DataScience®
foundation
MAKE BETTER DECISIONS

In Secondary Data We Trust: Secondary Data ''Trust'' Issues

Author, Michael Baron

A Data Science Foundation Blog

March 2022

----

www.datascience.foundation

For Analytics teams, working with Secondary Data is becoming increasingly common. In an ideal world, this would involve receiving structured/unstructured data and managing the analytics activities from that point onwards without having to worry about validity, accuracy and origins of the data received. Unfortunately, the world is far from ideal so prior to commencement of the data analytics activities, it is essential to ensure that Secondary Data (SD) at our disposal is fit for the purpose we require it for and not been tortured already. This article has been written to provide a brief overview of the SD validation issues as well as strategies for addressing those issues.

When validating secondary data, particular emphasis should be placed upon:

- Ensuring Data Quality
- Data Usage/Ownership Rules and Rights
- Data Currency

Ensuring Data Quality

With primary data, we can have all of data collection and sorting parameters tailored to our needs from the very start. For instance, we can automate the process of excluding data that does not pass sovereignty or formatting requirements. We can test-run formatting activities on freshly collected data to confirm that even if its ‘’Big’’ there will be no issues in bringing the data sets to common denominators required. Last but not least, we can adjust data parameters and conditions at any point in time so if the initial data collection needs are not established accurately or the analytics environment changes, our data collection and validation processes can be adjusted accordingly. With SD, we do not have such flexibility and have to accept/reject whatever data is available.

Furthermore, with SD, validation options are limited to contacting data owners/collectors to confirm the values provided or alternatively carrying out own validation activities. Needless to say, validation of all of the data sets is likely to become a lengthy and complex activity. Yet, there are still going to be plenty of potential discrepancies to deal with.

So what is it be done? First of all, it is essential to examine not only the data sets, data collection mechanisms and sources of SD but also do ‘’background checks’’ on SUPPLIERS OF THE DATA! SD should be coming from reputable providers that can be trusted both with good intent (e.g. low tolerance for data torturing) and professionalism with the data collection.

Data Usage/Ownership Rules and Rights

Prior to commencement of our SD analytics activities. It is essential to check whether there are limitations to how we are to use the data as well as whether the data has been provided to us on a legitimate basis. One common data ownership-linked problem that complicates data collection and usage activities in the digital age is sovereignty of the data. For example, analytics study on a UK-based/registered online retailer's (and digital ventures too have a physical home address) customers is likely to result in having to work with data of customers that are based across the globe. Likewise, it may be working with a range of distribution partners that deliver products to customers' premises. In both cases, there may be variations required for data usage and handling. The Analytics team needs to be aware of these variations and take them into account.

As well as the legal turmoils, there could also potential ethical issues to deal with if SD sources have not been acquired in a transparent manner. Many organisations collect data in perfectly legal ways ... but the ethical challenges continue to emerge. Large corporations such as Facebook or Microsoft come to mind. Once "Terms and Conditions'' are agreed upon and signed-off, the data collected may be available for further sharing or reselling. However, there are still questions to be raised. For instance, it is well-known that many people click on ''I agree'' on Terms and Conditions documents without having a comprehensive read through. Just because these Terms and Conditions may include permissions for sharing or reselling the Data, further negative implications should certainly be considered.

Data Currency

Data is becoming outdated fast. By the time analytical work on SD commences, there is always possibility (unless it's a historic data study) that the SD has become outdated already. The trickiest aspect of handling this issue is that it could have been valid at the time of acquisition but the time span between data acquisition and the data works (and it could be a very short time span) may have taken some, if not all of the validity away. This is why with SD, we should be particularly careful checking the ''expiry date''. Data validation should include not only dates for collection and processing but also processes for confirming that the data is still reflective of the environment investigated – just like it was at the time of initial collection!

To sum up, validation issues outlined above are no reasons to abandon usage of SD in favour of collecting your own ''where possible''. There are many analytical instances where SD provides is the only type of data to be used. There are also instances where primary data collection may result in greater discrepancies than SD processing. However, it is important to acknowledge that validity of SD data sets should not be taken for granted. If we are running analytics projects, it is OUR job to validate all the data used irrespective of whether it is primary (aka collected by us) or secondary (aka provided). If only the SD is not accurate or current, no matter how magically we work with it, in the end of the day, the analytics project will be a failure!

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email: admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation