Data Onboarding

Data onboarding—the preparation of unfamiliar data from disparate sources, both internal and external to the organization—is a complex process. Whether it's combining multiple streams of marketing data into a singular dashboard or augmenting customer data with third-party information, data onboarding involves huge volumes of data and repeats every time new information and sources are found—in some cases, daily1. To provide an example, data onboarding is one of those many processes in digital marketing that is, essentially, a black box. To help shed light into this particular black box, management consulting firm Winterberry Group has released its first report on data onboarding in the US, *The State of Consumer Data Onboarding: Identity Resolution in an Omnichannel Environment*2. Data onboarding begins with understanding the source data's purpose and structure, continues with the assessment of the data's quality and standardization across sources, and concludes with the combination of multiple data sources into a consistent view for analysis, or for import into your business applications. For most analysts, this process consumes the bulk of their time—on average, analysts report that up to 80% of their time is spent preparing data for analysis3.. Data onboarding is a laborious and time-consuming process because of *unruly* data. Gathering multiple types of data to combine into one standardized format is difficult for several reasons:

- **Data silos create duplicates**. Whether due to departmental differences in applications investments or mergers and acquisitions, data typically lives in silos. Data will often be duplicated in distributed data marts to respond to specific user demands, which makes it challenging to bring that data together
- **Siloed datasets have varying formats and standards.** Because of these data silos, each dataset often has a different format and standard, which makes them difficult to combine. This is made even more challenging when the data comes from third-party vendors, customers, or public data—in that case, analysts have no control over the data structure and norms used in the data, requiring them to decipher elemental data to make it consistent across sources to finally combine it together .
- **Outliers and errors are difficult to spot in large, unknown datasets.**Buried within all datasets are inconsistencies, such as an age value of 250 years, and errors, such as an invalid zip code or SKU format, but these are even more difficult to spot in large data volumes. If these anomalies aren't surfaced until analysis, they cause huge delays to accurate insights. Data volume and formats are growing—fast.
- **The variety of available data in most organizations has exploded**, but with more data comes an increased demand to manipulate and process that data. Adding to the complexity, these new formats are usually not tabular by nature, but hierarchical—value/pairs or free forms—making it difficult for analysts to get an immediate understanding of the data.

Unfortunately, overcoming these data challenges isn't the finish line, but the beginning of a marathon. In data onboarding cases, organizations receive new versions of the same data each month, week or even day, with slight changes from the previous batch that introduce new hazards to the onboarding process4. When onboarding data, analysts face these challenges again and again in the midst of high pressure from internal stakeholders or customers to get the data right as quick as possible. Hence, data onboarding is typically either delegated to an army of highly-skilled data engineers. To get insights faster, there has to be a way to standardize and automate the process of cleaning, transforming, and combining this data quickly and accurately, without investing in expensive engineering labor.

Open in RSVP Reader