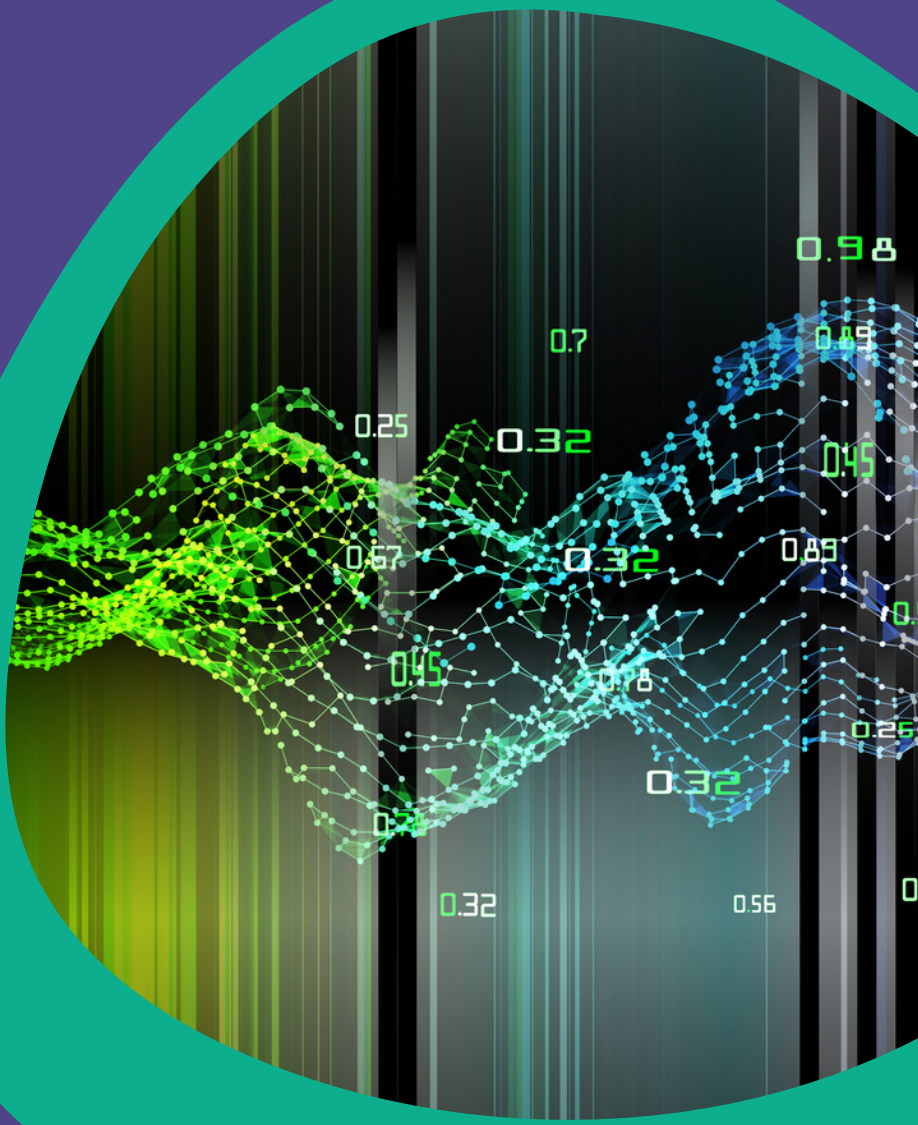
Getting AI Ready

The importance of data quality



Getting AI-Ready – the importance of data quality

Introduction

This article is intended to help you if you are interested in improving your confidence in the data your organisation provides to internal consumers, customers or suppliers. We explore why data quality is almost always lower than everyone would hope, why that is increasingly holding companies back and how to put a strategy in place to get to grips with the issue.

Everyone's challenge, nobody's problem

Increasingly we think of data insights departments as the engines of the business and addressing and maintaining data quality is the equivalent of checking the oil levels. We all know we ought to check the dipstick regularly, but it's a messy job and the engine seems to run okay, so we don't bother.

In my experience it is often everyone's job to groan about how poor data is holding them back, but rarely anyone's job to resolve the issue. Hands go up slowly when "fantastic opportunities" to do matching, data audits and such like arise.

Historically there has been an in-built coping mechanism. Users get to know the problem areas in the data, and they compensate by taking those elements of their reports with a pinch of salt. Sometimes it is even convenient to able to cite problems in the data to avoid explaining falling trend lines or unexpected observations.

In a recent study published on BI-Surveys.com (https://bi-survey.com/top-business-intelligence-trends), master data and data quality management came out the number one issue on people's minds.

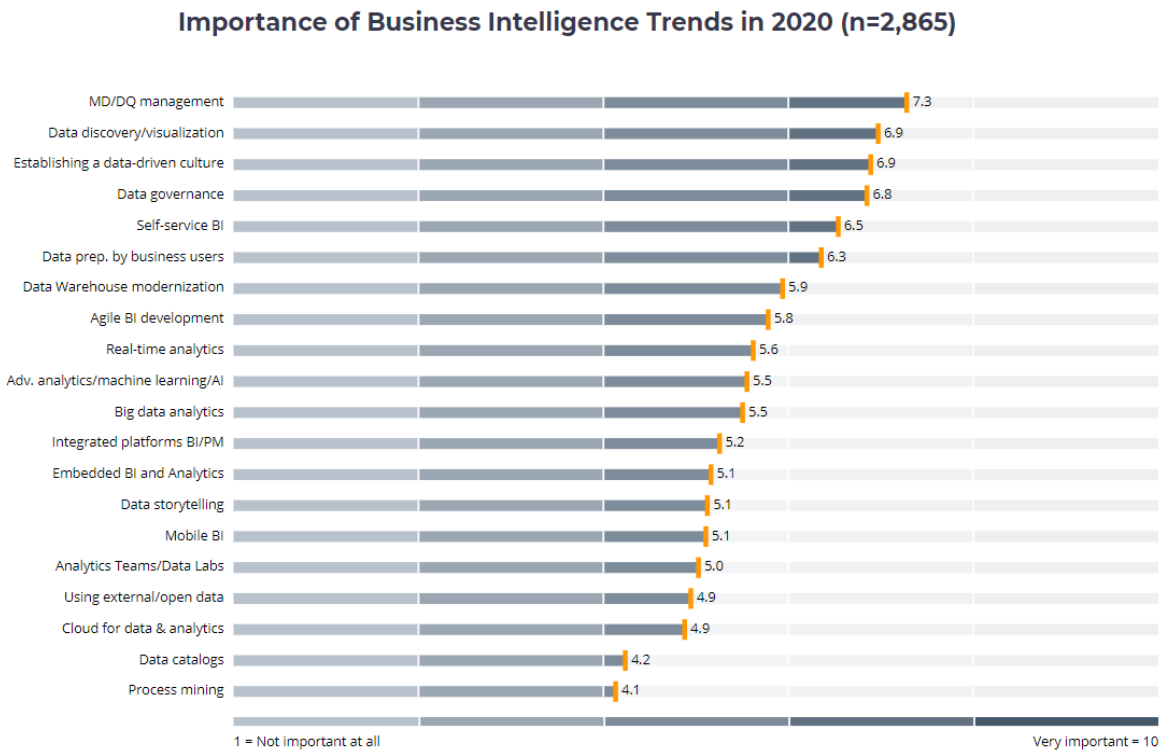


Figure 1: Source BI Survey.Com (https://bi-survey.com/top-business-intelligence-trends)

Many companies, perhaps including yours, are looking to increase the value they derive from data – to make it work harder for the business. However, it's when you start running the engine at high revs for long periods that you soon wish you'd checked the oil.

Tall buildings need solid foundations

It is hard to read a promotional email, website or have a conversation with a potential supplier at the moment without "AI" being mentioned. Many new cutting-edge solutions utilise Artificial Intelligence techniques and technologies to help businesses deepen understanding, make more informed decisions, or improve the efficiency of processes.

As we start to put more of our data into models, we need to be cognisant of the increased detrimental impact inconsistencies in our data can have on the results. In theory your models could be made complex enough to ignore or compensate for known weaknesses in the information, but that will be expensive and time consuming.

Duplicates, unmatched items, mis-matched items, uncategorised entries, irrelevant data, missing data, late data and outliers can all cause models to arrive at wrong conclusions. What's worse is that as models become increasingly complex, they are more of a "black box" even to internal analysts and there may be a lower probability of spotting the problems than with conventional analytics.

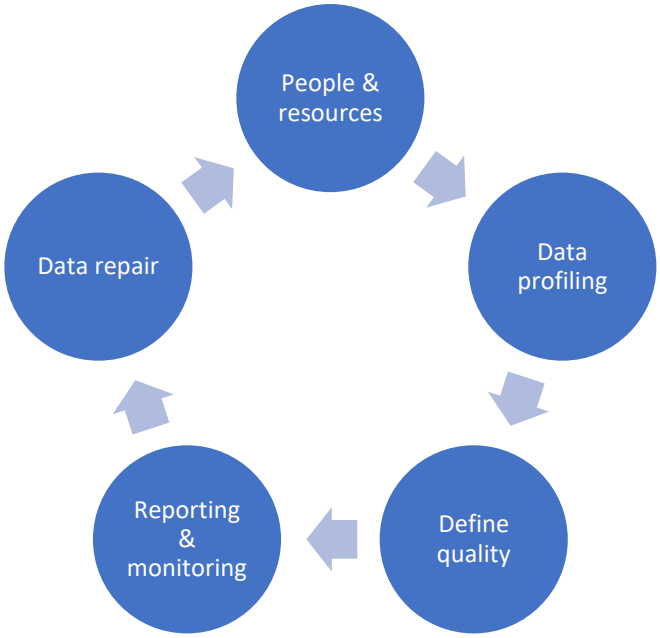
Many would argue you get a better ROI from improving your data quality than you do from many of the smart tools you might put on top of it. The advantage of getting your data quality in place first, is that you don't damage the trust of your userbase when the new tool you spent lots of money on delivers results that aren't expected.

Putting in place a data quality management strategy

I admit that as a sub-heading, this one sounds a bit dull. When little Jonny is asked what he wants to do when he gets older, he rarely mentions working for a company where he gets to put a data quality management strategy in place. The good news is that this is an area where small actions can reap big rewards.

The key to success is ensuring an ongoing process is in place. You keep a clean house by doing a bit of housework each week, not by relying on an annual spring clean.

The following process is generally considered to be a gold standard approach and is explained in more detail in Carl Anderson's book "Creating a data driven organisation".



The important first step is to identify someone who's responsible for data quality. Whether you employ someone, make it part of someone's role or engage with a third party to manage data quality, someone somewhere has to care.

That person can then go about the first step in the path – understanding the shortfalls of the current data landscape. What are the data issues perceived to be in the business? How do these affect how the data are used? What are the users' priorities & what's holding them back? What important data are we missing? This step gives you a starting point, and a sense of what needs to be achieved.

With this understanding in place, it becomes possible to define and agree the standards that need to be achieved. It is very expensive to seek 100% perfection in data, and probably not a necessary aim. Focusing on an acceptable level of quality that enables the data to be used confidently and widely is normally a reasonable and necessary compromise. For instance, you probably don't need all accounts in the system to be matched, but you might decide that all accounts with more than £1,000 of sales in the latest 12 months must be matched.

It is important that these goals are agreed with the business. By doing so, you can gain maximum benefits from the next step.

The next part of the process is often overlooked but is key to achieving and maintaining the trust of users. Only by tracking your progress against the quality objectives you set can you ensure things are improving towards those goals, and then being maintained at reasonable levels. Sharing this progress and achievement with the business helps give them confidence in the data, but crucially, it only does this if the business was on board with the quality goals set.

We are now in a great position. We know what needs to be achieved to get the business on board, we can monitor our progress, and we have resource in place to take action. Using what we have learnt along the way to prioritise our activities, we can start the process of data repair. This will typically be made up of processes such as:

- Deduplication
- Mapping unmapped items
- Repairing mis-matched items
- Categorising data – adding data dictionaries and keeping them up-to-date
- Removing or flagging erroneous data
- Resolving issues in the data supply chain
- Removing or flagging outliers

These are ongoing processes, not one-off tasks. It is important to continue communicating with the user base to inform them of progress being made and collect feedback on their satisfaction as well, as new issues that might have arisen. As data is used in new and different ways, quality objectives may need to change to reflect these.

Conclusion

In order, therefore, to strive for the skyscrapers of the Business Intelligence world we must first go through the arduous and rather mundane task of data cleansing. Only then will we be able to apply the new cutting-edge solutions that utilise Artificial Intelligence techniques and technologies with confidence. That said, ensuring business confidence in data and its subsequent analysis is a no-brainer if you are looking to get a return from the investment you make in your BI stack.

Visit our Data Quality Managed Service page to understand more on data quality management.