Is It Really "Bad" Data? An Argument for Usable Business Metadata

The Naysayers

We have all heard the lines. "I can't trust anything from [*Insert creative system name here*] it takes hours to scrub the data before we can use it." Or "Our department needs its own data warehouse because theirs is full of bad data." And of course my favorite, "Their totals are wrong, ours are better."

It's the problem of bad data. As a result, vast sums of money are spent building new and better custom warehouses and datamarts with the intent of 'cleaning up the data' so that it can be used by the business and analyst communities. Spreadmarts evolve with the sole purpose of creating SME (Subject Matter Experts) fiefdoms in which we can't lose certain associates because they are the only ones who know how to process the data for a report. Endless arguments about the accuracy of one report versus another will plague conference rooms, burning countless hours of wasted time.

But is the problem really with the data? Is any data really *bad*? Let's assume for the sake of this argument that when transaction systems are built, they process events correctly and that true data errors are corrected before a system reaches production. In this case data produced from business events must be accurate or at least accurately reflect the occurrence for which it represents. So what is the problem? This essay attempts to shed light on perceived data quality issues and how accurate metadata can eliminate the perception of bad data, without any fancy data quality tools.

To be clear, of course some data is bad. Certainly systems that allow; null fields, free text fields for codes that should be validated, numeric fields that allow characters, etc are truly bad and need correction. My favorite example involved two systems that tracked customer data for different market segments. Because they evolved from the same code base they both used a 6 digit numeric field for customer number. One system assigned the number randomly, the other sequentially. Whenever there needed to be analysis performed on the merged data set, there would be cases of customer number overlaps where the same number could have two different names representing different entities. A third system was created to resolve the duplicates. In that same case, product codes were classified differently in each system.

Obviously, there are cases where data has issues because analysts need to use the data in ways that the system designers had not foreseen. In my example above, each of the two customer tracking systems reported accurate data. It wasn't perceived as bad until it was merged. Another example, when I purchase an airline ticket, if I use "Josh" vs "Joshua" there is a potential for a logical duplicate in the booking system. The system processes my ticket quickly and easily; there is no need to complicate my purchasing experience with a complex verification process. Lastly, I confess that I sometimes use fake birthdays so that websites can't use that as a data element for a phishing scam. Data can be bad, often really bad.

This paper addresses the very specific case where transactional business data is made available to analysts. The analytical environments often enable different interpretations of the data when there is a lack of understanding of what the data is and the rules used to classify that data. In some ways, this eventuality is understandable. Any customer facing system must be designed with performance and flexibility to best suit a customer experience. There is little regard for incorporating analytical requirements. Anything that slows a rollout of new functionality is costing money, so we worry about analysis after the fact. On the analysis side, when development funds are tight, there is usually little appetite to spend the appropriate funds to do a full requirements and governance workup of a data warehouse only to support an activity that usually does not directly generate revenue.

But there can be a third way. By using metadata, we don't need to go overboard on cleaning data, and we don't need to build business rules into ETL processes to make sure the analysts see the data only in the way that has been approved. Interaction with transaction system designers and SMEs is important, but should not be a roadblock. We can let consumers see all the data with the transparency necessary to enable thoughtful dialog and meaningful analysis in the correct context.

Business Use Case

Most people have shopped for something from an online retailer. Recently I bought some heat packs for my daughters when skiing. Through the website I ordered 40 feet warmers and 40 hand warmers. I shopped for both items at the same time, checked out, paid by credit card, and a few days later 2 separate packages arrived in the mail.

From the retailer perspective much more happened. The 40 hand warmers went through without incident. A sub-vendor (with positive inventory) was matched by the online retailer. The order was placed and the sub-vendor shipped the product. But the feet warmers had a more complex path. The first sub-vendor indicated a positive inventory position (>1 @40 pieces), so the online retailer forwarded the order. Within a short time the sub-vendor cancelled the transaction, there was an error in the inventory system. The auto-match algorithm by the online retailer went to the second tier sub-vendor and finding positive inventory, placed the order and it was shipped.

Assume this is the only transaction for the month. There is a senior level meeting at the online retailer. Sales volume is the topic of conversation. The CIO reports 3 sales transactions, she ran a query on the number of transactions processed by the order-entry system. The marketing VP reports 2 sales; from her position they sold 2 different products. The finance VP reports 1 sale, he sees one customer, one sale of multiple items. Obviously, this is not a fun meeting.

Bottom line, there are multiple ways to view this simple and common transaction. Data can be over-simplified, viewed as incorrect, or even used to draw flawed conclusions. But none of these potential problems is because the data is actually bad or incorrect. In this scenario, "bad data" is perceived. In many organizations, the order-entry system will produce a 'common-feed' of sales data and provide it to the departmental analysts. They in turn will "scrub" the data to group it by the appropriate dimension, delete unwanted entries (like the failed sub-vendor record) and count the number of sales their group feels is correct.

Governance Must Be the Solution

A common solution to the problem described above is to convene a cross-functional data governance team. The team spends months collecting requirements, arguing business rules, translating them to technical rules and certifying a one-size-fits-all metric of "Sales Transactions". With this in hand, everyone can "compare apples to apples" as this is the only metric allowed to be reported and used as a denominator in various KPIs (Key Performance Indicators). But even if this governance effort is successful, are we serving the business?

This is a vast oversimplification, but the story probably bears some resemblance to a scenario at most large organizations. This is especially true among analysis teams. There will always be differences in the way various groups in an organization view similar data sets. Many times each will think that they all agree on a business definition for an operational metric. But when it comes to attaching data specifications to that business definition, problems arise. Governance often aims to simplify this reality. But I would argue that it need not. Complexity is good. The trick is to get people to understand how and why data appears to be complex. That training could be a much better use of time than data governance meetings.

Note that I am not a complete cynic. Governance can and should play an important role. Imagine 3 different distribution channels of a retail conglomerate. My use case above refers only to the online retailer, but there could also be a snail-mail catalog/phone channel as well as a brick-and-mortar store. When inventory is shared, how do you calculate cost per square foot at each of the three channels? Clearly there are cases where enterprise standards need to be applied to operational metrics and KPI's. The point of this essay is to illustrate that there can be strategies to minimize the need for the G&A (General & Administrative) tax imposed by governance efforts, simply by facilitating increased transparency through the use of metadata.

Verdict

Our goal is to enable data analysis that serves the business regardless of nuance. Imagine if the CIO labeled her metric "Sales Transactions Processed" and the marketing VP labeled hers as "Items Sold". We could then have a KPI called "Transactions per item sold". Right now the value would be 3:2 or 1.5 The business does better as this number *decreases* and approaches 1, indicating better vendor coordination as we are able to fill orders more efficiently. Neither of these two metrics, nor the derived KPI have any relationship to the KPI "Net Income Per Sales Interaction" which might be the preferred measure for Finance; ideally it is always *increasing* as "Sales Interactions" is the dividing metric.

It is imperative that operational metrics have unique naming, business definitions, and usage characteristics so as not to confuse the data consumer. Data lineage details in business terms must be tied to each record, so that there is no ambiguity as to why a data set is the way it is. Furthermore, usage of proper terminology when discussing or presenting results must be enforced so that the proper data set is used for the proper purpose. Avoiding contentious meetings, might be the single greatest outcome of a solid metadata campaign.

Knowledge Is Power

Let's face it, what we data-geeks do is really pretty simple. We consolidate data from one or many transaction systems, map it to coding that makes sense for analysis and put it into an environment that is friendly for reporting. All the rest of it just makes these basic steps easier. The problem is that business people don't understand what we do. In most cases, they don't need to know, and probably shouldn't. But what they really want to know is two things;

1. What is this data I am looking at? What is the definition, how is it calculated, how should it be used, etc
2. How did it get here? What is the source system, what rules were applied, who certified accuracy, etc.

We can start with the data for my heat packs. Figure 1 shows a sample of the common-feed data that might come from the order entry system.

Transaction Date	Invoice #	Customer	Item	Sub-Vendor	Sale Status	Cost
Jan 21, 2011 2PM	101654	Josh	Hand Warmers (40)	Joe's Handwarmers	SHIPPED	$22
Jan 21, 2011 2PM	101654	Josh	Feet Warmers (40)	Joe's Handwarmers	FAILED	$18
Jan 21, 2011 7PM	101654	Josh	Feet Warmers (40)	Aldo's Ski Shop	SHIPPED	$18

Figure 1. Common-Feed Data

At this point we can say the data is consolidated. From a cursory scan we start to see the interaction between the data and how the various departments produce their statistics. By now we should also understand the business rules from the departments. This is also a great space for a Governance team to operate. They don't need to verify the wisdom of different measurements, only that the logic is unique and unambiguous. Ideally business

rules stand by themselves and incorporate the English version of the rule as well as codifications of that rule that can be used in the ETL. (Rules driven ETL is the subject for another paper) The business rule indicates how the data was processed by the system. It helps answer the question "How did the data get here." Figure 2 shows the three Business Rules for our case.

Rule #	English version of code values (Where clause)
R1	Any transaction that has executed at least once
R2	Item level transactions with a positive Sale amount, where Sale Status is SHIPPED or ADJUSTED
R3	Positive dollar invoices where at least one item on the invoice has a sales status of SHIPPED

Figure 2. Business Rules

In this scenario, each rule aligns with the source of the data in a one to one relationship (respectively). But this need not be the case. Multiple rule and source combinations can exist. As long as the relationship between rule and source is stored with other metadata, it can be used by the consuming analyst.

Now that we have staged the data and established rules for data processing we can apply the rules to the data. This will generate a data set that looks to have duplicates, but because business rules (aka metadata) for creating each record are transparent, the analyst knows when each data set is appropriate for reporting. See Figure 3 below for a logical representation of this process flow.

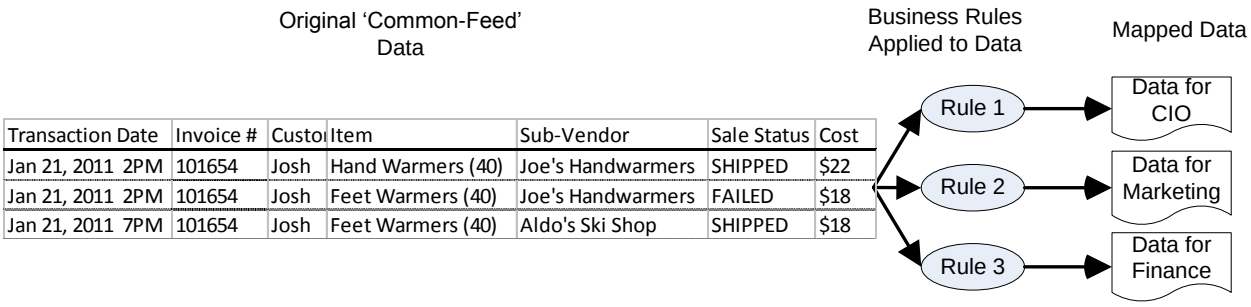
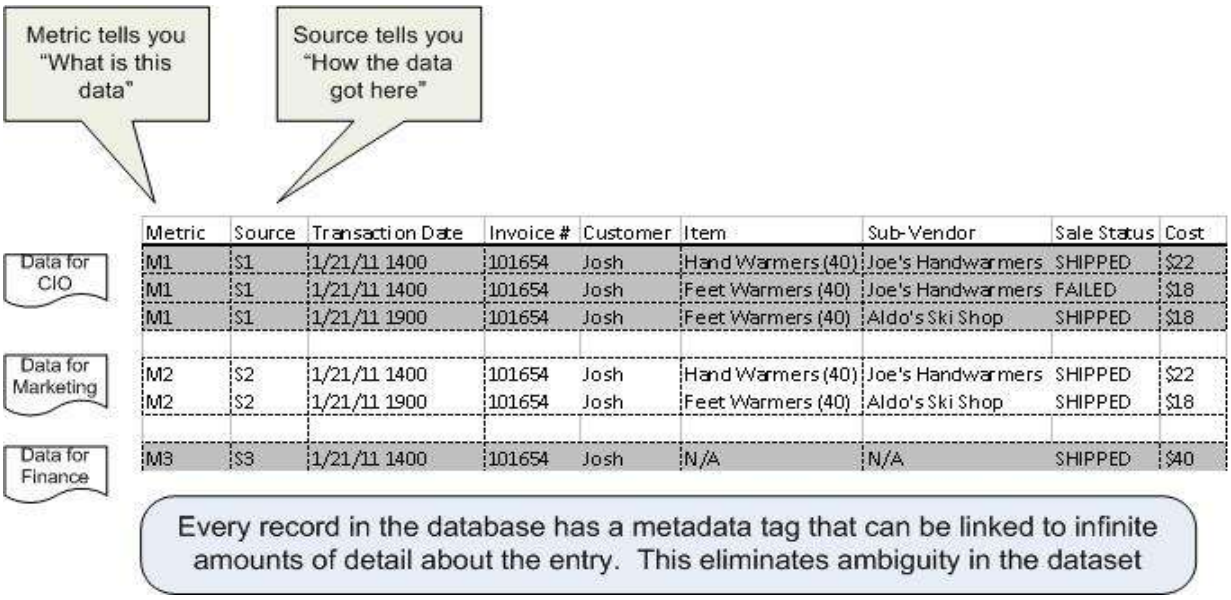


Figure 3. Process Flow

As the data is processed to "mapped data", each record gets two metatags applied to it. The metatag is a database field that allows each individual record in the database to be linked to metadata. The metadata can be stored in a variety of ways and should encompass all areas of interest for the consumer. See Figure 4 for an example of our transactional data with embedded metatags. In most cases, the consumer is not interrogating the metadata for each reporting or analysis project. But the important point is that if they need to justify a number or explain its derivation, the metadata is readily available. In this way, knowledge exists in the database and not with the analyst. That knowledge empowers all consumers of the data, not just the analyst and his fiefdom.



Metric	Source	Transaction Date	Invoice #	Customer	Item	Sub-Vendor	Sale Status	Cost
M1	S1	1/21/11 1400	101654	Josh	Hand Warmers (40)	Joe's Handwarmers	SHIPPED	$22
M1	S1	1/21/11 1400	101654	Josh	Feet Warmers (40)	Joe's Handwarmers	FAILED	$18
M1	S1	1/21/11 1900	101654	Josh	Feet Warmers (40)	Aldo's Ski Shop	SHIPPED	$18
M2	S2	1/21/11 1400	101654	Josh	Hand Warmers (40)	Joe's Handwarmers	SHIPPED	$22
M2	S2	1/21/11 1900	101654	Josh	Feet Warmers (40)	Aldo's Ski Shop	SHIPPED	$18
M3	S3	1/21/11 1400	101654	Josh	N/A	N/A	SHIPPED	$40

Figure 4. Transactional Data with Metatags

How It Works

The last step is to assemble the analytical data for the user community. Ideally we take the mapped data and make it easy for the analyst to consume. This is the art of the datamart (rhyme intended). Datamarts tend to be fluid objects as users change their mind continuously as to how they want to present the data and how best to store the data for a reporting tool. Regardless of the layout, carrying the metatags into the datamart is critical. In addition, we need to architect the mart so that it does not invalidate the metatags. What this means is that you cannot mix data values within a row of data that has an existing metatag representing a different source or meaning. For example, imagine we calculate "Total Sales per Month" and put it on each record right next to our transaction detail. Certainly this would enable an easy calculation of "Percent of Total" for each transaction. However, the *source* of the calculated total is the calculation routine. The *definition* of "Total Sales" is not the same as the definition of say "Items Sold". Adding this field would invalidate the link between the metadata and the data.

That said, a metadata model can be created and linked to each datamart object to enable the full transparency necessary for an analyst to describe, qualify and defend every number they use. An example of this model can be seen in Figure 5.

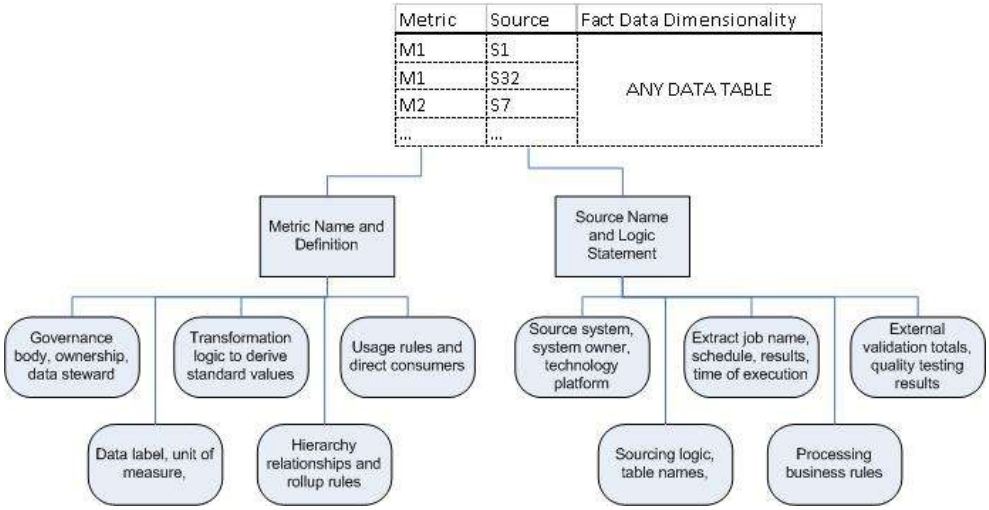


Figure 5. Sample Metadata Logical Model

Making it Real

All this theoretical mumbo-jumbo sounds great. But I know what you are thinking: In order to fund any sort of change to my warehouse, I've got to sell it to the business. Granted, generating ROI for a theoretical user experience is impossible at best. However, this could easily be quantifiable as an efficiency play. We know that people will be able to do their jobs better when there is less duplication of effort, less ticking-and-tying of data, and more certainty about the conclusions derived from data. What we are selling is *effectiveness*. So what does an effective data experience look like?

When you link metadata to every single record in your warehouse, you enable at least 3 distinct ways of accessing data that save time and make the consumers smarter as a result.

1) Parameter Driven Reports – The exact same report can be created for the 3 different metrics in my example. If I were designing the report, I would put the metric name, operational definition, business rules and other metadata right in the report header, forcing it to be printed. Since this is a simple join from the fact data to the metadata tables (joining on metric and/or source), this is very easy to develop and dynamically changes for each reporting metric. Picture a budget meeting to set funding for systems development. It would be possible for an 18 month trend of the CIO's "Sales Transactions Processed" to be increasing, even while Finance's "Sales Interactions" are decreasing- indicating more failed transactions. Instead of arguing 'how much' we strategize on 'why'. There might be justification for more IT spend, or we call in the head of supply management and fix the problem. Multiple statistics are good, let's figure out how they relate.

2) Deep Dive Analytics – As we look at more intricate analysis techniques we may use statistical tools to identify trends in the data. Six Sigma is a classic example. While we strive for operational perfection, the only way to make progress is to identify the outliers and understand why they happened. When metadata is attached to each record in the database, this becomes very simple. Out of 6 million transactions we might identify several hundred problems. Once these records are identified, the metatags create instant hooks that enable the relevant details to be reported for an operations person to research. Source system, transaction date, business rule, transformations applied, data steward of the transformation logic, etc are all joined to the data set instantly. This will make chasing and researching these data sets much easier.

The solution to the entire corrupt data set could be as simple as a bad transformation table from a new associate. Wouldn't it be better to know that before chasing down each individual record?

3) The MBA Quandary – Grad students love data. Imagine that a whole herd of them could be set loose on your data to build case studies for you: You take the credit for fixing problems that they found. In all seriousness, there is much to be said for enabling new associates and uneducated analysis teams to understand data sets without peppering the whole enterprise with numerous questions. Think of the value of separating the data consumer from; the technology team that built the warehouse, the SME who knows the transactional data, and the data steward who is responsible for enabling usage. Metadata is the way to make people self serviceable. A new user can query the metadata to find the data they want and more importantly, understand it. For example, he might search through the metric descriptions list for "Sales". The metadata tells you all the flavors of "sales", the definitions of each, where the data is stored and the dimensionality of that data. Then that same analyst can easily pivot to the actual data sets without having a single meeting where 10 people have to answer the same questions that they typed in a requirements document 5 years ago.

Conclusion

Effectiveness is hard to price, but surely it has value. If a well architected warehouse and metadata repository can make your team more effective and scale for all future data that might be added for analysis, surely we can find a way to justify it. IT resources are usually reserved for revenue generating development efforts. But enabling business people to do their jobs better, will reduce cost as efficiencies are exploited. Metadata is often overlooked in a data warehousing project. When it is addressed it tends to be a technical exercise aimed at cataloging the environment and facilitating impact analysis for DBAs and developers. Token business metadata (paragraph descriptions) is attached to tables and columns. But this is not how an analyst looks at data. They are focused on the rows of data, outlier transactions, interesting points in the trends, etc. We need to support that level of interaction to make the analysis efficient and effective. Metadata can be so much more than what is often created. It can and should be an integral part of the dataflow and the data that is consumed. More and more, business users are becoming data experts out of necessity and crave the understanding of ETL processes. Data management teams should embrace this sentiment and empower the analyst community with the information they need in self-service fashion. However, a business analyst should analyze data, not technology. Business metadata is the answer to most semantic questions of consumer data and it will usually allay any charges of inaccuracy. "Metadata is data about data", how often do we hear that? Maybe we could change it to something like: Metadata enables the knowledge of how data came to be. Not very snappy, but would it be useful?