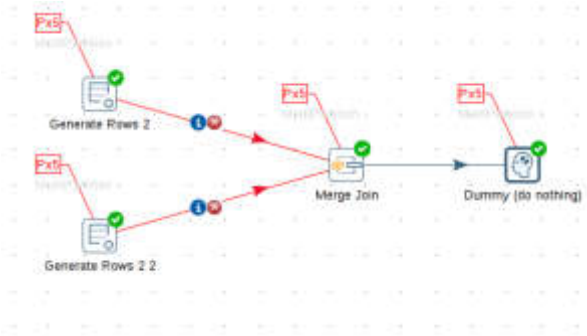
Scaling merge joins in Pentaho Data Integration

dankeeley.wordpress.com |

A while ago I was working on a project which used brutal amounts of partitioning within PDI so we could get a decent throughput on our jobs. The main reason for this was a constraint of using just a single box – Not a common architecture these days. Anyway we managed to get the throughput we needed (Getting on for 1M rows per second over the whole job) so that's fine.

We hit a minor issue though when it came to the merge join step. (A similar albeit different issue exists for stream lookup). Basically if you have a lovely partitioned stream coming in, and want to join with another already partitioned stream, PDI doesn't support this. In fact, it gives you a great big red hop just to make the point:



If you hover over the hop it tells you why you cannot do this.

That appears to mean you must de-partition (via a dummy step) single threaded merge join, then re-partition moving forward. Meh; That's just not nice.

There is however another way in some cases – this depends entirely on your data whether or not it will work. But for us it did – and for us it scaled wonderfully. The simple mapping.

This new(ish) mapping component has one important feature over the original mapping component (do we call it complex mapping?) – It can be partitioned. And the reason for that is because it can only have and must have 1 input and 1 output.

So our new transformation looks like this:

And the mapping itself looks like this:

Now; You must test this carefully because there's an awful lot more work in doing this – and I can imagine in some cases you'll actually get worse performance. But it's another handy tool to have in the belt! And YES; It does mean your input query for the right hand side of the join gets done as many partitions as you have – but that's a price you may have to pay. (And again, only works if one stream is an order of magnitude larger than the other.)

Don’t forget to checkout the free professional support sessions we’re offering at PLUG next week!