HITACHI
Inspire the Next

WHITE PAPER

Why Data Quality Is Essential for Content Analytics

5 Key Insights for Establishing Proper Data Hygiene

By Hitachi Data Systems

November 2016

Contents

Executive Summary 2

Introduction 3

Improve Your Data Hygiene 3

5 Key Insights for Establishing Proper Data Hygiene 4

Determine the Referential Value of Your Data With Profiling **4**

Establish Clear Data Ownership and Reject Myopic Thinking **4**

Remember: Data Hygiene Is Not a Step, It's a Journey. Don't Get Complacent! **4**

Avoid the "P-Hacking" Tendency **4**

Lead With the Business and Mature Data-Ready Managers **5**

The Data Quality Management System: 4 Requirements 5

Aggregates All Organizational Data **5**

Profiles Data in Place **6**

Ensures Data Quality **6**

Supports Data Security **7**

Hitachi Content Intelligence for Data Quality Management 7

Conclusion 8

About Hitachi 8

Executive Summary

What can your organization fund with an extra US$14.2 million dollars? This is annual cost that poor data quality has on an organization according to Gartner (ID: G00297356). The quality of data is fundamental to decision support and has, for a long time, been the reason why big data analytics and business intelligence initiatives exist. But if the quality of data determines the ultimate success or failure of an initiative, how do you ensure that you are getting the most accurate data to the right people when they need it most? Making better decisions requires the right information. Ensuring that you have the right information available, with verifiable quality, means taking the steps necessary to improve the processes and technologies used to store, manage, govern and mobilize data. It means aligning the business and IT stakeholders on what it means for data to be accurate and agree on a data quality management system to standardize and automate the process.

Introduction

In this age of big data, there seems to always be more data available than what an organization or individual can accumulate, analyze and use. Moreover, the quantity of data far outpaces the quality of data, and good data is a critical part of building transformative business strategies or making informed decisions. High-quality data can generate actionable insights, which are what organizations rely on to improve operational effectiveness, enhance the customer experience, and identify new business opportunities. But when the quality of data cannot be guaranteed, those business strategies and decisions are placed in jeopardy.

These difficulties should not be surprising, given the velocity, volume and variety of data available to us today. Naturally, the course of action would be to simply make the data better. Unfortunately, the main reason data is often inaccurate stems for poor data management practices, which stem from responsibility. Businesses often get in their own way by refusing to create a culture around centralizing data ownership and responsibility.

In many organizations today, data ownership, management and control is fragmented with priorities driven from competing stakeholders. Without a centralized and standardized approach to data management and governance, an organization will ultimately expend more capital and operational resources just to store the data, let alone prepare it for analysis. This is what creates inconsistences in the data and results in reactive processes to correct them, or proliferates data silos and point-based solutions that exacerbate the problem.

There must be a better way, and the good news is that traditional approaches are transitioning to new ways of thinking. A transformative shift in the market, industry regulations or an active part on the organization to change their data management strategies may lead to this change. The new thinking breaks down data silos, cleans and normalizes data, and considers business intelligence first rather than as an afterthought, as had previously been the case.

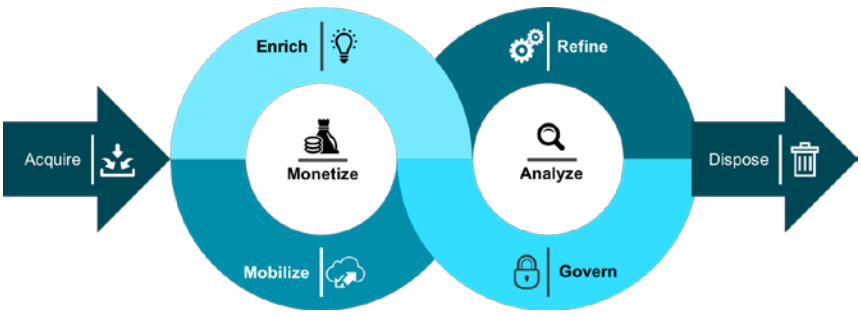
Many organizations know they need to improve the quality of their data and likely understand the conceptual benefits of utilizing their data. These same organizations often have a hard time with where to start: How do they define what quality data means and how do they measure it? How do they make the time and investment necessary to the current IT infrastructure? How do they implement a change while the business continues to operate and generate even more data? Regardless of these challenges, one fact remains consistent:

Accurate analytics drives actionable insights based on "good" business data that is based on the proper data hygiene.

Improve Your Data Hygiene

It is important to start this endeavor understanding that data ***must*** be viewed just like any other business asset. It means being proactive and diligent about what how data is developed, stored, managed, governed, accessed, analyzed and reported on. It requires defined processes that provide the guidelines on all the phases of your data's very dynamic life cycle, not just its creation to disposal (see Figure 1).

Figure 1. Phases of the Data Life Cycle



Getting started requires an understanding of what kind of data you have and where it's located. It is not only imperative to understand your data, but also to actively define how you are going to record, monitor and measure the quality of your data. Performing this step first is crucial to ensuring any future steps you take prevent proper data hygiene and management form being siloed and easily forgotten or discarded.

5 Key Insights for Establishing Proper Data Hygiene

Determine the Referential Value of Your Data With Profiling

Data quality means having the right data at the time you need it most in a format that you can trust. Data profiling is a key process in determining value. Data profiling is the process of gaining an understanding of the existing data relative to the quality specifications. This process consists initially of looking at the actual data and asking whether the data is complete and accurate. Profiling should be aligned to a business objective to ensure that you create the proper starting point for how data quality is measured.

To get started, audit your existing data sets with a business problem or objective in mind that your data is meant to solve. Enumerating your data with this in mind establishes the criteria for defining how the data will be used and the goals of its usage. This approach enables you to define a relevance or referential value for the data, based on its fitness for its purpose. Finally, the profiling performed can be used to define how future data is processed, to ensure the same level of scrutiny is applied to data quality.

Establish Clear Data Ownership and Reject Myopic Thinking

Data quality can vary significantly depending on its intended use, how it was collected, where it will be stored and the processes used to enrich and cleanse it. It is important to understand that not all data is created equally. Some data is for immediate use and quickly discarded, other data may be critical to the operations of the business, and a large portion may be legacy data that must be retained for a set period of time. Given that not all data is created equally, ensuring its quality does not mean that a single data management approach applies to all data.

It's extremely important to establish clear ownership and accountability over your data as soon as possible as lack of ownership equates to a lack of control. Use the metadata of the data repository to profile how and where organizational data is stored, and who is responsible for the repository. Audit its contents to validate that the data has the identified and required qualities defined in the first insight. Data ownership mapping is equally important to mapping where the data is located in the organization. The benefit is that any future data stored in these repositories will inherit the management and governance policies set forth, and data owners are operating under a consistent set of policies and expectations.

Remember: Data Hygiene Is Not a Step, It's a Journey. Don't Get Complacent!

Organizations get overwhelmed quickly when it comes to managing data quality. This is because data is not always seen for what it is: a tool for changing the way business is done. Therefore, data quality needs to be maintained.

There are a lot of things that can impact data quality: user-introduced errors, inaccuracy in how it is collected and processed, moving the data from one location to another, and so forth. In other words, the accuracy of data can diminish over time. This makes the improvement of data quality an ongoing pursuit, not a project. A project-based approach to data quality is singular and often introduces more quality issues than it solves.

Consider enacting a data custodian role and making that role responsible for documenting and preserving the findings regarding data sources and processes. This step continues to refine your data quality approach and provides a way to oversee and communicate the data quality findings to appropriate stakeholders.

Avoid the "P-Hacking" Tendency

P-hacking is the concept of using inaccurate data or purposely manipulating data to achieve the desired results. Simply grabbing any kind of data, analyzing it and arriving at a conclusion without first understanding the reasons why data quality can vary, can result in bad consequences. Pausing to take time to understand the data, where it

originated, how it is related to other data sources, and its relevance to the business, results in greater accuracy in the decision made or action taken.

Continually evaluate data relevance as much as data quality. The best quality data accurately describes something in the real world. Ensure that data is normalized to arrive at a common format, enriched to keep it fresh, and validated to ensure accuracy. Mastering your data, metadata and reference data management are the most important steps when considering the quality of your data.

Lead With the Business and Mature Data-Ready Managers

IT-centric data quality practices will not work effectively since they are often misaligned with business requirements. Similarly, business professionals will not always understand the importance of data quality if they are not equally responsible for it. The most common mistake made is leading with a data quality initiative rather than leading with a business initiative. Data quality isn't a business outcome: It's about the business objective that data quality enables. This is why all leaders need a working knowledge of data science, and why data quality is a business problem as much as it is an IT problem.

Data quality can remain a responsibility of IT, but only if the IT infrastructure has been modernized and the business shares in the responsibilities of defining data quality, value, usage and life cycle. As additional data sources are adopted, the ability to navigate among the data sets using a flexible IT infrastructure that is guided by defined quality standards enables fluidity versus conformity. It makes the business nimble, ready to act with decisions based on accurate data.

While these data quality insights can seem daunting, they are practical steps that businesses of all sizes can use to establish a flexible data quality management system. Data is an organization's most strategic asset and, as such, must be invested in like any other critical system, to leverage and improve organizational performance. It is imperative that organizations have a proper data quality management system to keep data fresh and guarantee proper data hygiene. A business can enable the rapid surfacing of meaningful insights by ensuring its system can review and evaluate data quality, integrity and completeness, so that when required, it can create business rules or gap strategies to fix your concerns.

The Data Quality Management System: 4 Requirements

A data quality management system entails the establishment and deployment of roles, responsibilities, policies, procedures and technologies concerning the acquisition, maintenance, dissemination and disposition of data. A partnership between the business and technology groups is essential for any data quality management effort to succeed. The business areas are responsible for establishing the business rules that govern the data and are ultimately responsible for verifying the data quality. The IT organization is responsible for establishing and managing the overall environment. It must handle the architecture, technical facilities, systems and software that acquire, maintain, disseminate and dispose of the electronic data assets of the organization.

Choosing the right data quality management system can be challenging, but it is not impossible. Any solution being considered must not only be flexible and scalable, but it must also consistently learn and evolve with the new data types and sources impacting your organization. The following four requirements must be met for any solution being considered to improve workforce productivity and data quality without compromising the objectives of the business:

Aggregates All Organizational Data

Data often exists in multiple locations within your organization at the same time. These locations are often physically or logically disconnected, resulting in data sprawl that negatively affects both your resource utilization and the overall quality of your data. Successful initiatives hinge on having a complete view of organizational data; to obtain that view, ensuring data quality is key. Now, more than ever, organizations are placing ever-increasing importance and value on the data they hold, and the shortcomings of fragmented and siloed repositories are becoming more apparent.

The data quality management system must support the aggregation of any data type from any data source. This is important whether you intend to centralize the data on a standard architecture or simply want know what you have and where you have it. Without the features to support connections to any data source with any data type, you'll never have a complete view of your organizational data. More importantly, any efforts to profile it, measure its quality, enrich and augment it, or apply the necessary governance and security controls will be ineffective and lack completeness.

Profiles Data in Place

Since data about the same item can exist in multiple locations, it can take virtually any form. This results in difficulties with respect to identifying and connecting it to the intended use, that is, unless you have a means to profile the data to evaluate its completeness and accuracy.

Data completeness concerns the analysis of its contents to ensure it contains everything necessary to support the intended analysis, as well as the appropriate metadata that describes the data in more detail. For example, making sure the data is the most recent version and contains details of a product being reviewed is a prerequisite for the data to be considered complete. Data accuracy focuses more on the values the data contains. In other words, just because the data contains a version number does not mean it is correct or the most recent version. With data profiling, we can look at the data as a whole and discover a more about the data than what is described by its file name.

The data quality management system must provide the appropriate processing steps to analyze data, regardless of format, to properly determine the data type, extract all or part of its contents, and collect all of the metadata (standard and custom) used to describe the data. This approach allows you to be selective in the data that you process, and ensure the data meets the agreed upon standards for inclusion or exclusion in any analysis.

Ensures Data Quality

Once data completeness and accuracy are validated, ensuring the quality of the data allows an organization to transform it into valuable and relevant business information. Data can be transformed with cleansing, normalization, and augmentation techniques, all of which are not meant to change the data's original context, but rather to make it more useful to its intended audience.

Data cleansing or scrubbing is the process of detecting and correcting inaccuracies and either excluding them or correcting them from the final data set. For example, if the business rule is that all references to states should use the official two-character designator, data cleansing can use pattern matching to find the inaccurate uses and replace them with the correct data format.

Data normalization is a way to organize and standardize similar data constructs to reduce redundancy and improve the integrity of the data. For example, if you have files from different geographies they likely contain dates that are formatted specific to the region they originated from. Normalization allows you to set one date format and ensure it is applied to all processed files.

Data augmentation allows the system to further help the overall data quality. With data augmentation, new key-value identifiers, metadata, relationships and so forth, can be applied to the data. This process is a means of further enhancing the data's relevancy to the business objective and the user of the data.

The data quality management system must provide the capability to interpret the business objective as a set of rules for which profiled data can be excluded, accepted, corrected, enriched and transformed on a data-by-data basis. Interpretation is needed because the specific approach taken may differ for each data object, and the decision on the approach needs to be made by the business area that is responsible for it. Encapsulating the business rules enables the system to automate the data quality assessment process and reduce the available data to only the usable or relevant data.

Supports Data Security

It is important to deliver the most accurate and highest quality of data available to the right person at a time when they need it. This practice both increases its value to the business and reduces the risks the business might face, depending on how sensitive the data is.

The data quality management system must support integrations into the organization's security services to validate a user's role with their request to access the data in question. Also, the system should support those unique situations where user access to data must be more granularly controlled. For example, users in one group have rights to see all of the data, while users in another group can only see a redacted subset of the data.

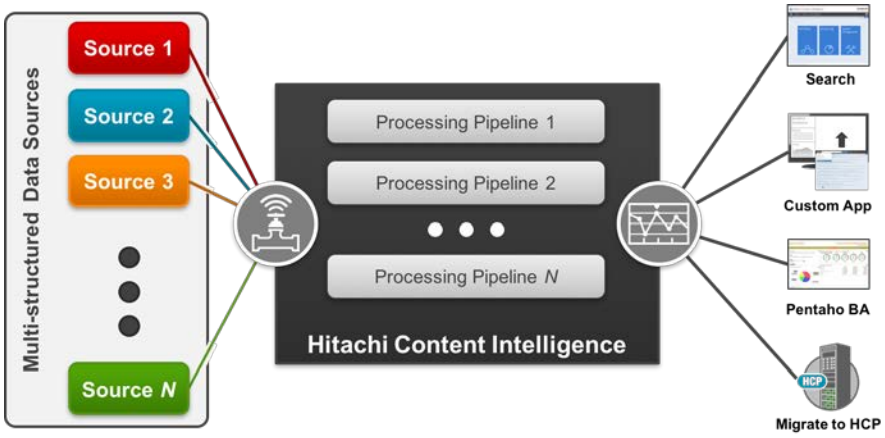
Remember, quality data in the wrong hands can be as detrimental to the business as making decisions with inaccurate data.

Hitachi Content Intelligence for Data Quality Management

Organizations must answer the challenges of exploring and discovering relevant, valuable, accurate and factual information across a growing number of data producers and siloed repositories. To do so, they need a less time-consuming solution that surfaces rapid insights from data that is big, deep and dark by enabling the identification of facts, trends, events and other relationships.

Hitachi Content Intelligence enables organizations to connect to and aggregate all of their organizational data, regardless of where it resides and the format it is in. With more than 20 different data processing capabilities, ensuring the completeness and accuracy of the data is not a balancing act. Data can be profiled, enriched, cleansed, augmented, transformed, excluded, included and more to ensure it is of the highest accuracy and the most relevant for its intended use. As shown in Figure 2, Content Intelligence is capable of preparing the data for multiple uses. For example, it can migrate the profiled data from one location to another, supplement business intelligence tools, or provide self-service guided data exploration services to business stakeholders.

Figure 2. Hitachi Content Intelligence: Multiple Data Uses



HCP = Hitachi Content Platform, BA = Business Analytics

As a part of Hitachi Content Platform portfolio, Content Intelligence can enable businesses to centralize all of their data into an enterprise data fabric that is both flexible and scalable to meet their retention requirements. From a single place, all organizational data can be effectively managed, governed, mobilized and analyzed to address improve operational effectiveness, enhance the customer experience, or search for new business opportunities.

Conclusion

While these data quality insights can seem daunting, they are practical steps. Businesses of all sizes can use them to establish a flexible data management program, to invest in data as a strategic asset and to leverage it to improve their organizational performance. It is imperative that organizations have a proper system to ensure data quality by keeping it fresh and guaranteeing proper data hygiene. The proper system also enables the rapid surfacing of meaningful insights: It reviews and evaluates data quality, integrity and completeness, so that when required, it can create business rules or gap strategies to fix your concerns.

Ultimately, data quality is a balance. Using two of the more common metrics of data quality, completeness and accuracy, we see that to have data that is both 100% accurate and 100% complete can be very expensive, and is not always achievable. We often find ourselves sacrificing one or the other. For example, when there is a data error, we need to decide whether completeness is more important, in which case we may include the data with an error, or whether accuracy is more important, in which case we may omit the data. But a balance can be achieved with the right data quality management solution: Hitachi Content Intelligence.

About Hitachi

Every business must improve cost-efficiency, time to market, customer experience, and revenue. Digital transformation promises these gains through better management of your business's data, the common ground among business and IT leaders. No one knows data like Hitachi Data Systems. We help the world's largest organizations with one thing – data. Our integrated strategy and portfolio enables digital transformation through data, helping enterprises to manage, govern, mobilize and analyze data to uncover insights for better outcomes. Use what you have today and define your data-centric roadmap towards digital transformation. Hitachi Data Systems is your partner for digital transformation today and tomorrow.

Hitachi Data Systems

Corporate Headquarters
2845 Lafayette Street
Santa Clara, CA 95050-2639 USA
www.HDS.com | community.HDS.com

Regional Contact Information
Americas: +1 866 374 5822 or info@hds.com
Europe, Middle East and Africa: +44 (0) 1753 618000 or info.emea@hds.com
Asia Pacific: +852 3189 7900 or hds.marketing.apac@hds.com

WP-556-A S. Baker November 2016.