203RSAN-150-1DL : Data Governance

Week 5 Discussion

W5Q1 Due Saturday, midnight EST

W5Q1 Due Saturday, midnight EST
by Travis Dawry - Tuesday, 11 August 2020, 7:57 PM

Discuss the methods that organizations can use to attain high data quality.

Permalink Mark read Reply

Re: W5Q1 Due Saturday, midnight EST
by Kevin Swenson - Saturday, 15 August 2020, 12:46 PM

Discuss the methods that organizations can use to attain high data quality.

There are seven steps that organizations can use to ensure the quality their data is continuously high and reliable. Reliable data provides a foundation for the analysis' complete within the organization, which will ultimately be turned into business decisions. Bad data runs the risk of missteering these decisions and ultimately bringing adverse results to the organization, losing money and misusing resources.

There are 5 criteria used to judge data quality:

- Accuracy- data must accurately describe its objective (1)
- Relevancy- data must meet the requirements of the intended use (1)
- Completeness- data should not have missing values or missing records (1)
- Timeliness- data should be updated and kept up to date (1)
- Consistency- should have established format and be used cross-functionally (1)

Organizations have to measure the data quality to ensure it is consistently achieving the goal of being high quality. In order for an organization to do so, they can follow the seven steps listed below:

1. Rigorous data profiling and control of incoming data (1)
2. Careful data pipeline design to avoid duplicate data (1)
3. Accurate gathering of data requirements (1)
4. Enforcement of data integrity (1)
5. Integration of data lineage traceability into the data pipelines (1)
6. Automated regression testing as part of change management (1)
7. Capable data quality control teams (1)

When working for Dell-EMC, I was on a team that was implementing a new data system for our R&D engineers. In doing, I was part of the data quality control team. We had dashboards made to show the amount of incomplete records that had the ability to be report on multiple ways, including the ability to see what regions were lacking data quality, what buildings, equipment type, etc. further to help ensure the data quality we had a ticketing system in place that was routed through our team to review all new records before they were established in our data system. This helped to limit the amount of duplicate and incomplete records.

References:

1. Shen, S. (2019). *7 Steps to Ensure and Sustain Data Quality.* Retrieved from https://towardsdatascience.com/7-steps-to-ensure-and-sustain-data-quality-3c0040591366

Re: W5Q1 Due Saturday, midnight EST
by Travis Dawry - Saturday, 15 August 2020, 2:25 PM

Kevin,

Thanks for getting us started this week. Even a variable as fundamental as incomplete/complete can be immensely useful. There is also a meaningful difference between Missing Not at Random (MNAR), Missing Completely at Random (MCAR), and simply Missing at Random (MAR).

Did you see examples of missing data that fits into one of these categories? Did your team interface with the engineers in defining quality?

Class,

What are the differences between data MNAR, MCAR, and MAR? How can you be sure of the classification?

How stringent are controls regarding duplicates in your organization?

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Frances Tang - Tuesday, 18 August 2020, 1:36 AM

Hi Travis,

Missing Not at Random (MNAR), Missing Completely at Random (MCAR), and Missing at Random (MAR) are the three categories of missing data distinguished by the underlying reasons for why the data are missing.

Missing Completely at Random refers to the fact that missing data are unrelated to or independent of the observed and unobserved data, meaning that there are “no systematic differences between participants with missing data and those with complete data” (Mack, Su, and Westreich, 2018). For example, if a bank sends out customer service satisfaction survey to all customers who visited one of the 100 branches for transactions and services in the past six months. If only two out of the 100 branches did not upload survey results into the information system, but the other 98 branches all uploaded a complete set of survey data. Missing data from the two branches can only reduce the analyzable population of the analytical study. Survey results of the 98 branches are considered a random subset of the full data set of interest; the missing data do not introduce bias into the study.

Missing at Random means that the missing data are “systematically related to the observed but not the unobserved data” (Mack et al., 2018). In other words, “there might be systematic differences between the missing and observed values, but these can be entirely explained by other observed variables” (Bhaskaran and Smeeth, 2014) or the known factors about missingness can provide guidance on handling the missing data in a way that will not introduce bias into the conclusion of the analytical study. Continuing from the previous example, if the data received from the 98 branches show that survey responses are mostly from customers of age 45 or older, not much from millennial. Millennial customers are more tech-savvy as they were born and raised in the digital era, so they are more likely to conduct banking transactions through online banking, much less at the brick-and-mortar branches through face-to-face direct interactions; whereas many of the survey respondents (age 45 and above) may still prefer the traditional banking channels. The probability of participation in the customer service survey is related to customer’s age, but not the quality of customer service at physical bank branches, so the missing data might not introduce bias into the analytical study of branch customer service quality.

Missing Not at Random means that the missing data are “systematically related to the unobserved data” (Mack et al., 2018). In other words, the missing data are related to important factors that were not captured or not measured. To extend the previous example, if survey respondents are mostly the customers who were dissatisfied by branch services and needed a channel to vent their disappointments, then missing the data of responses from those neutral or satisfied customers can lead to biased analysis, concluding that branch services are bad.

Different causes of data missingness can pose different implications on the conclusions or decisions made based on an incomplete set of data. Expecting absolute completeness in data can be unrealistic sometimes. Therefore, understanding and accounting for the factors of missingness is important for effective measurement of data quality and for alerting data consumers to be aware of bias caused by missing data.

Regards,

Frances

References

Bhaskaran, K. and Smeeth, L. (2014). What is the difference between missing completely at random and missing at random? *International journal of epidemiology*, *43*(4), 1336–1339. Retrieved from https://doi.org/10.1093/ije/dyu080

Mack, C., Su, Z., and Westreich, D. (2018). Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide. *Agency for Healthcare Research and Quality (US)*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK493614/

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Bob Cheek - Saturday, 15 August 2020, 4:25 PM

Discuss the methods that organizations can use to attain high data quality.

In order to define methods for achieving high data quality one must first define what high data quality is. In business processing data is collected for a reason or purpose. One has expectations on how that data can be of use for those purposes. In that regard, high data quality would mean there is a high match of expectations and use to what the consumer of the data is trying to accomplish. As Zhang states, "Therefore, we can define data quality as the satisfaction of the requirements stated in a particular specification, which reflects the implied needs of the user." [1]. It is about the consumer of the data, the expectation and the use of the data and how all they all match. "The level of quality of data represents the degree to which data meets the expectations of data consumers, based on their intended uses of the data. Data quality is thus directly related to the perceived or established purposes of the data" [2]. Hence the first problem in defining what is high data quality is very subjective because it is in terms of each consumer. Each organization will have its own standard of quality.

To determine the degree of which the data is at high quality, one needs to be able to measure the quality and how it is improving or degrading over time. There are various ways of measuring the quality which revolve around the dimensions of data. These dimensions refer to aspects of the data such as the measuring the timeliness of the data, its accuracy and completeness and the ability to the trace the data back to its source. "High data quality should contain many dimensions like accuracy, completeness, integrity, consistency, timeliness, and traceability. "[1] Knowing the desired result (matching the expectations) and how to measure the journey, one has various of methods of achieving the same results.

All these methods share in common having a data governance program, being able to measure the quality, making data an important part of the culture and the constant monitoring/adjusting of the process. It is not a one and done solution. To achieve a level of high-quality data one must be able to have predefined expectations and plans of collecting and using the data (a strategy and roadmap). Following that roadmap, a data governance programs and methods for measuring data continuously must be established and driven into the culture of the organization.

Zhang gives an example of best practices in the example of pharmaceutical industry a 5-step program. This 5 step program consists of:

1.) Determine a quality strategy and plan how to design, collect and process the data

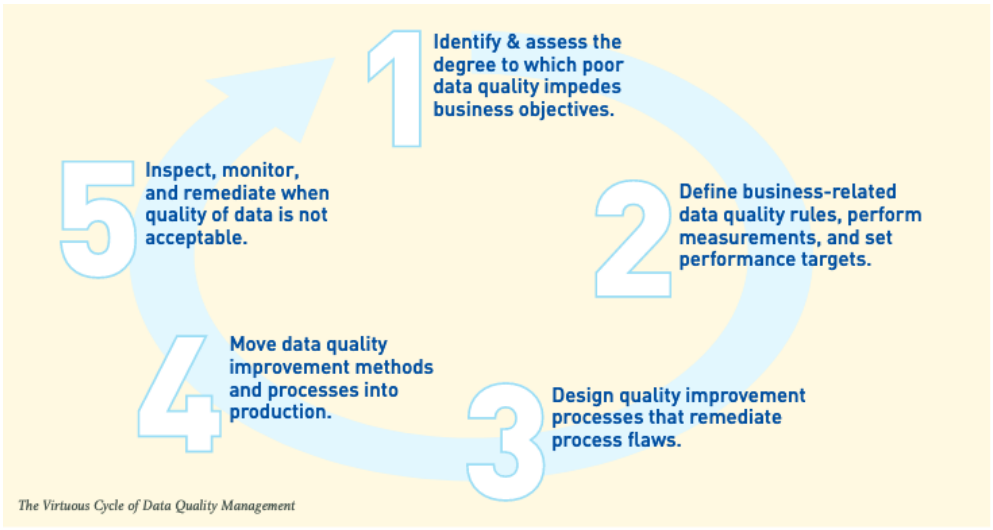
2.) Follow the data governance and data quality roadmaps

3.) Establish data quality as a structure

4.) Develop and implement necessary standards. It is through the use of standards that one can develop automation of data quality

5.) Develop a data quality tool/system [3]

This is following the same basic principles of data quality of being able to assess where one is with the data, measuring, integrating data quality into every process, and monitor and improve. These principals are shown in the following diagram:



[4] Cycles of Data Quality

References:

[1] Julia Zhang, "Chapter 4: Operationalizing, Data Quality through Data Governance." *Data Governance: Creating Value from Information Assets,* by Bhansali, Neera, CRC Press, 2014, pp. 66

[2] "Chapter 4: Data Quality and Measurement" *Measuring Data Quality for Ongoing Improvement: a Data Quality Assessment Framework*, by Laura Sebastian-Coleman, Elsevier, Inc, 2013, pp. 40.

[3] Julia Zhang, "Chapter 4: Operationalizing, Data Quality through Data Governance." *Data Governance: Creating Value from Information Assets*, by Bhansali, Neera, CRC Press, 2014, pp. 83-85

Striving for High Data Quality using MDM and MDR
by Graham Waters - Saturday, 15 August 2020, 6:30 PM

Striving for High Data Quality using MDM and MDR

Graham Waters

Prompt: Discuss the methods that organizations can use to attain high data quality.

Bhansali (2014) asserts that to measure the quality of any set of data, one must evaluate how well that data represents what was expected of it and if it meets the needs of the intended recipients. She goes on to list "accuracy, completeness, integrity, consistency, timeliness, and traceability" (Bhansali, 2014, p.66) as integral components of high-quality data. A different definition offered by Sebastian-Coleman (2013) states that data quality measurements have one overarching purpose. They serve as a form of backtesting that, using consistent terminology, give data analysts a way to compare guesses about data to the real state of that data (p. 53). In-line quality measurement and reassessment are also important pieces in the quality machine, according to Laura. Another consideration is that to attain something is to reach it as a goal, and the essence of data quality is that it is just abstract enough to be unreachable. "Data Quality is a journey, not a destination" (Bhansali, 2014, p. 82). With these definitions and concepts in mind, we may benefit from redefining the question as follows: *What are some of the methods organizations utilize to maximize the expectability of their data, minimize the differential between expected characteristics and reality, and measure how well it meets the needs of intended consumers?* This question necessitates an answer that is far more multifaceted and comprehensive than one discussion can supply; In fact, an entire book could be written on it. For this discussion, I primarily consider two of these methods: master data management and metadata-driven architecture.

Maximizing Expectability of Data

Suppose a data model is created with the intention that it will spit out blue circles; that is what is expected of that model. If it begins to produce red triangles, then it is highly likely that there is a data quality issue somewhere in the model. Applying standardization methods will help to maintain expectability within data.

Enforcing Standardization of Data

A big step towards making data more expectable as well as more useful is the standardization of that data. To standardize a data set simply means that that data set is internally cohesive or homogeneous (Ferguson, 2018). Two methods that are used for standardization are metadata-driven architecture (MDR) and master data management (MDM).

Metadata-driven architecture (MDR)

Organizations use a metadata-driven architecture to generate a single voice that speaks the truth within the organization. This *single source of truth* is built on metadata, which is attached to data within the system, and it is charged with all aspects of managing the organizational datasets (Bhansali, 2014).

Master Data Management (MDM)

In a broad sense, master data management is in charge of making sure each business unit has a name and that every other entity in the business knows that name. MDM also provides the information needed to exchange data (without compromising its integrity) securely to the correct recipient regardless of which system that person, unit, or entity resides in. It also serves a categorization role corralling individual data into smaller tagged cohorts within the system (Bhansali, 2014). MDM also gathers and collates companies' various, disparate data. It integrates data that was previously siloed into a cohesive location under a standardized framework (Janoschek, 2018).

> **MDM Software Solutions.** When companies are storing large volumes of data with potential duplicates or inconsistencies, an MDM software solution is usually in order. The beauty of an MDM platform is that it functions like a well-intentioned helicopter parent. It is always hovering right there just out of sight waiting for the first opportunity to purge the database of duplicate data, correct any inconsistencies, and most importantly maintain the integrity of the data ("What is Master Data Management Software (MDM Software)? - Definition from Techopedia", 2020).

Conclusions and other methods

Stephanie Shen, a columnist for the online publication *towards data science,* believes there are seven steps or methods, that can allow organizations to achieve high data quality and maintain it: rigorous data profiling and control of incoming data, careful data pipeline design to avoid duplicate data, an accurate gathering of data requirements, enforcement of data integrity, integration of data lineage traceability into data pipeline, automated regression testing as part of change management, capable data quality control teams (Shen, 2019)[i]. Using MDM and MDR to standardize data are two conventional methods used to increase organizational data quality, and modern, forward-thinking companies would do well to consider their potential benefits. What experiences have you had with these technologies in your different industries? I would love to connect on this.

References

Bhansali, N. (2014). *Data Governance - Creating value from informational assets*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Ferguson, K. (2018). Why It's Important to Standardize Your Data - Atlan | Humans of Data. Retrieved 15 August 2020, from https://humansofdata.atlan.com/2018/12/data-standardization/

Janoschek, N. (2018). Data Quality & Master Data Management: How to Improve Data Quality. Retrieved 15 August 2020, from https://bi-survey.com/data-quality-master-data-management

Mayer-Schönberger, V., & Cukier, K. (2014). *Big Data - A revolution that will transform how we live, work, and think*. Boston: Mariner Books.

Sebastian-Coleman, L. (2013). *Measuring data quality for ongoing improvement*. Amsterdam: Elsevier.

Shen, S. (2019). 7 Steps to Ensure and Sustain Data Quality. Retrieved 15 August 2020, from https://towardsdatascience.com/7-steps-to-ensure-and-sustain-data-quality-3c0040591366

What is Master Data Management (MDM)? - Definition from Techopedia. (2019). Retrieved 15 August 2020, from https://www.techopedia.com/definition/840/master-data-management-mdm

What is Master Data Management Software (MDM Software)? - Definition from Techopedia. (2020). Retrieved 15 August 2020, from https://www.techopedia.com/definition/30185/master-data-management-software-mdm-software

---

[i] See paragraphs 4 through 20 for details.

Permalink Mark unread Show parent Reply Export to portfolio

Re: Striving for High Data Quality using MDM and MDR
by Bob Cheek - Saturday, 15 August 2020, 11:11 PM

Graham,

I am glad you talked about the metadata management as part of data quality. Last week, we discussed how important it is for data governance for organizations to focus on both the data and the metadata. One of the basics of data governance is having clearly defined and enforced standards. Zhang states, "Standards are necessary for interoperability, portability, and reusability, and are the most efficient way to facilitate the development of cost-effective,

interoperable systems." [1] Metadata is one of the key factors in establishing standards and having standards creates the uniformity one needs to develop automation to maintain data quality. Metadata not only helps in automating data quality processes but also defining the rules needed for reacting and transforming the data to meet quality standards. "The benefits of technical meta data use include source system identification, data quality measurement, improved management of ETL processes and database administration. Use of these technical data tags offers warehouse administrators and business users a means for measuring the content quality of the data in the warehouse." [2] To achieve high data quality it is more than managing the data on a daily basis, it is about managing the metadata as well. The definition, structures and

standards are as important as the data. As Anne Marie Smith, data management strategist, author and instructor stated, "Good metadata management can lead to good data quality since having and relying on the metadata can identify poor data / incorrect data / missing data. Also, having good metadata shows an understanding of data management and shows that the organization is committed to good data – hence an improvement in data quality almost always follows."[3]

References:

[1] Julia Zhang, "Chapter 4: Operationalizing, Data Quality through Data Governance." *Data Governance: Creating Value from Information Assets,* by Bhansali, Neera, CRC Press, 2014, pp. 75

[2] Marco, Author David. "Implementing Data Quality Through Metadata, Part 1." *TDAN.com*, 1 Nov. 2015, tdan.com/implementing-data-quality-through-metadata-part-1/5024.

[3] Jones, Dylan. "Data Quality Through a Metadata Strategy: Interview with Anne Marie Smith." *Data Quality Pro*, Data Quality Pro, 5 Oct. 2012, www.dataqualitypro.com/blog/data-quality-through-metadata-strategy-anne-marie-smith.

Permalink Mark read Show parent Reply

Re: Striving for High Data Quality using MDM and MDR
by Travis Dawry - Sunday, 16 August 2020, 7:08 PM

Graham, Bob, and class,

Even something as seemingly straightforward as labeling data may not scale up all that easily. Having a role dedicated solely to metadata management will create a niche expert, but it may also cause quite a bit of turnover! Institutional knowledge is not all the helpful if the SME keeps leaving the institution. The current pandemic has illustrated how vulnerable many jobs are to automation (Semuels & Alana, 2020), but much of this automation is surely illusory (Solon & Olivia, 2018). Labeling of non-critical, or even critical but non-sensitive (McNamara, 2020), data may be outsourced to humans external to the organization.

Does anyone have any other examples?

References

McNamara, A. (2020, August 13). Project Discovery: Could computer games help find a cure for COVID-19? Retrieved from Science Focus: https://www.sciencefocus.com/the-human-body/project-discovery-could-computer-games-help-find-a-cure-for-covid-19/

Semuels, & Alana. (2020, August 6). Millions of Americans Have Lost Jobs in the Pandemic—And Robots and AI Are Replacing Them Faster Than Ever. Retrieved from Time: https://time.com/5876604/machines-jobs-coronavirus/

Solon, & Olivia. (2018, July 6). The rise of 'pseudo-AI': how tech firms quietly use humans to do bots' work . Retrieved from The Guardian: https://www.theguardian.com/technology/2018/jul/06/artificial-intelligence-ai-humans-bots-tech-companies

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Frances Tang - Saturday, 15 August 2020, 8:13 PM

According to Loshin (2009), there are five fundamental practices that organizations can adopt for more effective data quality management: data quality assessment; data quality measurement and metrics; integrating data quality requirements into the business application infrastructure; operational data quality improvement; and data quality incident management.

First, data quality assessment involves processes for identifying current data-related issues, understanding their root causes, assessing the negative impact on business performance, prioritizing data quality management resources to address the issues based on business impact, and applying the assessment results to set baseline against which data quality improvement will be measured.

The second step is to determine the dimensions of measurement based on the assessment results from the previous step, and then define specific metrics. This process involves determining the tools, techniques, and skills needed to capture the data quality measurements (in terms of measurement dimensions); creating data quality metrics; defining the data validity rules and acceptance thresholds that can be incorporated into business applications and processes as controls to handle data errors and ensure conformance to business rules; and lastly, devising data quality scorecard to monitor and track data quality improvement progress (Loshin, 2009).

Third, to integrate data quality requirements into the business application infrastructure, organizations need to first conduct data quality requirement analysis to "determine the most appropriate points [where data sets are extracted, transformed, exchanged, or integrated] for inserting data inspection routines" and making data corrections (Loshin, 2009). Data inspection or data validation is a systemic process, enabled by technology, that "compares a body of data to the requirements in a set of documented acceptance criteria" (Bhansali, 2014, p. 86).

Fourth, the operational data quality improvement phase involves establishing a data quality service level agreement between data suppliers and data consumers. It is an agreement that "specifies data consumer expectations in terms of data validity rules and levels of acceptability [in terms of accuracy, completeness, consistency, timeliness, integrity, traceability, etc.], as well as reasonable expectations for response and remediation when data errors and corresponding process failures are discovered" (Loshin, 2009). Based on the agreement, data suppliers will proceed to develop enterprise-wide data standards and incorporate them into the information architecture to determine how data are stored, processed, and delivered. To facilitate the enforcement of data standards across the enterprise, metadata development and management is crucial. Metadata provides clear and standardized definition of data elements applied in the business applications and interfacing systems, thereby enhancing data sharing and system interoperability in a way that guarantees interactive parties of the organization can share the same understanding of the presented information (Bhansali, 2014, pp. 75-76).

The last part of Loshin's data quality management cycle is to carry out the data quality service agreement, identify and track data quality issues, conduct root cause analysis, and then develop and apply remediation plans. To achieve effective data quality management, remedying data errors should be proactive that starts at the root of a problem, instead of being reactive to address the downstream effects through iterative and time-consuming data cleansing. Process remediation is recommended because it "encompasses governed process for evaluating the information production flow, business process work flow, and the determination of how processes can be improved so as to reduce or eliminate the introduction of errors" (Loshin, 2009).

In addition to the methods recommended by Loshin, Julia Zhang also suggests cultivating an enterprise data quality culture through proper training, which can help encourage "a collaborative effort that arms business process experts with the right technical tools [and knowledge] to make cost-effective decisions about identifying, reacting to, and anticipating the types of data errors that lead to negative business impact" (Bhansali, 2014, pp. 84-85).

Methods for attaining high data quality are not limited to those mentioned above. Organizations also need to develop and implement different data quality metrics, data standards, and management requirements specific to their industries and based on their business priorities.

References

- Bhansali, N. (Ed.). (2014). *Data Governance: Creating Value from Information Assets*. CRC Press.
- Loshin, D. (2009). Data Quality & Data Integration: Five Fundamental Data Quality Practices. Pitney Bowers. Retrieved from https://moodle2.brandeis.edu/pluginfile.php/1625713/mod_resource/content/1/five_fundamental_data_quality_practices_Week9

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by William Hiraldo - Saturday, 15 August 2020, 9:42 PM

Discuss the methods that organizations can use to attain high data quality.

An organization's financial stability, and credibility are immensely dependent on the quality of data. A good analogy is the "quality of a product produced by a manufacturer, for which good product quality is not the business outcome but drives customer satisfaction and impacts the value and life cycle of the product itself." (Shen, 2019) The standard for good data quality can differ depending on the requirement and the nature of the data itself. Accuracy, relevancy, completeness, timeliness and consistency are five characteristics that embody high quality data. Good data quality consists of a well-organized data governance policy, a comprehensive management of incoming and outgoing data, standardization, thorough testing and design of the data infrastructure.

Bhansali points out three stages that all organizations should deploy with their quality strategy if they want to achieve high quality data. The first stage is a planning phase for enterprises data. In this phase, goals are defined, and a blueprint is designed to determine the overall end to end process of dataflow for the organization. "A good quality data program is an investment in the long-term success and profitability of your business." (Bhansali, 2014) The second phase encompasses the deployment of the processes developed during the planning phase. It is often referred to as quality control, and it is defined as a set of procedures intended to ensure that a manufactured product i.e. data adheres to a defined set of standards or requirements.

This stage specifically entails the standardization and technologies that are used within the enterprise to monitor, oversee, and manage their data. Bhansali states that the standards and technologies implemented should be tailored to your "specified business objectives and that focus will achieve a more sophisticated and effective solution." (Bhansali, 2014) That last step is about improving upon what was established and measuring performance.

Data is the foundation of the key performance indicators that drive business decisions taken by a firm's management. Designing these performance metrics, organizations must answer the where, what, and how for their metrics. The first step is to determine where you will be assessing the performance of the data. A thorough understanding of the operational processes that will be leveraging the metrics is necessary. For example, in the distribution one key performance indicator used is fill rate. Fill rate is the ability for a warehouse to satisfy an order in its entirety on the first attempt. It consists of a multitude of data elements such as sales order info, item information, location information etc. A firm grasp of these process is imperative to effectively manage the data. The next question to be answered is what (data dimensions) will be measured. Organization must determine what data will be used and what rules govern that specified data i.e. security. This happens frequently in the healthcare industry because many data elements are highly sensitive and are heavily regulated to ensure organizations manage and secure it well. The last question that should be answered is how to measure under the specified rules defined by the previous question. Bhansali sums it up nicely, data quality is the satisfaction of the requirements stated in specific business needs. (Bhansali, 2014) In distribution, this can be a dashboard that shows receiving productivity, picking productive, and vendor performance. All the data elements working in perfect harmony to further the success of the organization.

References

Bhansali, N. (2014). *Data Governance.* Boca Raton: Taylor & Francis Group.

Shen, S. (2019, July 28). *7 Steps to Ensure and Sustain Data Quality*. Retrieved from Towards Data Science: https://towardsdatascience.com/7-steps-to-ensure-and-sustain-data-quality-3c0040591366

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Travis Dawry - Sunday, 16 August 2020, 7:20 PM

Class,

William outlines some metrics important in logistics. Thinking about their characteristics, in the vein of Monday's prompt, do any of these stand out as particularly important to focus on? How about less important?

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Timothy Senstock - Tuesday, 18 August 2020, 12:15 AM

Hello William, Professor Dawry, and class.
Thinking about logistics where there are a significant amount of moving parts, the most important characteristic of measurement - in regard to the dimensions of data quality - would be timeliness. William mentioned how the fill rate is a significant measure of the organization's ability to fill an order entirely on the first try. Measuring the timeliness of the data is essential to understand whether the filling of an order is actually complete. While there could be a case to be made about the each dimension of data quality and their associated measurements, data timeliness and accuracy are going to play a significant role in the quality of data the organization relies upon.

Cheers,

Tim

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Greg Irwin - Monday, 17 August 2020, 11:50 PM

William, I like how you really laid out the three points for organizational deployment. For the first, planning, I really like how you mentioned that this part is the blueprint for the road ahead. To me, the planning phase is always the one i put at the top of the important list because a 'well' oiled plan makes the doing that much easier. If the design is well laid out, an enormous amount of energy is put in then the actions to be had are there for the taking. An organization that is in it to clearly define a

well structured plan for data quality will surely be happy to make the plan come to reality. Though the procedures may be daunting since they're the literal actions, procedures that are easily set out aren't difficult for a collaborative effort to tackle together. I would've like to see you put more mention into the last phase as the reevaluation should have nearly as much weight as the planning; checking over work is critical from turning in a homework assignment to a company going over its data quality procedure. While we hope to clearly lay out an issues in front of us, putting as much effort in the tail end will only highlight how diligent we are throughout.

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Greg Irwin - Saturday, 15 August 2020, 11:04 PM

When we look at the measurements or assessment of data its necessary to first construct the dimensions of what we're looking for. There should already be the obvious standards of what presentable data would like accessibility and legibility - you need to be able to understand what you're looking at to judge its value. After clearing these necessities we arrive at the dimensions that are critical - there needs to be some form of scope in our determinate of data quality. We need some form or a tangible setting/scale for our data or our judgement of it will never meet the expectations we're looking for. Our evaluation of data has now been set into a relative scope, from here we are looking for said data to be the following: comprehensible, reproducible and allow us to make comparisons between objects at different points in time.

Overall data quality involves a perception of assessment - a business would be holding a certain bar as the necessities that need to be met. Acceptable quality is necessary for an organization to show its effective work to all involved. Implementation of methods for quality will naturally make people feel like there will be an extra eye always watching and would hopefully encourage the standard to be raised. In addition to the watchful eye understood, companies should implement the necessary steps to implement a method for quality to meet its full potential.

1) Measurements must be Comprehensible and Interpretable:
I mentioned the obvious standard - you have to be able to comprehend what you're reading in order to properly assess its legitimacy. Like any high school english teacher, if you can't understand what a student may be writing then there's no way to properly asses the value of the work. A company must articulate the necessary measures to have a medium in understanding - having some sort of medium of understanding allows people to effectively understand the language of the data.
2) Measurements must be Reproducible:
There needs to be a level of trust in the consistency of measurements - without this uniformity of trust, the numbers and literal data can't be understood by all. Similar to the SI units of measurement, individuals need a standard bar of consistency in order to replicate measurements all over the world. Comparisons are the necessary tool in evaluation and being able to reproduce results is the next important pillar in that process.
3) Measurements must be Purposeful:
You need to have some sort of basis as to the reason for your assessments. In addition, you need some sort of counter in order to compare objects - you have a definite starting point. There is an understood direction and purpose to the process you're after - the purpose is making it worthwhile and pushing the direction.
4) Data Quality Assessment:
The end of the circle where we see if the steps taken have provided the necessary outcomes - this part is the necessary "drag" we see at the completion of any work. Checking our work is the necessary evil and the assessment is the evaluation of all the work we've put in. If we've been diligent throughout the process then the assessment will be quick and easy; the work will be much heavier if there are a lot of errors however it will show that the assessment is just as necessary in the end.

Reference:
Sebastian-Coleman, L. (2013). Measuring Data Quality for Ongoing Improvement, Morgan Kaufman

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Travis Dawry - Sunday, 16 August 2020, 11:15 PM

Greg and class,

This provides a good transition to Monday's prompt. What standards that seem obvious might be overlooked with regard to assessing data quality?

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Timothy Senstock - Sunday, 16 August 2020, 1:54 AM

Discuss the methods that organizations can use to attain high data quality.

When it comes to identifying the methods used to establish high-quality data, one must define *high-quality data*. Sebastian-Coleman discusses how, ultimately, high-quality data is that which meets the expectations of the organization (2013). In the most basic sense, data must serve the needs of the organization in order to constitute *high-quality data*, but the expectations of the data must be defined. The reason that an organization must possess defined expectations of their data is because this will allow the organization to properly measure the data quality, which is a key to helping the organization understand the extent to which their data quality needs improvement.

Sebastian-Coleman describes the ways or *dimensions* in which data can be measured – these dimensions include: "accuracy and validity, completeness, consistency, and currency or timeliness among them" (2013). For the purposes of having a reliable data quality, the measurements of the data quality must be comprehensible and reproducible, meaning that there must be an understanding of what is being measured and why as well as consistent or reproducible in regard future measurements (Sebastian-Coleman, 2013). One method an organization will need to perform would be a data quality assessment, which is a determination as to whether the expectations of the data matches the business needs. *Data Profiling* is a method to identify the features of the datasets (data types, field lengths, etc.) and to quantify the distributions of those values within databases before the data is stored (Sebastian-Coleman, 2013). *Data Issue Management* can help to identify problems in the datasets and serve to report, track and resolve these issues (Sebastian-Coleman, 2013). *Data Quality Thresholds* can be used to identify variations or trends in the data, which are outside of the 2-3 standard deviations from the mean. These deviations could indicate a data quality issue that can be resolved before the data might be used or analytic purposes. Sebastian-Coleman explains how these *controls* act as a "form of feedback built into the system to keep it stable" (2013). If there are any significant changes in the data caused by a data quality issue, it will be the controls that will help to identify these issues (Sebastian-Coleman, 2013).

Resources:

1. Sebastian-Coleman, Laura. (2013). Measuring Data Quality for Ongoing Improvement: *A Data Quality Assessment Framework*. Morgan Kaufman: Waltham, MA.

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Bob Cheek - Sunday, 16 August 2020, 9:12 PM

Hi Tim,

You started your post by talking about what high-quality data is. It being a match of expectations to use makes it a somewhat subjective measurement. By which, each organization must find its own path through the data quality journey. There are is plenty of material and guidelines but each organization must define its own roadmap. Next step was the discuss of data assessment. Once one can define where one wants to be the next obvious step would to determine where one is. Now the organization has something to measure against to verify the quality of their data. With measuring and profiling of data, it is difficult to balance the amount of effort in trying to measure the data and amount of effort to apply in cleansing. "So, when measuring data accuracy, you should find a good balance between the required efforts and the value expected from the action." [1].

One common pitfall is that an organization makes with the assessment and defining high data quality is to believe they are done. They believe that if they are measuring and tracking their progress to data quality they are achieving high data quality. However, what is important to remember is that there should be periodic re-assessment. Since it is about meeting expectations and for purpose of use, over time expectations changes and needs change. If the use of the data has changed one may not have high quality for the new use. "The status of people and context can and do shift over time, thus the quality of something can change even if the thing itself does not change." [2]

References:

[1] Bekker, Alex. "Data Quality Assessment Is Not All Roses. What Challenges Should You Be Aware Of?" *KDnuggets*, Sept. 2019, www.kdnuggets.com/2019/09/data-quality-assessment-challenges.html.

[2] Bach, James. "Assess Quality, Don't Measure It." *Satisfice, Inc.*, 29 Feb. 2020, www.satisfice.com/blog/archives/487091

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Timothy Senstock - Tuesday, 18 August 2020, 12:43 AM

Hello Bob,

I definitely agree with you in regard to the re-assessment aspect of data quality. While I have not personally been a part of an organization that has had to re-assess data quality as it related to changing needs and requirements, I can certainly understand how it could change. I think about baseball (mostly because I am obsessed with Sabermetrics) and how organizations had to quickly re-assess and get as creative as possible in their data acquisition and quality in order to embrace the changing big-data age in baseball analytics. Because so much data can be derived from a single pitch, hit, etc., MLB organizations have to have a significant stake in data accuracy and completeness in order to run effective analytics when it comes to arranging teams based on win equity or trading or signing players on expensive contracts.

Tim

Permalink Mark read Show parent Reply

Re: W5Q1 Due Saturday, midnight EST
by Greg Irwin - Sunday, 16 August 2020, 11:55 PM

Hi there Tim,
I enjoyed reading your piece particularly when you stated how some of the challenges in attaining high data quality involves first defining high-quality data. Everything involving our overall evaluation needs to be set in standards that make sense. Creating these standards are necessary because we need to remain realistic to meet expectations. In this direct sense, like you mention, data can then serve the needs of the organization - the more definitive we get in the definition of data then the easier it is to understand if we've met the standard for it.
To go in line with the methods of meeting high data quality, I think its hugely important to emphasis the comprehensible aspect of the evaluation. For me its the most part, and the first door you open in the process. If you're unable to understand what is in front of you/the data from somewhere else then you're at an immediate dead end for evaluating this quality.

Permalink Mark read Show parent Reply

◄ W5Q2 Due Monday, midnight EST

◄ Five Fundamental Data Quality Practices

Jump to...

Week 5 Participation Feedback ►