Data Quality Manager - you need one if you take data science seriously

Time is ticking for big corporations to make the most of their data assets. Competition is well awakened and a species selection process has already started. Companies not able to adapt soon enough their steering models and embrace an insights driven approach are not going to survive for much longer.

The need to make the most of own and foreign data introduces new challenges in the traditional organization and requires the creation of new roles.

One of the key success factors many companies fail at implementing is a data quality strategy with a clearly defined ***data quality manager***. Let me share with you why:

7 Reasons why you need a Data Quality Manager (DQM)

The intrinsic nature of the *data* imposes the need of this role. As you are going to see below, without a DQM it is going to be really challenging for any company to exploit its data assets in an efficient way:

1) Data expires

And that's the problem *survey* based data providers are constantly facing. For example let's say I run a phone survey 5 years ago on a population segment to gain insights about the place where they are living (e.g.: say to determine the chance of people there driving a BMW or being in the high-end affluence segment or having a bachelor degree). Let's say I run different surveys on the same population once a year and then I aggregate the information and I sell it as a segment (companies such as *Acxiom* do it all the time). Obviously the information I gathered 5 years ago needs to be treated in a different way that the information I gathered this year…

A data quality manager takes care of auditing this information, applying expiring policies, establishing metrics to inform about the recency and trustworthiness of the information.

2) Data is often not consistent across reporting systems

If you are presented with a systems landscape of a big corporation, you can't happen but shake your head. After the initial shock, if you manage to identify which system does what, you are going to certainly discover, that many systems do more or less the same thing... different enough to justify the presence of both but close enough to report the same

kind of information.

Sometimes the problem comes later, in the reporting layer, where different *"cubes"* end up reporting different values for the same metric or KPI... Why that? because of several reasons: different assumptions, different definitions, different ways of aggregating the results...

I'm sure there are plenty of good reasons to do it like that, but the damage caused by diverging information is huge. You end up wasting time you should invest in real analytics to understand which KPI value is the right one and often you have to reverse engineer your way back from the report to the source...

A data quality manager makes sure this situation never happens, by homogenizing definitions, aligning assumptions, deciding which source is the right one... or if the prior is not quickly achievable, at least a data quality manager should be in a position of explaining the different values in the short term.

3) Data does have gaps and inconsistencies

System outages, system releases or just a hard drive running out of space... The reasons why you might have gaps on your data are countless. Yet you don't want each and every data scientist in your organization to create their own gaps filling method. Sometimes gaps are even natural -a bank holiday- and you shouldn't even fill the gaps.

This is again an item in the portfolio of a data quality manager. Ideally, gaps are identified with standard gaps detection methods running as cronjobs and acted upon. The incidence is typically documented along with the method used to solve it.

4) Data can't escape the past

And data scientists cannot make sense of the data without having enough history. Depending on the local data protection regulations, certain information cannot be stored for longer than X months, at least in a raw form. Sometimes you just need to save space... Archiving and aggregating procedures need to be implemented to minimize the value loss.

The data quality manager needs to inform the data custodians and archivers about the value of each data asset and quantify the loss after applying irreversible aggregation procedures. Additionally, the DQM is also responsible for facilitating the access to historical data for the data scientists to implement the intelligence

5) Data value is time dependent

When a piece of insights leads to an action (and it should be a MUST), the action is subject to the so called *"opportunity window"*. Understanding today that I should have lower the prices of a given product to counter fight a competitor's aggressive campaign might be too late... the campaign might even have been finished... and you didn't react.

A data quality manager shall understand how the *time-to-information* impacts the value of the data, flagging data in categories depending on how real-time they are for the business.

6) Data needs to be documented and explained

Without a proper data sources documentation, a company is going to face efficiency issues. Data scientists and data business analysts are going to get their hands on each and every source that provides context to extract intelligence from... These folks don't wait for longer data facilitation processes and tend to grab and scrap whatever they need...

Sometimes what they need is available as a report and the scraping continues...

If you have 5 different teams of data scientists, you can be sure that you end up with many different data gathering procedures or *hacks*, none of them "bullet-proof" and none of them documented and optimized for a general use.

With a catalogue of data sources and the proper process to request new data if not available, under the supervision of a data quality manager, analysts can focus on extracting intelligence and data engineers can work with the proper requirements.

It is responsibility of the data quality manager, to maintain this catalogue and to make it available and accessible, as well as providing transparency on the status of new data requests

7) Data can create other data (inference)

None all the data available in a company is factual information. Equally relevant are inferred data assets, that are made available after applying some data manipulation or inference algorithms on the factual data. For example, let's say you compute the affinity of a customer to buy a given product; taking this score for further analysis requires to be certain about its correctness: *how was it computed? when was last time it was refreshed? what are the possible values of the score -domain-? does this score contradict other information available for the same customer?, etc*

The DQM should provide answers to all questions related with inferred information and establish the link to the ones behind the inference procedure.

Your take away: you need one Data Quality Manager

After such a long post, there is only just a message I'd like you to take away: you need a DQM!
If these 7 reasons didn't persuade you... if you don't have any DQM in your organization, please help me understanding how you manage your data...

In upcoming posts I'm going to tackle the problem of measuring the data quality and the data usage in a corporation...
Just a teaser for now: *Yes, there is a way of measuring it!* and *No, it is not easy to implement*