Raw Data Collection 2020: Principles and Challenges

Author, Michael Baron

A Data Science Foundation White Paper

January 2020

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: admin@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

Raw Data (also known and often referred to as Primary Data) collection is the starting point of any data analysis. Once the RD (Raw Data) is collected, it is processed to turn it into *Information* that can be converted into *Knowledge* further down the analysis track. The purpose of this White Paper is to explain the key RD collection principles and challenges and how these principles and challenges are evolving in the light of the emerging technologies and business processes.

Raw Data Collection: The Traditional Approach

All of the traditional approaches to the RD collection involved a fairly straightforward 5-step process:

1. Establishing what information/data needs to be collected
2. Setting a timeframe for the collection
3. Establishing a method for the collection
4. Collecting the Data
5. Sorting the Data

For many years, Steps 1 to 4 have been considered trivial to carry out. For each of the data types, there have been standard collection processes that could be implemented without much difficulty. Within many industries, set industry-wide standards for data collection and formatting were available. Some of the industries where data collection has been standardized to a dominant extent were: Banking & Finance, Travel, Resource Industries, International Shipping & Trade. Data Scientists could develop so-called ‘’templates’’ and apply these templates as required. On the uncommon occasions where no template appeared to be fit for the purpose of the data collection due to project-specific requirements, new templates could usually be developed by the means of re-engineering the currently available ones. Building templates from scratch was rarely a requirement.

Quantitative data collection was particularly easy where the data sources were standardized. Therefore, the simplest way of collecting such data was:

Step1: identifying relevant RD collection standards and practices within the respective industry followed up by:

Step 2: Establishing the RD parameters based on these standards and:

Step3: Commencing the RD collection from the data sources selected

The data collection and processing tools were easy to link to the data sources based on unifying features of these sources. Identifying such features involved finding common denominators across the data sources/sets. When the problems, did occur – it was usually in cases of cross-industry data collection projects since consistent denominators were difficult to establish. Otherwise, even in projects where multiple data collection sources were required, the denominators seldom varied.

Test of Time: Challenges to the Traditional Approach to the Raw Data Collection

As time goes by, new challenges to the traditional RD collection processes keep emerging. One particularly significant challenge has been *growing persity of the data sources*. The persity is hard to eliminate as it tends to take place NOT only due to inconsistencies in the data handling but at a ''higher'' level. Diversity of the data sources alone would not be too difficult to manage if it would not be linked to *persity of the business processes and practices.* For example, many organisations are selling their products and services both online and via traditional channels. Not only online and traditional channels are likely to have different data sources and to require different data collection methods, but they may also be based on totally different business processes. If the business processes behind the data sources differ, the data integration becomes significantly harder.

Establishing Primary Keys

Integration of the data sources and bringing these sources to common denominators has traditionally been done by the means of establishing Primary Keys. A Primary Key is a pre-defined choice of a minimal set of attributes that could be used to identify an object or a record. In simpler cases, it could be linked one single attribute. Back in 1973, when Charles Bachmann defined concept of the Primary Keys in his legendary Turing Award Paper ''The Programmer As Navigator'', database development and consequent data management practices made a giant leap forward. Contemporary databases still utilize Bachmann's concept where possible and having such Primary Keys makes identification of data sources/data collection from the sources easier.

However, to be able to set up highly effective Primary Keys, data scientists need to be able to

determine the core attributes. In his recent DataTalk article, the author already discussed some of the difficulties in dealing with the data set disparities and pointed out that extreme efforts to integrate poorly compatible data sets with one another is no different from ''comparing apples and oranges''.

When dealing with the RD, the challenge could be even greater. RD is unsorted data. It is uncategorized, unsorted and unclassified. It is not arranged via specific groupings or preliminary attributes (not to be confused with the attributes that analysts establish when they commence the data-information-knowledge process. Such data would be likely to come in a very perse range of formats. The greater the differences between the formats are, the harder it is to establish the Primary Keys.

Data Collection Tools: A Single Tool is Rarely Enough!

As evident from the discussion above, differences across the data sources could be quite dramatic. A wide range of types of the data sources is likely to lead to the need to use a proportionally wide range of data collection tools. It should be noted that *quantity* (overall number) of the data sources is not a significant factor but *quantity of types* of the data sources is!

Having to use many data collection tools presents both technical and operational challenges. The technical challenges involve identifying ways to streamline all of the RD collected towards a unified structure. Operational challenges for the RD collection could turn out to be even more stressful to deal with. With every additional type of a data source, cost of the data collection will keep increasing and so will the time frame required to complete the collection.

If RD collection involves usage of a significant number of perse data collection tools, it may challenge accuracy of the analysis by the fact that data scientists/analysts are likely to be familiar with some of the tools but obviously not all of them. Therefore, if the RD collection is carried out across a variety of data types, it may require involvement of larger work teams that include expert users/practitioners for each of the tools. The total cost of such multi-sourced data acquisition may turn out to be several times higher than in cases of single-sourced data collections.

Establishing RD Collection Time Frame & RD ''Life Expectancy''

In the light of the emerging technologies, time is starting to run ''faster'' and frequent updates

to the data sets are required. After all the efforts we put into multi-sourced data acquisition, we are still going to have data sets that are likely to become outdated fast! Currency of the data should be ensured throughout NOT ONLY the RD collection processes but also while the consequent data related activities (knowledge generation, data-driven business process re-engineering or development etc.) are taking place. There is little point in carrying out all the analytics activities to come up with findings that are no longer representative of the environment analysed.

It is becoming increasingly common for Data Analysts to turn RD collections into never-ending processes that include continuous updates to the data sets. Such ''Live'' updates do have a lot of merit. However, they also make it more challenging to get the data analysis projects completed and the outcomes established. While on the one hand, Live updates to the data sets guarantee ongoing currency of the data throughout duration of the data analysis projects, on the other hand – it is hard to ''close'' projects that keep receiving ''updates''. If the RD collection is part of a timed business project rather than an ongoing study, there got to be some time allocated for the data analysis ''beneficiaries'' to utilize the findings. Should the findings continue to receive adjustments based on the new data emerging, they may become more difficult to act upon.

RD Collection: Future Trends and Challenges

As John Legend pointed out: ''The Future has started yesterday, and we are already late''!

As time continues to goes by, we are likely to face a number of future challenges that (at least for now) appear to be inevitable, namely:

- Data Life Cycles will keep getting shorter and shorter
- Variety and Diversity of the Data Sources will continue to increase
- Data Ownership and Privacy- related complications will emerge to an even greater extent
- Live Data accumulation pace will be getting faster and faster

To sum up, as evident from the trends discussed above, RD collection and pre-processing Is getting more complex. In the light of these trends, Data Scientists need to continue looking out for ways of optimizing the RD collection and handling processes!

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