a warehouse full of sealed barrels and a note apologising for the absence of an opener.

### 4. What It Actually Looks Like on the Ground

Set the theory aside and count the hours, because the hours tell the truth.

Consider a team I would recognise in almost any sector: six analysts, each expensive, assembled over a year with real internal fanfare and a mandate to "unlock the value in our data". Watch where their days actually go. Of every five, four are spent not on modelling but on dragging the data into a state where a model is even conceivable —

extracting it from operational systems that were never designed to give it up, reconciling a customer count that stubbornly differs depending on which system you ask, chasing down why last month's revenue figure quietly moved when no one will admit to changing anything. The modelling — the thing they were hired for, the thing they are genuinely excellent at — happens on the fifth day, if the week is kind and no source system breaks. Four-fifths of a scarce and costly capability is spent doing work that a well-built pipeline would have done silently overnight.

Then there is the second, quieter failure, the one that does the more lasting damage to the credibility of the whole endeavour: the model that works beautifully and never runs. A churn model is built, cross-validated, and presented to a delighted steering group; the accuracy is genuinely impressive and the room is won over. Six months later it has changed precisely nothing, because the features that fed it were hand-assembled on an analyst's laptop from three separate extracts that no scheduled process will ever reproduce. There is no path from the notebook to the system that actually contacts a customer. The model was never a product. It was a slide, and slides do not retain anyone.

---

*"A model that cannot be refreshed is not an asset. It is an expensive photograph of a moment that has already passed."*

---

This is what analytics ambition without analytics infrastructure reliably produces: our most brilliant people doing janitorial work, and their most brilliant work dying in a steering deck. Neither failure is a failure of talent. Both are failures of sequence.

## 5. The Strongest Objection — and Why It Only Half-Holds

---

The serious counter-argument deserves to be put at full strength, because it is not a straw man and I have watched it defeat the lazy version of my own case more than once.

It runs like this. Infrastructure built in the abstract is infrastructure built wrong. If you wait until the data is pristine before you hire, you will wait for ever — and worse, you will most likely spend three years and a small fortune on a grand data-warehouse programme that delivers an immaculate schema and not one useful insight. We have all seen that death march, and its corpses litter the industry. So hire the talent, the argument concludes, and let their concrete demands pull the plumbing into being behind them. Necessity is a better architect than any committee.

I have real sympathy for this, because the warehouse-to-nowhere is not imaginary and the instinct to avoid it is sound. But the argument proves far less than it claims. It is quite true that infrastructure should be pulled into existence by real analytical demand rather than built speculatively against a diagram. It does not remotely follow that the person you hand the pipe-laying to is the data scientist. A data scientist building

extract-and-load routines is a Formula One driver out resurfacing the track: you are paying the highest day-rate in the building for the task it suits least, and collecting mediocre plumbing and no models for the privilege. The choice was never "perfect warehouse first" against "scientists first" — that is a false binary that flatters the decision we had already decided to make. The real question, the one the objection skates past, is whether anyone whose actual job is the pipeline exists in the organisation at all.

## 6. Build the Pipes Before You Anoint the Priesthood

---

The corrective is not to stop hiring data scientists. It is to stop hiring them first, and alone, and calling that a data strategy. Three propositions I would defend to any board:

- The pipeline is the product; the model is a feature of it. A modest model running on data that refreshes reliably every night will out-earn a brilliant model that ran impressively once. Sequence the investment to match that truth.
- Hire the plumber before — or at the very least alongside — the priest. The person who builds reliable, documented, repeatable data flows is not a junior support act to the scientist. In most organisations today they are the scarcer and the more valuable hire, precisely because nobody is competing to make them an offer.
- Judge the function by what reaches production, not by what reaches the steering committee. A model in a slide is ambition; a model wired into a system that acts on its output is value. Count only the second, and watch how quickly priorities rearrange themselves.

None of this is a call for a grand infrastructure programme before any analysis may begin — that is the very death march the objection rightly fears, and I want no part of it. It is a call for the opposite discipline: build the smallest pipe that reliably delivers one clean, current, documented dataset; put a single real business question through it from end to end; and grow the plumbing behind proven demand rather than ahead of an imagined one. Thin, reliable and in production beats broad, brilliant and stranded, every time.

The organisations I see actually getting value from their data at the moment are not, as a rule, the ones with the most impressive hires or the largest clusters. They are the unglamorous ones that can answer a deceptively simple question honestly: when a model needs data tomorrow morning, will it be there — clean, current, and the same as it was yesterday? Everything else we have been buying so eagerly — the talent, the tooling, the ambition itself — is waiting on that answer.

We were sold the scientist. What these two years should have taught us, and mostly have not, is that we needed the plumbing first.