WHITE PAPER

Data Quality at 'The Edge'

The Build vs. Buy Challenge:

Insight into a Hybrid Approach

melissa

GLOBAL INTELLIGENCE

EXECUTIVE SUMMARY: ADDRESSING THE BUILD VS. BUY CHALLENGE

Data is complex. There are domains of data covering an entire universe of information, including customers, products, locations, finance, employees, assets, and more. But data doesn't exist for its own sake. For it to be useful, it has to be trusted. The data steward must ask of these domains: "Can I trust the insights that I'm getting? Are operations going the way I expect them to?" Trust is a keyword in all data quality considerations.

Too often, the accuracy of data lets us down. There are abbreviations added where they shouldn't be, company names entered in the contact name field, and main values left blank, creating multiple parsing problems. Hand-coded rules introduce more problems, requiring extensive debugging. And, with international databases, non-English words add to the challenge for U.S. data stewards.

Those who work with data are aware of the cost of bad data. It ranges from strategic problems (poor data = poor analytics = poor business decisions), to lost sales (bad addresses = returned shipments = angry customers). But how bad is it really? In a recent report in the Harvard Business Review citing IBM statistics, bad data costs U.S. companies $3.1 trillion annually. Yes, that's pretty bad.

Avoiding major pitfalls due to bad data often comes down to a choice—On the one hand, deploying in-house expertise to build data quality solutions, or, on the other hand, turning to a qualified vendor for modular off-the-shelf solutions. Relative costs are important, as is company expertise. More essentially, you must measure the potential quality and accuracy of any solution and the roles it is expected to play.

The guiding principle of this white paper is understanding the relationship between rules-based data quality, where internal subject matter knowledge is necessary, and active data quality, things that are constantly changing outside the organization. This is in the critical zone indicating when data quality is best addressed, whether in house, or with off-the-shelf solutions, or by a combination of both.

We call this zone of decision "The Edge."

UNDERSTANDING THE CHALLENGES

Deciding between Build or Buy means grappling with what data quality actually means and for whom. Data quality itself can be broken down into six dimensions:

1. **Completeness.** Is all the necessary information available, or are data values missing or in unusual states? Do you have everything you need in a particular data field?
2. **Validity.** Do you have all the values that conform to a specified format? Street addresses, for example, have specific formats. Do your fields match the proper way of doing it?
3. **Consistency.** Here, data should align with understood patterns. A common example is MM/DD/YYYY, the U.S. standard for date of birth, but may be different in international markets.
4. **Accuracy.** How well does the data reflect the real-world person or object being identified. It's probable that "Mickey Mouse" is not a real sales prospect.
5. **Uniqueness.** Do you have multiple, unnecessary duplications, for example, the kind that mistakes one contact for multiple contacts?
6. **Timeliness.** How stale is your data? Considering the enormous number of changed e-mails, household moves, and recycled phone numbers, it's no surprise that 25% of marketing data goes stale over the course of a single year.

Many roads to perdition

How do data systems go awry? Faulty data entry is a major culprit.

People mistype and misspell all the time, but sometimes a system crash can corrupt files as they're processed through the system. And, several other of the six dimensions of data quality can be encountered. There's a record violation where a birthdate doesn't match the contact's stated age. Or, perhaps it's a rule problem, with a date that references a 13th month, for example. Missing fields are common. Uniqueness problems arise when you have two separate individuals supposedly sharing the same social security number. And, sometimes things are just flat wrong. A data entry staffer enters John Smith in Dept. 127 … but there is no Dept. 127.

Equally common are transposed telephone numbers, or nulls like a computer default phone number of 9999. Multisource problems occur with the relationship between two different tables. For example, two separate entries—Christine Smith and Chris L. Smith—may be the same person, or not, or indeed be of the opposite sex. There are multiple inconsistencies in the way names are spelled out.

The data challenges are seemingly endless. Here are few more common ones:

- **Appearance.** Lower case letters may be used when upper case is called for.
- **Duplication.** Address composition seems straightforward, but streets and roads might be mistakenly entered one for the other. Now you have the same person entered twice, once on a street and again on a road.
- **Formatting.** With international databases, phone numbers and street addresses can become confusing and varied. Transliteration and the actual script being used (Cyrillic, Arabic or Kanji anyone?) compounds the issue.

The benefits of clean data are obvious. Once it's standardized, cleaned, geolocated, formatted correctly, and more, it becomes easy to blend that data with demographics and firmographics to enhance analytics, business drivers, and decision making.

TO BUILD OR BUY? LOOK FOR 'THE EDGE'

In arriving at the right decision whether to Build or Buy, companies typically create a data stewardship team to analyze and assess the organization's data, and to understand what they're going to do with it. They may start with a master repository of everything, migrate it into a data warehouse, and identify the data and its purpose, where it came from and all the governance around it. They also may look at what levels of problems they have, and define some workflows or mapping rules to dive down a little bit more.

Rules-Based vs. Active Data Quality

A key factor in deciding between building an in-house data quality solution or seeking solutions from a vendor—and it's perhaps the biggest single issue—lies in the relationship between internal rules-based data quality and active data quality. For rules-based data, Build may be best. For active data, a rule of thumb is to consider the Buy option.

A hybrid approach based on these practicalities is key. Determining where rules-based data quality ends and active data quality begins is essential to determining the DQ "Edge" within your company.

RULES-BASED DATA: KNOWING WHAT YOU ALREADY HAVE

Just because data is internal—relatively "passive," well-managed by rules and without suffering the many unforeseen changes that befall active data—doesn't mean it's any less important. In fact, rules-based data quality is essential to any organization, and may indeed be handled best by the Build approach. Here, companies track and monitor data that already exists within internal domains, identifying and repairing any variations to its own data rules. This validates, repairs and protects data quality.

We're talking about data quality that helps companies understand and analyze best-practice KPIs, among other things: Internal data that reports on, for example: employee performance; metrics and retention; moving new product introductions more quickly to market; correcting internal customer records and billing information; optimizing supplier payment terms; and reducing inventory, among many. It's about continuing to enforce data rules throughout an organization that data stewards can keep errors under control. This is rules-based data control.

You can build it but should you?

- How does the cost of building and maintaining your data match up with the cost of a commercial solution?
- Consider time-to-implementation; which is quicker and more effective?
- How does the quality and accuracy of a home-grown solution compare with a refined, tested solution?
- Is the issue unique to you (Build), or one that's been solved many times already (Buy)?

Where Build can go wrong.

Many of the common pitfalls of the Build solution, no matter how appropriate it might be for the task at hand, are obvious to most data stewards: The program began six months ago and is already late; the budget is stretched; it's difficult to meet the changing needs of users and departments. Sometimes the home-grown solution doesn't meet user-acceptance testing, or (as bad) doesn't work well with new, up-to-date systems. Overconfidence in the face of truly tough challenges is another Build pitfall. Writing fuzzy matching or an address validation engine in-house is an enormous undertaking often done by generalists without years of specific experience.

Where the Build option most often goes wrong, however, is in trying to manage active data, the kind that's always changing.

ACTIVE DATA QUALITY: A WORLD IN MOTION

Active data quality concerns information that has a relationship with the "real world" of ever-changing circumstances, things that are constantly moving outside of the organization. Included here are things like customer or prospect addresses, phone numbers, emails, names, companies, nationalities, and job titles. In the real world, contacts and their information change constantly: Has someone gotten married or changed names, or perhaps even died? Did they move to a different company, or is their business defunct? These are things that typically are outside of the control, or even knowledge, of an organization.

And, active data is enormously complex. Here, the concept of "fuzzy matches" rears its head to make sense of apparently non-similar things. For instance, these two addresses are really the same: 6801 Hollywood Blvd, Hollywood CA and 6801 Hollywood Blvd, Los Angeles CA. Hollywood is a vanity city name, and Los Angeles is the USPS® preferred city name. Failing to match these two records would lead to duplication problems.

Standardizing company names, which may have any of dozens of appropriate suffixes, is particularly difficult with international databases. As well, the honorifics of particular persons (Mrs., Mrs. Ms., Dr., The Honorable), or the variable contact peculiarities in Berlin or Moscow or Beijing, make for constant hair-pulling. And there's the issue of householding, that is, managing things like marriages and divorces. Yes, some of this can be done via a Build process, but it may be handled better by plugging in the right tools that have been in development and constantly improved over years and years.

Change forces change

There are times when a changed environment drives the Buy option:

- It may be an urgent need that needs to be addressed immediately, such as excessive losses in mailing and shipping.
- Or perhaps it's a new compliance rule or best-practices guideline, common in the financial services and healthcare arenas.
- International expansion triggers the need to get truly serious about data quality. Naming, addressing, and even script differences need to be parsed.
- Is the issue unique to you (Build), or one that's been solved many times already (Buy)?

Where Buy can go wrong.

Buying a data quality solution also has its pitfalls. One of the most common is when companies overbuy, meaning they acquire an entire suite of data quality tools and APIs when they're not ready to adopt them into a coherent data quality solution. A company can go through an entire year "sitting on" the solution before desperately reaching out for technical assistance—or worse, simply giving up and abandoning the solution entirely. Also, while budget is always a factor, buying a data quality solution on the cheap may not scale as the organization's size and needs grow. The product may then have to be abandoned entirely.

Another Buy pitfall is departmental siloing. A particularly innovative department may reach out to a vendor to solve its data quality issues, but leave out other departments in the company that also need help. Many e-commerce operations require very good contact data, but so do their colleagues in marketing and sales. Buying for one while ignoring others doesn't do the entire organization very much good.

Just as the Build versus Buy question depends on how the data quality solution is to be used, there are pluses and minuses to each approach. Understanding these tradeoffs can provide useful guidelines as to which approach to use. Or, much more likely, how to blend the two into a cohesive, effective whole.

WHEN THE 'BUY' OPTION IS RIGHT FOR RULES-BASED DATA QUALITY

While any decision between Build and Buy must weigh circumstances with the rules-based data versus active data being perhaps the most essential consideration it doesn't have to be either/or. Even rules-based data quality can benefit from the Buy option when it's determined that off-the-shelf rules-based tools can manage internal databases more efficiently than in-house solutions.

Data stewards looking at the Buy decision here will want to look for vendor tools for parsing, formatting, verification, fuzzy matching, and generalized data cleaning. Specific examples include the following:

- Data completeness rules: If the address is present then the postal code must be present.
- Data quality rules: If the phone number is correct then it must be callable.
- Data domain rules: In a column of U.S. states there must only be 50 values. If a field for a person's age exists it cannot contain a number higher than 150.
- Attribute dependency checking: If a loan is funded, its amount must be greater than 0.
- Data consistency rules: IBM, I.B.M. and International Business Machines all refer to the same company, and will be standardized during processing.

Melissa offers rules-based solutions that are packaged into single tool sets. For example, Melissa's MatchUp de-duplication tool has some 16 different fuzzy matching algorithms that can be mixed and matched. Melissa's Generalized Cleansing tool bundles six different cleansing operations, including punctuation, abbreviation, search & replace, and regex (regular expressions) to help clean and prep data. The convenience and efficiencies of rules-based tool sets available from a qualified vendor may prompt a Buy option.

A GUIDE GOING FORWARD: TAKE A HYBRID APPROACH

Strong data quality outcomes are best realized using a hybrid approach, both building and buying depending on the situation. Data stewards should build and buy tactically. Anything that is internally rules-based or deals with the organization's own data may be better managed by a Buy approach. Perhaps a generalized cleansing and a few parsing engines will be adequate here, using custom rule-building. Build where internal information and subject matter knowledge is necessary.

Everything that's active data is much better with off-the-shelf solutions and vendor products. Buy when external reference data is involved, anything having to do with the continual verification of information.

Leverage cleansing and validation tools. If a vendor has a powerful parsing engine then buy it. Likewise, buy customer matching—it's a very difficult job. Golden Record Survivorship, which is the consolidation of duplicates after they're identified, is a challenging process often beyond the abilities of even proficient in-house teams. That's something you buy.

Re-evaluate your options every year or even more frequently, making sure your criteria are still being met.

Finally, even internal rules-based data can benefit from the Buy option, with the convenience and mix-and-match capabilities of ready-made tool sets from a best-of-breed vendor.

VENDOR RECOMMENDATIONS FOR THE 'BUY' OPTION

When choosing the Buy option, vendor research is critical. Here is a compendium of guidelines and rules of thumb that can help you negotiate the marketplace for not only the right solution, but the right vendor partner:

- Does the vendor have customer references?
- Do they have case studies that apply to your industry?
- Does the vendor offer a one-stop solution for more than one domain of data quality?
- Ask about the vendor's entire range of product information management tools.
- Use social media to research vendors. Peer-review social ratings can be invaluable, in particular in a technical environment. Ask about vendors, ask about their particular tools, ask about negatives.
- Consider if your prospective vendors have security or compliance certifications. Likewise, consider data standards that the vendor abides by.
- Ask about the age and sources of the vendor's reference data.
- Ask about how the vendor resolves duplicates.
- Determine how the solution can be delivered (cloud, on-premise or both).
- Ask about customizable and scalable capabilities.

Go for small wins

Starting too big often means you can't quickly demonstrate that your decisions make sense, and that you're spending time and money well. Strive for small wins. Solve one problem at a time in discreet phases, focusing on a problem or two, then a third and fourth later to start showing results quickly and getting buy-in from management. Then rinse and repeat.

ABOUT MELISSA

Melissa is a leader in data-driven solutions that help organizations leverage Big Data and People Data (name, address, phone and email) to unite customer insights, analytics, data quality, and cross-channel marketing. We profile, cleanse, verify, enrich, and consolidate data assets, providing more than 10,000 brands in over 20 countries with accurate, reliable, and trusted information that can be utilized throughout the enterprise. For more than thirty years, our extended legacy in data quality, ID verification, and data enhancements has earned the trust of organizations from around the world.

1-800-Melissa (635-4772)

www.melissa.com