COMMSCOPE®

The dirty data problem and data center infrastructure management

Volney Douglas, Ph.D., Principal Systems Engineer | iTRACS

Contents

Intended audience

This document addresses factors and concerns related to the use of data in an organization, such as those that support data centers and are interested in improving their overall data quality. This document is written for and intended for use by technical engineers and managers with some background in the data sciences and familiarity with data centers and related engineering principles.

For more information on iTRACS, please visit https://www.itracs.com/.

Overview

Big Data, automation, and the application of deep learning are rapidly becoming critical components of organizational success. Still, with the increase in data quantity, the problem of data quality (i.e., the so-called "dirty data" problem) has grown in complexity and urgency. Dirty data is incorrect or misleading qualitative or quantitative values or variables (e.g., facts, statistics, or measurements) with some quality issues. For most organizations, the volume, variety, and velocity of data-gathering processes are increasing. Still, unfortunately, the size of Big Data rapidly outstrips the rate at which it can be processed and stored without quality issues.

Diving deep into this data is critical to future organizational success, but incomplete or "dirty data" can hinder, misdirect, slow, or even mislead these efforts. Dirty data is the Achilles heel of modern Big Data, automation, and machine learning because it can hinder understanding or even provide the wrong insight. Organizations benefit most by having processes that help ask the right questions at the correct times. These systems can recognize patterns, reduce known and unknown risks, predict events, automate, inform the best decisions, and optimize the organization for innovation. The value of Big Data is directly related to the quality of that data. Therefore, understanding, managing, and improving data quality using systematic processes, frameworks and systems are critical to success.

Many techniques exist to process and improve dirty data. This paper discusses high-level strategies and offers techniques and tools for deep diving into data using context and frameworks. In specific areas such as data center management, software solutions such as a digital twin can provide organizations additional support when they manage and process dirty data. Often organizations view these tools as instant solutions to dirty data problems instead of frameworks that help guide the organization toward quality improvements. If applied using improved methods and analysis—and with the application of specific organizational processes—a digital twin can help resolve these dirty data issues. One of the critical paradoxes of dirty data is that organizations tend to want to hide the problem because reviewing and fixing this data is viewed as an unknown-unknown issue that might highlight other organizational, process, or data-gathering problems. Dirty data wants to stay dirty, and this aspect of dirty data can cause a negative data quality feedback loop if the data process is unmanaged. In addition, because organizations tend not to want to discuss issues concerning dirty data, they tend not to understand the scope of the problem. They might even falsely believe they do not have a dirty data issue.

Summary

Reviewing, analyzing, comparing, and evaluating data as part of an organizational system using tools specifically designed for those purposes is the first step toward improving data quality. Improving data quality (e.g., fixing the dirty data problem) should be a fundamental part of any organization's continuous improvement process. Using purpose-built tools such as a digital twin can improve data quality on specific aspects of organizational data and is critical to leveraging future technology in areas such as machine learning (ML), Big Data, automation, and process improvements. The success of these larger Big Data projects relies solely on whether an organization can overcome its dirty data problems.

What is dirty data?

Dirty data comes in many forms, but, at the core, it is incomplete, inconsistent, or inaccurate information. Dirty data is more than just poor-quality data; it is an indicator of organizational data health. Unfortunately, most dirty data results from organizations not treating their data as a strategic asset. Dirty data is not just bad data but can lead to the pollution of other data and information. Dirty data can act like a virus or corrosion that can corrupt other data sources and cause issues with continuous improvement, decision-making, automation, strategic thinking, and organizational processes. Integrating multiple sources of real-world data sources may lead to significant errors. Dirty data can also be data silos (e.g., millions of pieces of data stored in thousands of different spreadsheets or even in notes written down but not shared). Each organization should view dirty data as a strategic priority since the causes and solutions tend to be organization specific. The organization should manage the dirty data issue with processes developed, tailored, and implemented strategically.

Most experts agree that data cleaning remains one of the most time-consuming steps in data analysis. Fixing and collecting data often require manually inspecting each dataset, which can quickly become costly and time-consuming. Currently, there is no simple way to fix some errors without targeted processes and a toolset that fits into the organization's strategic planning process. The causes and solutions to an organization's dirty data problem must be defined and managed strategically. For example, what is an acceptable level of dirty data? What existing tools and systems exist to improve the data quality? What is the value of Big Data, machine learning, and automation to the organization? Since the scope of dirty data covers many tools, organizations, processes, and strategic planning, this white paper discusses the impact of dirty data on data centers and tools such as digital twin used to resolve those issues.

What are Big Data and deep learning?

Big Data is the systematic extraction of information from significant data sources to develop predictive analytics and business informatics. The Big Data sources have increased in size and complexity each year as data storage and processing costs have decreased. The size and complexity of these data sources have required novel methods, such as deep learning to process and understand the information. Unfortunately, Big Data and deep learning need good data to be effective and valuable. For example, an organization could collect data from every device on the network every minute and collect massive amounts of data. But what does that data mean, and how can it be used? Or an organization could manage power used on all devices in all data centers, but what does that data tell us about device efficiency? Big data and deep learning can provide some insight if used correctly, but even in training ML models such as R and Python, the first step is to download training datasets that have been cleaned and carefully organized. Users who have then attempted the same processes on their data usually get confusing results as the first step. In theory, these tools work fine. But why is this not working on my data? The answers are varied, but often the issue is that you must first validate your dataset to use those methods. Applying Big Data processing and machine learning to dirty data can provide unfortunate results and incorrect, confusing, or contradicting answers. Often the promise of Big Data and machine learning only comes to an organization after it has solved its dirty data issues.

The big V's of data

The three big V's of data are volume, variety, and velocity. As the field has evolved, it has expanded to include three more big V's: veracity, value, and variability. Big Data starts with a large volume of data, but it also should contain various data sources and types and a large velocity in the overall data flow.

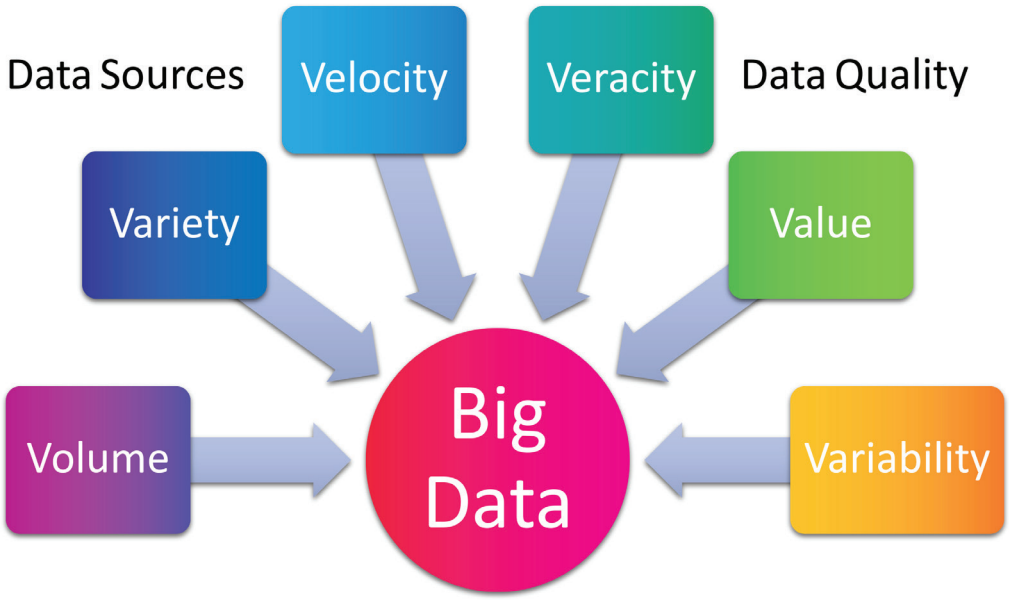


Figure 1. Big V's of data quality

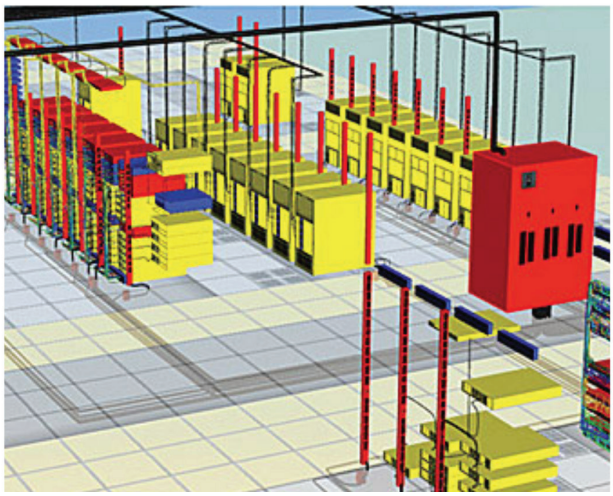
The second set of Big Data elements deals with data quality, where dirty data has the most impact:

- Veracity is the ability to identify the relevance and accuracy of the data and apply it to the organization's purpose.
- Value understands how that data can provide revenue or unlock opportunities for the organization.
- Variability provides context and structure to the data for predictions and analytical reporting.

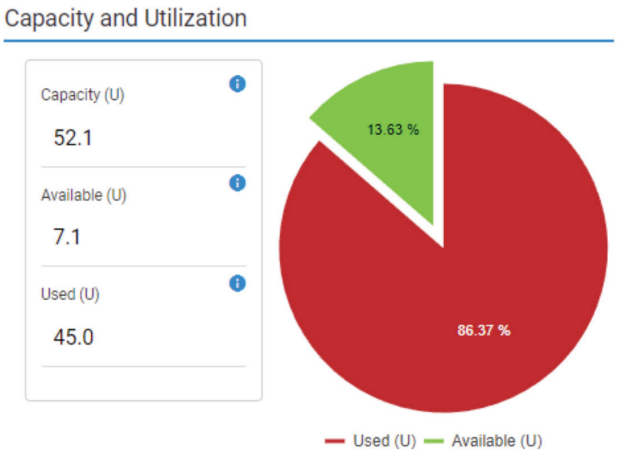
Most organizations make the mistake of focusing on Big Data sources and not on Big Data quality because quality usually requires the application of systems and processes. Big Data quality also requires strategic management to balance the cost of tools and techniques against the potential benefits and gains of higher-quality data. In our case, we will use specific examples and tools framed in the context of an organization that manages data center(s). In the context of data centers, one tool that can help Big Data, machine learning, and automation improve data quality and reduce dirty data is digital twin software.

What is a digital twin?

At a strategic level, managing data centers is a complex activity. Data centers tend to be a core feature of organizations that deal with Big Data, and organizations that deal with Big Data tend to develop data centers to manage that data. Data centers and management go hand in hand and should be evaluated strategically. What type of resources does an organization need to collect data? Should that organization build and operate its data centers or outsource these functions to the cloud? Are the resources being used effectively, or are there better methods? The development of software tools such as a digital twin has focused on centralizing, monitoring, managing, and improving data center critical systems such as power and



space planning. To simplify how dirty data impacts these types of data center systems, we will focus on two main areas: space and power. These two areas are typically the starting point for any organization attempting to improve data quality for their data centers because they seem less complex.



A data center has the equipment, and some or all that equipment tends to be powered, so, in theory, that should be a good starting point for data management. Space management encompasses the location and placement of assets, while power management deals with the consumption and control of electrical power. Space management can be as simple as understanding the place and order of equipment and can also include additional aspects like fault tolerance, risk management, future planning, and strategic objectives. Power is one of the most expensive parts of a data center and, therefore, a good target for cost reductions or improvements. Organizational leadership should evaluate data center space and strategic power management objectives at the strategic level. For example, cutting power costs by 10 percent might not be realistic if a goal is to grow by 10 percent without new investment in technology or removing existing equipment. Digital twins can help provide a framework to support strategic planning by providing a realistic dataset on which management can evaluate past, current, and future conditions.

Space management

Space management is typically the first concern for most organizations starting a digital twin project. Often organizations will start with IT equipment because it is a critical resource in the data center. Examples of spatial data are a data center's position, status, and equipment assignment. Space typically starts with collecting the current state of the asset and maturing into data center asset allocation and future development. Since the current state of asset management of a data center is the backbone of data quality, it is often the first place dirty data shows up. Without high data quality, using advanced space management features in a digital twin has uncertain value. For example, engineers cannot effectively plan for future assignments if current assignments are unknown. Current space data is critical to planning future space assignments. Improvement of space data is a vital starting point for a digital twin installation. Space data is a good indicator of data quality. If the space data is poor, it is a good sign that other areas of the data quality are also insufficient. For example, if the location of a server is unknown or incorrect, often the associated power data might also have issues.

Power management

Power management is the second aspect of a digital twin that attracts most organizations attempting to manage data centers. Power is often also combined with cooling or thermal management since data center equipment produces a lot of heat, which requires energy to remove from the facilities. Advanced computational fluid dynamics (CFD) techniques, thermal management, and cooling management are all tools that organizations can use either as a part of digital twin or with integration with other similar systems. The value of these advanced management tools relies on the quality of the power data, e.g., if the power consumption of all the servers is unknown or incorrect, the cooling load of these servers will also be wrong. The value of the analysis and improvement of power management relies heavily on good data quality. Poor data quality can impact the value of any process evaluation, especially in the areas of power management in digital twin.

Strong foundations

For example, power usage effectiveness (PUE) is a ratio that compares the data center's practical energy use to the use of energy on supporting systems. PUE is one of the most popular metrics for calculating a data center's energy efficiency. PUE is the total facility energy divided by the IT equipment energy. The total energy usage of a facility is easy to gather since it is often the direct cost associated with the data center in power use. The other IT equipment energy is more difficult to pick since the data center must know either the energy use of just the devices specifically dedicated to the functional IT purpose of the facility or the energy use of all the support equipment. Many companies used to market their lower numbers. Power engineers often assumed anything below 1.2 was impossible, but many companies advertise numbers as low as 1.09. PUE has always been problematic in how data is collected and how accurate that data is in aggregate. A PUE calculation will always be questionable without engineering data quality analysis. For example, failure to include power consumers like lighting, and having incorrect power consumption values on devices, can lead to errors in the rollup values calculations for PUE. Management must resolve the dirty data problem for rollup values like PUE to have any organizational value. Often rollup values such as PUE are a double-edged sword that can highlight administrative data issues and mask the real problems. For example, why do we care about accurate power information if the PUE number is excellent?

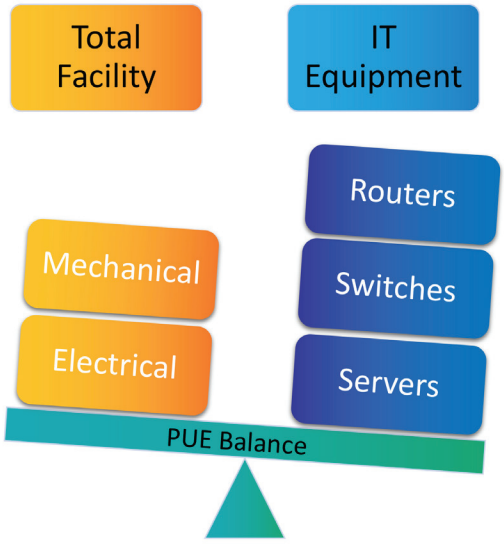


Figure 2. PUE balances on good data

Understanding dirty data

How good is your organization at spotting dirty data? If an organization does not have a strategic definition, process, and method for measuring data quality, it might not be good at spotting dirty data. How can you measure something without some form of measurement? For example, how good would an organization's finances be if there were no financial department, no financial planning, and no strategic method for evaluating the financial health of an organization? Once again, that might not be a good sign. Still, it doesn't mean the organization was financially unhealthy or didn't manage financial data correctly, only that this might be a sign of potential issues. Many organizations manage product and service quality using quality management methods such as Six Sigma, Kaizen, TQM, PDCA, and BPR. Still, they often do not include organization data in those same processes, frameworks, and systems. Big Data grew faster than many organizations' operations, procedures, frameworks, and methods and are often overlooked or misunderstood by many quality management organizations.

Is this dirty data?

Often customers overestimate data quality and their ability to detect and resolve dirty data. Measuring an issue's scope and complexity is often challenging without definition and context. For example, what is the distance between three points if we do not have each location or a method to measure distance? The first step is, therefore, to collect and evaluate the current information. Comparing the existing data will often help highlight gaps, outliers, missing data, potential issues, and areas of improvement. Step zero is data collection since some data already exists but might need to be combined or evaluated for quality in additional steps. Data collection can also be part of the dirty data problem because organizations often start with the assumption that the data collection is not a part of the quality of that data. How information is collected is vital to both existing data and adding to data quality as we advance, just as clean water in a clean pool keeps the pool clean. In contrast, dirty water makes a clean pool messier. Management of the origin of the information is vital to data quality.

Data quality maturity paradox

Another major issue with customer data is the lack of mature processes developed for data quality evaluation. The first step is to create an understanding of your organization's data and the development of strategies and tools to manage that data. The second step of data processing is to simplify the collection of data. Most organizations start with the second step because it is often easier to understand the problem if they have some data to work with, but this can cause additional confusion. The dirty data maturity paradox is that the organization must evaluate data to develop strategies and techniques that best fit that organization to establish processes and systems to manage dirty data. Management can mitigate this issue by using continuous improvement to improve data quality. Existing data quality is used as a baseline and evaluated for maturity concerning strategic objectives that are refined and enhanced as data quality increases.

Data collection

Collecting data is a critical step because the quality of the data and the process used to collect that data usually heavily impact the overall data quality. If the organization does not have any good data sources or is uncertain about the data quality, collecting the data can serve as a good starting point from which an organization can measure improvements and provide a target for future data quality improvements. Data collection in systems like digital twin can be as simple as an asset audit. If a list of assets, location, placement, and conditions does not exist, the organization should invest in an asset audit as an excellent place to start for initial data collection. This audit data can become the baseline to measure data quality and serve as a guideline for future objectives.

Data preparation

The second step is data preparation. The organization should use tools that allow the data to be formed for further analysis and processing. Often this might be as simple as placing the asset list or audit data in a tabular spreadsheet file and then reviewing the data for accuracy and completeness. Figure 3 shows an example of processing typical data issues of completeness or inconsistency.

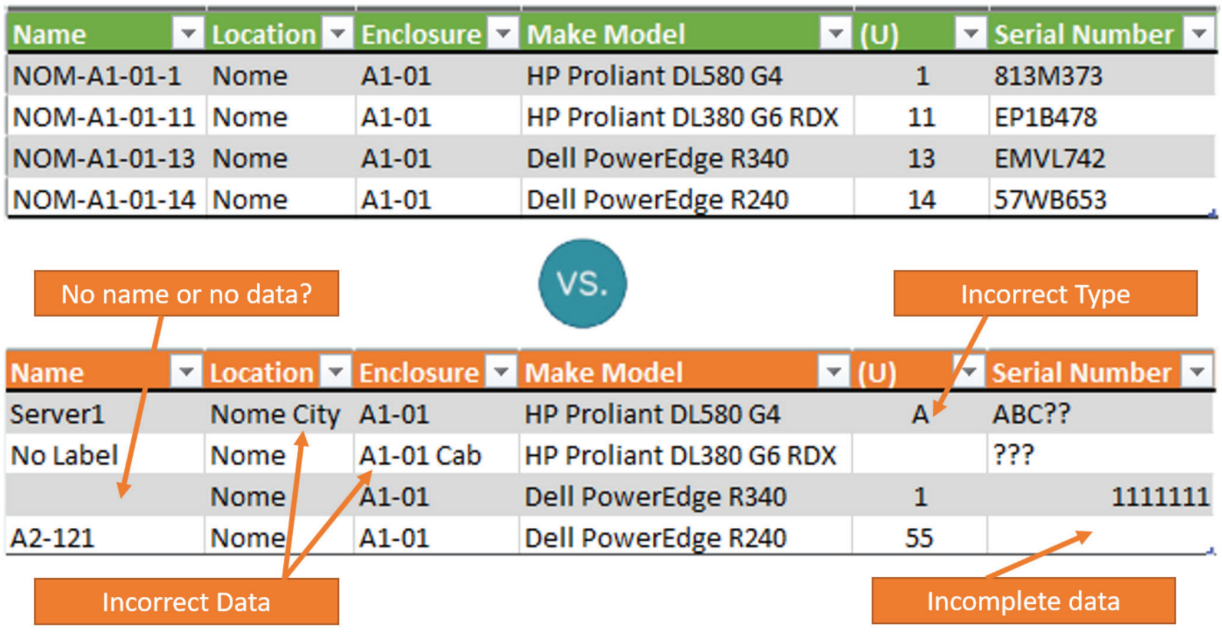


Figure 3. A traditional example of dirty data with easy-to-spot errors and incomplete data

The green-colored table shows an example of complete asset data, with each of the required data fields having the data collected and the data in a consistent format. The orange tab shows incomplete data in names, missing U positions, and serial numbers. Other issues like inconsistent naming of the "Enclosure" field can be found using spreadsheet features like column filtering or pivot to address topics such as the Enclosure name A1-01 Cab. These shared data quality issues can be addressed and resolved as part of the standard data preparation step.

Unfortunately, many organizations might not have addressed or reviewed this issue since spreadsheets can often allow bad data quality to grow and spread due to the lack of constraints, relationships, or other more powerful features found in modern database systems. In addition, a hidden step of knowing which fields are required and which areas can be ignored highlights the importance of understanding data from a strategic perspective. An organization can collect infinite data points, but that collection, evaluation, and management have costs that the organization needs to evaluate strategically. Collecting all data without understanding data quality is spending resources without considering cost and impact. Data can still be collected, but the source should have a quality measure associated with each source. For example, if we collect serial numbers of servers from a physical audit, that data might or might not be more accurate than data collected using software or other network methods. Methods should be compared for accuracy and graded based on the results. For example, suppose the physical audit is 99 percent accurate and 100 percent complete, and the software method is 100 percent accurate but only 30 percent complete. In that case, the software method should be used to validate the physical audit. The physical audit should be used since it offers more data coverage. The software method might be more accurate, but if the coverage is not 100%, it cannot be the primary method used to collect data. The software method would be the secondary method that could be used for data validation since the quality is higher.

Data quality dimensions

Data quality has many dimensions: accuracy, completeness, consistency, and currentness. Data preparation in a spreadsheet system is limited because spreadsheets lack most of these features. A standard audit of data using a system such as a spreadsheet has limited capabilities to check data for accuracy, completeness, consistency, and currentness because the user can quickly enter whatever data they want. This comfort level is a double-edged sword that often starts and ends in creating and maintaining dirty data. digital twin provides the user with methods to measure accuracy, completeness, consistency, and currentness in importing and preserving user data. These systems utilize database frameworks that provide methods, processes, and systems for ensuring accuracy, completeness, consistency, and currentness, which are missing in spreadsheets and other frameworks. These frameworks and processes guide the user in what data is required and essential without understanding strategic objectives or overall organizational data quality standards.

Don't kill the messenger

Unfortunately, the phase between data preparation and data evaluation is where organizations might discover one of Big Data's biggest hurdles. This aspect of dirty data is called the "bad news problem" (i.e., people don't like bad news). Coping with bad news is not an issue related to dirty data, but it shows up more often and in much clearer ways when managing data. It is usually a hurdle toward improving data quality. Effective organization management often focuses on risk factors and processes associated with negative information. An organization should have strategies, methods, and tools to manage negative information so it doesn't kill the messenger who comes bearing bad news. Killing the messenger encourages people and processes to hide bad news and does nothing to discover the root cause or provide corrective action.

The organization should also avoid the runaway messenger who drops the message off and leaves before it is read. If a system presents awful news without processes, people, or procedures to evaluate the message, this situation is also problematic for the organization. The bad-news problem is often best handled by providing a neutral messenger or a system designed to take that news and direct it toward analysis and action. If the system directs the user toward action and not blame, the outcome is often better in the long run. Suppose the organization relies on the messenger to deliver

a negative message. In that case, the messenger is incentivized to spin the information to put them in a positive context or frame of reference. Data can become dirty again because the context has been changed to put a spin on a specific message.

The best method to solve the messenger problem is to provide a neutral approach for both good and bad news to be presented to the system and a plan to process each message from the initial report, analysis, action, and completion.

The bad-news problem

Something is often missing even when an organization believes it has resolved the issues in data accuracy, completeness, consistency, and currentness. Expert systems like a digital twin provide better frameworks and systems than tools such as spreadsheets. They are purpose-built to give the organization systems and processes that evaluate data quality. digital twins also allows for contextualizing data aspects found in specific data relationships and understanding particular data structures applied to specific contexts such as a data center, cabinet, or power configuration. These tools can also provide a method of managing the bad-news problem that often is hidden by using tools such as spreadsheets. For example, in Figure 2, both tabs appear correct regarding completeness, consistency, and currentness. Data accuracy is often hard to determine because accurate data is defined as being error-free. In the case of Figure 2, the green tab is valid, but the orange section has errors and is, therefore, inaccurate. What is the difference, and how can we figure this out?

Name	Location	Enclosure	Make Model	(U)	Serial Number
NOM-A1-01-1	Nome	A1-01	HP Proliant DL580 G4	1	813M373
NOM-A1-01-11	Nome	A1-01	HP Proliant DL380 G6 RDX	11	EP1B478
NOM-A1-01-13	Nome	A1-01	Dell PowerEdge R340	13	EMVL742
NOM-A1-01-14	Nome	A1-01	Dell PowerEdge R240	14	57WB653

VS.

Name	Location	Enclosure	Make Model	(U)	Serial Number
NOM-A1-01-1	Nome	A1-01	HP Proliant DL580 G10	8	813M373
NOM-A1-01-10	Nome	A1-01	Dell PowerEdge MX7000	10	EP1B478
NOM-A1-01-10a	Nome	A1-01	Dell PowerEdge MX740c	13	EMVL742
NOM-A1-01-10B	Nome	A1-01	Dell PowerEdge MX750c	14	57WB653

Figure 4. This example is much more typical of dirty data, making errors harder to spot

Data accuracy can often be measured as a relationship to other data elements. Spreadsheets do not manage context or connections. So minor data issues can easily be hidden if the news is not optimal. In addition, if we do not have all the data to analyze, we often cannot determine if the data is good or bad. Therefore data will tend to be presented in a positive light even in the case where the data might be hiding issues. The spreadsheet, as the messenger, has limited methods and tools to provide unbiased information; therefore, negative information is often lost in the data management process. Digital twins can have similar issues if organizational procedures and methods allow for unmanaged negative information. However, a digital twin often provides a better framework for supporting and providing a neutral messenger for the bad news.

A better way to deliver bad news

Understanding aspects of the data, such as naming conventions, can help answer the bad-news problem. a digital twin can provide a method to evaluate asset naming against an organizational standard. Unfortunately, organizations often draft naming standards without assessing and collecting data. Outliers and edge cases can cause many issues, and often users will attempt to make something fit into a model that was not developed using real-world data. In this case, what is the server installer to do when the naming standard does not match reality? They make the best decision for that case, record the data, and move on. But who is responsible for updating the naming standard for the new edge case? Who gets to tell the standard naming group the bad news that their standard does not work in the real world? Who gets to be the messenger of the bad news?

Evaluation in context

In this case, in Figure 2, we have two assets: NOM-A1-01-10 and NOM-A1-01-10a. The naming standard, in this case, is the site name code followed by the cabinet name followed by the U position of the device, i.e., [Site Code]-[Cabinet Name]-[U Position]. Since the dataset contains both the name and the U position, this can be a helpful key to understanding the position. The naming of this item might indicate something about the position and location of that asset. For example, an asset NOM-A1-01-10a with position 13 might be a naming issue or a sign of another potential data issue. The data point 'a' at the end of the name might lead to additional confusion about the actual location. For example, either the device's naming is wrong, the device's position is inaccurate, or there is some additional missing data point. If we dive into the data, we discover that a Dell PowerEdge MX7000 is a blade enclosure and a Dell PowerEdge MX740c is a blade server fitting into it. Blade servers are installed in blade enclosures, so there might be a missing relationship between the devices that explains this naming issue. What can we do to provide a framework to present the user with more information? The bad news is often hidden in tabular formatting. Using a digital twin's visualization can provide additional insight into what is happening in this case and allow a method to present and resolve the issue in the future.

Providing more context

The simplification of the tabular format often abstracts too much critical information and might give the user a sense of overconfidence in the data quality. Visualization or other UI tools might help provide more context and help clarify the issues. A good example would be to model the data shown in Figure 2 and visualize the data to highlight the potential problems. At a high level, there does not seem to be anything incomplete or incorrect in the tabular format. Using a non-standard ending such as A and B in NOM-A1-01-10a and NOM-A1-01-10b might highlight some issues with the asset's position. Visualization of the data can help provide more context and allow for correction.

Figure 5 shows the visualization of the valid data from Figure 4. This visualization provides the user with more information and additional context. The size and position of the devices are shown. Visual data is suddenly available in the context missing from the tabular format. The user can see the device's position as it appears in the actual data center. Additional information includes open areas for new rackmount items, placements, and weight distributions. Some servers look to be older models or legacy devices that should be decommissioned.

Figure 5. Here is an example of good data with visualization

Visualization provides context

Now let's visualize the data from the second part of Figure 3 to understand why there are issues with this data and what makes this data dirty compared to Figure 4. Visualization of the data provides additional context that might help provide insight into the problem with the naming. Figure 5 shows that NOM-A1-01-10 could only fit at U position 12 and above because NOM-A1-01-1 is listed at U position eight and takes up four U positions. The naming also suggests that the device's actual position is at U position 1. There could be many possible issues with this data. The model of NOM-A1-01-1 might be incorrect, the placement of the device might be wrong, or perhaps even NOM-A1-01-10 is actually at U position 12, and the machine was not renamed. In addition, both NOM-A1-01-10a and NOM-A1-01-10b are listed as servers at U positions 13 and 14. When modeled, the servers are shown to be blade servers inside NOM-A1-01-10.

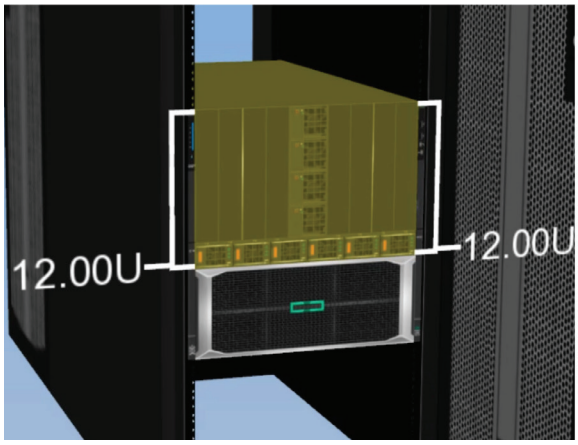


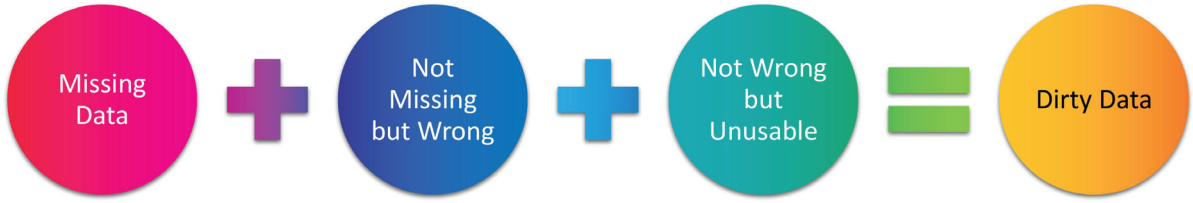
Figure 6. Both servers are larger than the data suggest and overlap each other

Figure 7. NOM-A1-01-10a and NOM-A1-01-10b are blade servers inside NOM-A1-01-10

The visualization of NOM-A1-01-10a and NOM-A1-01-10b shows that the U position data is incorrect. The data entered was either a mistake, a misunderstanding—or perhaps an installer trying to fit into the existing naming standard but uncertain what to do in this specific case. Since no framework existed to provide the installer with support or context, the user may have attempted to guess the best naming standard in this case. In addition, there may have been no method to find or correct the naming method or feedback on the edge case to the standards group using a tabular data collection method. The bad news was that the naming standard did not fit, did not apply to blade servers, was misapplied, or several other potential issues. Users that viewed this issue after installation also might not have been able to correct the naming or not had a method to report the problem. Management would not have a process to find this issue or apply context or a framework to view other outliers, so the dirty data remains in plain sight in the tabular data.

Common customer issues

The focus on tabular data and relational databases has resulted in filling in the blanks rather than understanding the context and applying organizational processes in the data-cleaning journey. Part of the problem is that most organizations do not view dirty data as an issue that is a critical part of strategic planning. Dirty data is a typical customer issue that needs to be addressed strategically. Tools, systems, resources, and processes should provide a framework for improving data quality and evolving toward higher levels of data quality maturity in the scope of organizational strategic objectives. One group or individual cannot address the issue of data quality without the resources and support that come from strategic alignment with corporate goals. The first step is understanding that this is a common issue and has a place in organizational strategic planning.



Missing data

Organizations often miss the importance of the difference between a null and a blank value. Is the value missing, or is it blank on purpose? Let's take, for example, a list of servers in an inventory. What is the issue if most servers in the data center have serial numbers and one does not? Is this missing data? Is this because someone did not capture it? Maybe this server does not exist? Perhaps the server is misconfigured? Often servers' serial numbers might be linked to licensing or other installation factors. Therefore, if a piece of data exists, it is best not to use blank or null as a starting condition. The data can be linked to enforceable constraints or default values, but often, data should be connected to the process. For example, instead of having a blank serial number, it might be better to default the value to "Pending installation" or "Pending licensing." This condition helps flag error conditions or pending actions. If the server is installed and the status is "Pending installation," an automated task can be created to have a technician audit that serial number on the new regularly scheduled audit event.

It is also OK to delete data points. In the world of Big Data, it is often said that we should keep all the data. Suppose the data is empty or has the wrong structure, or does not provide some value to the organization. In that case, it should be removed from the overall data structure since dirty data tends to form a sort of data pollution that can have a corrosive effect on the overall data health of an organization. For example, suppose the organization does not include barcode items but has four servers out of 1,000 with the barcode information. In that case, it can be OK to delete that data point or mark the other barcode field with a default value of "Missing barcode." If the barcodes still exist on the server, deleting the barcode from the data structure will not remove any helpful information that cannot be gathered again if the organization changes direction.

Not missing but wrong

Now that the blanks have been filled with some data or process steps—or unhelpful data points have been removed from the overall structure—the organization can take the next step toward data maturity. The focus is now on data that is not missing but is wrong. This data type is often tough to detect, but expert systems can help fill the gaps. These systems can often provide new points of view to help that organization improve its data. Tabular systems like spreadsheets often inject these "not missing but wrong" errors due to the non-enforcement of constraints. The most significant areas of concern in tools like spreadsheets are the addition of user-specifiable conditions. Users can put whatever they want in the data without understanding the data structure, such as:

- Invalid data types: When a user puts a string in for a number or a number for a text
- Dangling data: Such as if a single server was placed into two or more cabinets, e.g., the server was moved, but the reference to the old server still existed in the data and the old location
- Duplicate data: Multiple entries of the same information or multiple entries of data with minor differences
- Mutually inconsistent data: Such as when various servers have the exact name location but different serial numbers and configurations

Often these duplications might result from multiple systems storing information about a device.

Other areas of concern are lost updates, wrong categorical data, outdated temporal data, inconsistent spatial data, erroneous data entry, misspelling, extraneous data, entry into wrong fields, and incorrect derived-field data. Many of these types of inaccurate data require data audits, pattern matching, or periodic physical audits so the data structure can be understood and programmatic methods can weed out areas of concern. An expert system can solve many structural and procedure issues with invalid data associated with these integrity constraints. Still, these issues often start and spread in tabular systems such as spreadsheets and are harder to correct without specialized tools. Understanding the data structure and then pivoting the data to look at it in new ways can help identify and fix these issues. For example, checking the server installed date data for outliers might find erroneous data entries if most of the entries have dates within the past ten years and two servers contained install dates in 1921. These two servers should be flagged for further data analysis, and if issues are found, the system could be improved to prevent this extraneous data entry. 3D visual verification of server placement can also show potential problems if two servers overlap in U position placement.

Not wrong but unusable

The last step on the road to data maturity is to develop processes, methods, and systems to deal with data that is not wrong or missing but unusable. Once techniques, procedures, and strategies have been implemented to deal with missing and erroneous data, the deep dive into usability begins. Often customers ask: If the data is not wrong, how can it be unusable? The first part of incorrect data should be solved as part of the "not missing but wrong" step, which addresses different data associated with single devices. There should not be one server with the same serial number that shows it is installed in two other places in two different systems.

Another insidious issue is ambiguous or nonconforming data. For example, inconsistent abbreviations such as "20W" might be 20 watts, the floor position for space 20xW, or a typographical error. Non-standard conforming data, different representations of non-compound data, and various abbreviations for different data types or nicknames instead of formatted names can all cause data to become unusable. Algorithmic transformation issues, such as importing different data types into other systems using various methods, can cause invalid data. Even problems such as the use of leading zero and their programmatic removal issue in a spreadsheet can cause loss of information when saving to different formats such as CSV. Measurement unit differences, such as the conversion between imperial and metric measurement, can also cause data to become unusable if they become mixed. In addition, using special characters, such as the difference between using a comma and a period as numeric space separators in different countries, can cause data usability issues. Other more complex examples are the representations of compound data, such as concatenated data and

abbreviated versions (e.g., John Kennedy for John Fitzgerald Kennedy). Another example is using special characters such as space, no space, dash, and parenthesis in data such as Social Security numbers or phone numbers.

Data that is not wrong but unusable is often tough to clean and requires special tools, processes, and systems. This issue also requires strategic investment to clean the existing data and develop strategies and techniques to prevent or reduce these issues in the future. The most challenging data quality issues to resolve usually require strategic oversight. For example, a single individual in an organization below the executive level usually does not have the authority to determine the official organizational measurement unit used by that organization. For example, if that user cannot select the dataset's type of measurement, such as imperial or metric, the user might ignore the issue. The issue goes beyond the scope of the authority of the end user to resolve, and the invalid data continues to exist. All organizations should have a process to report more significant strategic data issues, providing a root cause and corrective action process. This process should be developed at the strategic level.

Journey, not a destination

Cleaning dirty data is a continuous improvement process and is more of a journey than a destination. An organization's first step to combat dirty data is to understand that this is a common problem that is often ignored, misunderstood, mismanaged, or even hidden. Dirty data usually does not develop overnight; it is an evolutionary process that has grown over time. Any user who stumbles upon dirty data might be at first eager to resolve minor issues. Still, as the size and complexity of the problem mount, it becomes something a single individual in an organization cannot manage. It is often difficult to articulate the complexity and nature of the issue to executive management or even about a problem that might be perceived to be associated with individual blame. For example, suppose a manager is hired to manage a data center and finds that many servers do not have a position or serial number information. The manager might try to resolve the issue by collecting the data in that case. If the scope becomes too large for one individual, that user might not be willing to escalate the matter to higher management due to cost or other concerns. The issue continues to plague the organization without a path toward improvement. Dirty data continues to grow in the darkness of miscommunication, misunderstanding, and suboptimal processes. The issue is highlighted when aggregate data is analyzed, and conflicting information is suddenly presented.

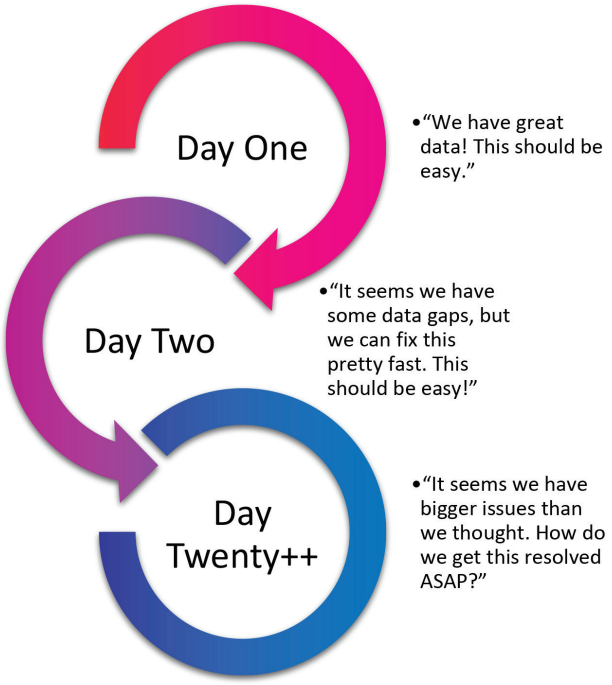


Figure 8. Dirty data usually hides on day one

Delivering bad news

Delivering bad news is never easy. At each step in the dirty data problem, we have the potential for the bad news. Customers, partners, and stakeholders never like to get bad news. The bad news often includes increased costs, implementation delays, satisfaction decreases, and overall discontent. The first step is understanding that bad news is a critical part of the dirty data problem, and hiding or not understanding the value of bad news is usually one part of the equation that leads to dirty data. The first part of delivering bad news is framing feedback. In an ideal world, customers, partners, and stakeholders openly accept corrective feedback. They would gather input, ask questions, develop action plans, and work toward improvements. Unfortunately, this is rarely the case—very few people thank their dentist for finding cavities.

Figure 9. Methods of presenting bad news

Framing terrible news is simply providing a workable set of issues to help provide an actionable plan to guide the decision-makers image of the situation. In the example of missing asset names, the organization might think only about the cost of gathering more information or performing new audits. It might be hard to sell the value of collecting more data if all the data is bad news. The value is in providing better information that can help decrease costs, reduce future confusion, and allow for more automation and system integration. Reframing bad news allows for more action-oriented planning and enables decision-makers to scope these actions' impact on other organizational aspects.

Digital twins allow for data improvement because they allow methods of reframing the dirty data problem into actionable items. Decision-makers can use a digital twin to frame the scope of the issue, generate actions, and provide a baseline for future improvements while helping to control the size and, thus, the cost of these changes. A digital twin allows a structured method to evaluate data and pivot different datasets around a common framework. Some digital twins contain visualization, tracings, reporting, dashboards, templates, and other tools to help reframe data so it can be evaluated for gaps, missing data, or other abstract issues.

Yellow brick road to data maturity

Solving the dirty data problem does not happen overnight. The road to good data quality should be understood as a data, systems, and organizational maturity process. The first step is understanding that data should always be evaluated for quality. The dirty data problem starts with failing to treat data as a strategic business asset. Each organization should answer the question: What does good data look like to us? With a strategic focus, what does good data look like as an organizational asset? With this strategic mindset, the first step is to start small and focus on the highest areas of impact that can help reduce dirty data. Step zero is often skipped at this stage because it should be a fundamental question for any organization. What data do we have? The current dataset should be collected and evaluated to understand the basic structure of that information: What data fields are gathered now, and where does the data currently exist? This process is called "step zero." Often this is the first activity an organization does that highlights the potential dirty data problem. The challenge is that, frequently, evaluating a single data source might highlight the dirty data issue and, therefore, the evaluation of only that data source when the scope should be increased to assess all data sources on a strategic level. Data should be understood as a strategic business asset; therefore, the limitation to a single data source will minimize the overall impact of cleaning that specific data. In addition, data points might be missing or removed that could benefit other areas of the organization. For instance, the cost of an asset like a server might not interest the IT group but might be attractive to the financial or accounting departments.

An excellent first step is understanding the data required for strategic organizational success and filling those blank areas. The data areas of concern should be part of the overall corporate strategic plan and focus on planning, execution, monitoring, and continuous improvement. The best first steps tend to be an overall strategic look at the importance of the data to the organization and the areas of the data that are linked to key performance indicators (KPIs). High-level strategic KPIs tend toward financial indicators, which should be easy to determine. Most organizations have an excellent method to review, evaluate, and improve this data, so the key is to determine how other data points connect toward the high-level KPI. It is often challenging to associate the collection and storage of server serial numbers with high-level financial objectives. Still, it might be in operational stability or failure rate analysis. The question for the organization should be: Why do we collect this information? What benefit does this information provide to the organization? The organization must look at the data to fill in the blanks and pivot the data to see if there are obvious issues.

Structure of the data

Often data structure is built organically. What is the design of our data? There was a reason someone wanted to collect, document, or store some bit of data about some object or asset. The list grows until often the structure of an organization's data becomes like an amoeba, with many small arms linking back toward the main data body. The first step is to pull all these data sources and look at the bigger picture. The Big Data problem does not need to be solved, but a basic map of where the data lives, who or what updates the data, and who owns the information is critical to understanding the data structure. To walk down any road, you must first take a step. The first step in the yellow brick road of data maturity is finding these data sources. Once this first step of understanding the data structure is taken, the organization can begin the journey toward increasing data maturity.

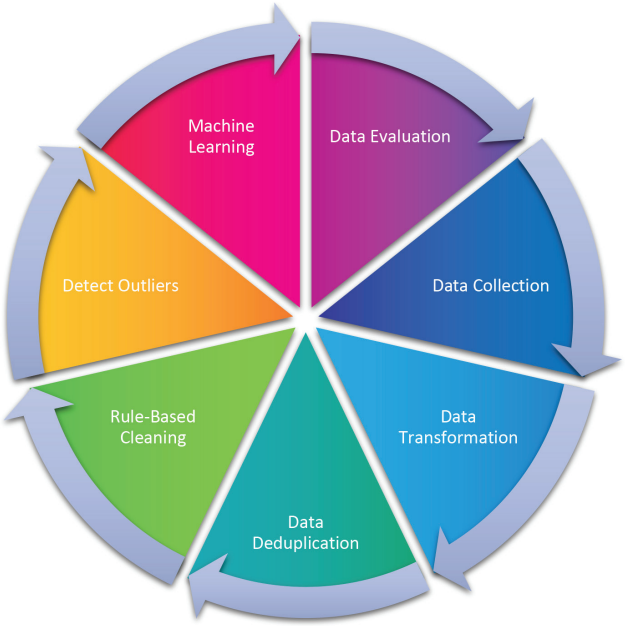


Figure 10. Data quality is a continuous improvement cycle

Single source of truth

The best method to resolve these issues is to have a single source of truth (SSOT) for specific data types. For example, space and position data on servers might be stored in a digital twin, and the IP address and network configuration information might be stored in a network management system. Each system contains information about the server, but the SSOT on location and placement is stored in the digital twin, and the network information is stored in the network management systems. The question is: Why have a digital twin and a network management tool? If an organization needs an SSOT, shouldn't there be one system that can be both a digital twin and a network management tool? For some organizations, these systems can be combined since some digital twin products have network management tools and some network management tools have location and placement. Unfortunately, due to the complexity and variety of organizations and their strategic objectives, there is no universal best practice and best tool that fits all use cases and environments. Each organization should evaluate the tools needed to form an SSOT related to its strategic plan. Just as there is not one single tool or system that can do all an organization needs to accomplish strategic objectives, there usually is not a single system that can be an SSOT.

If there cannot be a single system that can be the perfect repository for all organizational data, how can there be an SSOT? The answer is that organizations need to adopt systematic thinking on how data can be effectively managed concerning that organization's strategic objectives. The SSOT does not need to be a single system such as a digital twin

or network management tool but can be the application of a process that attempts to design, integrate, and manage these complex data systems in relationship to the organization's strategic objectives. SSOT is a systems problem and not a tools problem. Data can be in multiple systems—each managing the data that best fits that system's use case. Each SSOT should be linked or integrated to ensure data quality. Each design helps to improve the overall data quality of the organization. The proper use of integrations can help provide a continuous improvement feedback loop that helps generate a systematic SSOT.

Data cleaning steps

There are nine major areas of managing dirty data and essential questions that must be answered to develop robust processes to push an organization toward a high-quality data process. The overall order of the steps should be part of a continuous improvement process.

Data Cleaning Steps	Key Questions
Data Quality Evaluation	• How do we know we have good data? • How do we measure good data?
Data Collection	• How is data collected? What is the source of the data? • Does the current process add or remove dirty data?
Data Transformation	• How do we get data from one source to another? • Is something lost or corrupted in the change, e.g., leading zeros?
Data Deduplication	• Is the duplication a duplication, e.g., the same name on two devices? • What caused the duplication? Can it be fixed?
Rule-Based Cleaning	• What are the boundaries of good data? • Can data errors be repaired using a rule-based system?
Detect Outliers	• Is this information correct? Is this a one-in-a-million example? • Does the outlier show us something new, a potential issue, etc.? • What is the difference between an outlier and a mistake?
Machine Learning	• Can automated systems be used to clean the data? • Are there patterns that can be understood and corrected?
Human UX	• Can we use the UX to help clean data and find errors? • Does visualizing the data help to highlight patterns?
Continuous Improvement	• How can the data cleaning process be improved? • Can we automate the collection and improvement of data?

Figure 11. Data quality is a strategic organizational question

The overall data cleaning steps need to be viewed in the context of the organization's strategic objectives related to data quality. Often this framework and questions can be used to help guide the overall strategy's development. One of the more significant issues is that it is hard to determine the best areas and places to improve data quality. What data quality objectives would provide the organization with the most significant benefits? Improving data quality will consume some organizational resources, so the best method would be to find the areas of data quality that could best benefit that organization.

Data quality maturity and heat mapping

One method to help streamline the organizational data quality strategy would be to import the data into a system that could help provide context and visualization of that data quality. For example, let's use a small data center to resolve the missing data issue. We have imported all the data into the system. We have included data points such as purchase order, service designation, serial number, owner name, owner ID, owner email, owner phone number, and various other data points. The spreadsheet used to import the data had many missing data points. How does the organization decide which area to target first? Which data points have the highest fixed value for the organization?

Figure 12. Example data center

If we take a heat map of the items just based on if any items are missing data, this task suddenly seems overwhelming. Everything is red, besides a few outlier devices containing all the information.

Figure 13. Heat mapping for 100 percent data quality is usually always red at first

Organizations are often stopped at this point and locked into indecision. This bad-news problem causes some organizations to become overwhelmed and stuck at the first step of data quality evaluation. Therefore, the best practice is to understand that each organization starts with a specific level of data maturity. Data quality maturity is a strategic paradox. The heat map that shows all red should be used only if the organization has a high level of data quality maturity. If the heat map is red, the organization does not have high data quality maturity. If this happens, remember not to kill the messenger and to understand that data quality maturity is part of the journey. In this case, we would often recommend lowering the data completeness score to a lower maturity level and attempting to assess what is different.

Figure 14. What areas of data quality can be improved first?

By lowering the data quality maturity level, we find that a couple of outliers might explain the difference. Many data points do not include the owner's name but contain an owner ID. Does the digital twin need to hold the owner's name if the owner ID is available? If the owner's name and contact information are contained and updated in an SSOT for contact information, those data points can be removed. Unless the SSOT for contact information is the digital twin, the owner's name, email, and phone number can be removed. The focus can be on ensuring that the owner ID is always provided, so the link to the SSOT on contact information is available to the user.

Once again, we run the evaluation. The overall data quality has improved, but there is still room for additional evaluation and improvements.

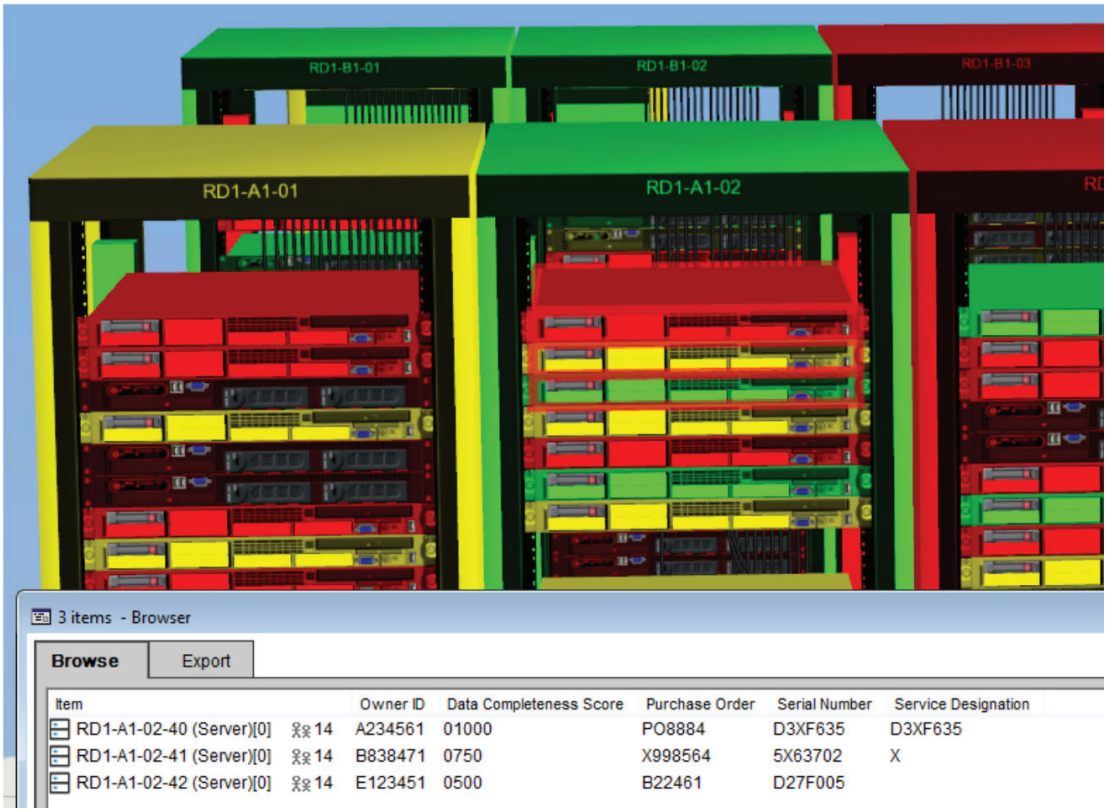


Figure 15. Target specific aspects of the data and remove duplicate data

In this case, we sample three servers and find there seems to be an issue with the service designations data. One server shows the service designation matching the serial number, while the other has a value, but it does not match the expected pattern, and the last server is red since that value is missing. A review of other servers shows that the service designation is a legacy value carryover that matches the serial number in all cases. Since the information is duplicated and is no longer used, that value can be removed from the system. Information about the relationship can be stored in the knowledge base and outlined as part of the data quality strategic plan.

We can remove the extra data point, rerun the evaluation, and improve the data quality heat map.

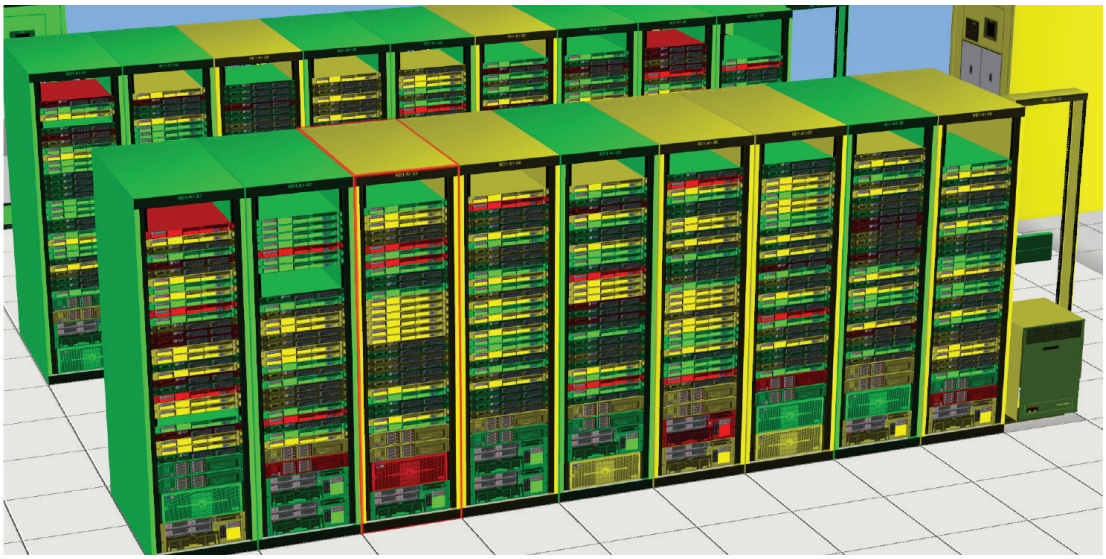


Figure 16. Focus on taking action to collect missing data points like serial number

Data is still missing or of unknown quality, but there seems to be a better sense of what might be an excellent focal point to invest in improving specific asset data quality. The missing data can be evaluated to find patterns applied to actionable tasks. Many items are missing serial numbers that can be collected, or items with missing serial numbers can be assigned an organization-specific asset number to replace those data points. After the serial numbers have been collected, the heat map shows improvements. This point might become more complicated if the information is more difficult to correct, such as missing purchase orders or owner IDs.

Figure 17. The remaining data quality tasks might take more time and resources but should not be ignored

Additional work might be done to track down the owners of the devices who might have information about the purchase orders. Devices with missing owners and purchase orders might be in line for a decommissioning process since the use and purpose of the device is unknown. Some devices might not be functional, or end of life, and decommissioning devices can save the organization resources in terms of power and space. Each evaluation of the data quality heat map should improve the overall data quality of the data center.

Figure 18. The remaining data quality issues should be reviewed and improved as part of the new data quality process

Some devices might have outstanding actions to complete. These devices would not be 100 percent yet, such as if an owner is still working on providing the purchase order or new equipment recently installed with a purchase order has not yet been assigned an order. As new equipment is installed or removed, the information quality improves. The organization can increase the data maturity level as data points are added or removed. The use of heat mapping can be an effective method of targeting data quality improvements in specific functional SSOT systems such as digital twin.

Pivot and improve—using different perspectives for data analysis and cleanup

Dirty data is the Achilles heel of modern Big Data and machine learning. Leveraging good quality data for analytics, predictive models, continuous improvement, and innovation can be critical to providing organizations with an edge over their competitors. Since data quality is a fluid and changing aspect of any organization, systems and processes must be implemented to determine how high data quality is maintained in this organization. An organization should have methods to evaluate and improve a specific data model to check for data quality and continuous improvement. There needs to be a strategic analysis of data quality, organizational objectives, systems, and processes. SSOT tools such as iTRACS can provide organizations with tools and methods to help solve these complex problems.

References

Alexandrov, Alexander, et al. "The stratosphere platform for big data analytics." The VLDB Journal 23.6 (2014): 939-964.

Ardito, Lorenzo, et al. "A bibliometric analysis of research on Big Data analytics for business and management." Management Decision (2018).

Cai, Li, and Yangyong Zhu. "The challenges of data quality and data quality assessment in the big data era." Data science journal 14 (2015).

Côrte-Real, Nadine, Pedro Ruivo, and Tiago Oliveira. "Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value?" Information & Management 57.1 (2020): 103141.

Cichy, Corinna, and Stefan Rass. "An overview of data quality frameworks." IEEE Access 7 (2019): 24634-24648.

Dong, Xin Luna, and Divesh Srivastava. "Big data integration." 2013 IEEE 29th international conference on data engineering (ICDE). IEEE, 2013.

Franklin, Mike. "The Berkeley data analytics stack: Present and future." 2013 IEEE International Conference on Big Data. IEEE, 2013.

Gardner, Lauren, et al. "A need for open public data standards and sharing in light of COVID-19." The Lancet Infectious Diseases 21.4 (2021): e80.

Kim, Won, et al. "A taxonomy of dirty data." Data mining and knowledge discovery 7.1 (2003): 81-99.

Marsh, Richard. "Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management." Journal of Database Marketing & Customer Strategy Management 12.2 (2005): 105-112.

Perez-Castillo, Ricardo, et al. "Data quality best practices in IoT environments." 2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC). IEEE, 2018.

Swartz, Nikki. "Gartner warns firms of dirty data." Information Management Journal 41.3 (2007): 6-7.

Qi, Zhixin, et al. "Impacts of dirty data: an experimental evaluation." arXiv preprint arXiv:1803.06071 (2018).

Wang, Jiannan, et al. "A sample-and-clean framework for fast and accurate query processing on dirty data." Proceedings of the 2014 ACM SIGMOD international conference on management of data. 2014.

CommScope pushes the boundaries of communications technology with game-changing ideas and ground-breaking discoveries that spark profound human achievement. We collaborate with our customers and partners to design, create and build the world's most advanced networks. It is our passion and commitment to identify the next opportunity and realize a better tomorrow. Discover more at commscope.com

COMMSCOPE®

commscope.com

Visit our website or contact your local CommScope representative for more information.

WP-117359-EN (01/23)