Data hygiene: maximize your data

CCA Europe

The 2022 Experian survey revealed that inaccurate data hinders the ability to respond to market changes, as reported by 77% of interviewees. Managers highlighted that poor data quality negatively impacts the customer experience (39%) and noted that 84% of employees are not skilled in data analysis.

Effective data management is crucial for providing efficient customer service and delivering valuable products. A key aspect of data management is data hygiene, which ensures data reaches its full potential and integrates seamlessly with AI.

In our e-book, we focus on how to keep your data clean, accurate, consistent, and up-to-date to maximize its value in business processes. We present strategies for maintaining data hygiene that enhance the reliability, integrity, and usability of your company's data.

TABLE OF CONTENTS

What is data hygiene?

Data hygiene involves the **intentional and careful handling of data**. Within your company, you have both structured and unstructured data scattered across various databases and files.

Data hygiene encompasses the actions and activities required to keep all your data reliable, up to date, and error free. It's not merely a technological issue; it's also an integral part of your organizational culture. This means thoughtfully collecting, structuring, and managing data to provide **an accurate picture of your business's current status**.

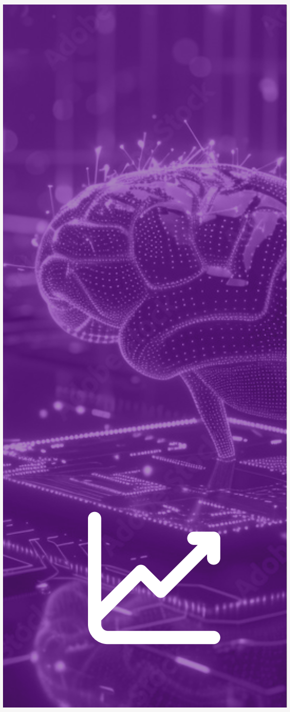
Why is data hygiene important?

Because it promotes:

safety

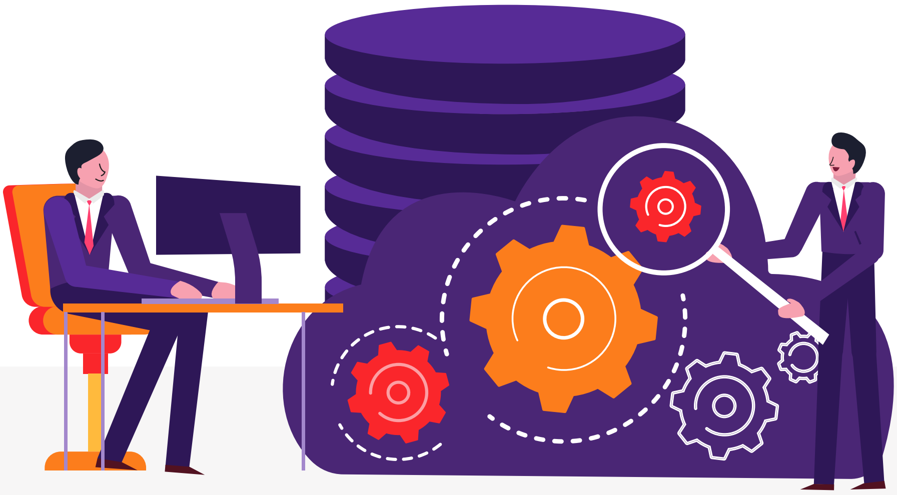
productivity

regulatory compliance



efficiency

Your company collects a range of data **to make informed decisions** that minimize risks, predict product demand, and control costs. This is possible when your business applications and processes **use only clean, accurate, and relevant data**.



How to obtain such data?

1. Determine your business goals.
2. Identify key data: Determine the essential data needed to meet your goals. Plan how to collect and process this data. Identify critical stages in production where important data is generated, ensuring it is not lost.
3. Secure your data: Ensure that data is protected from unauthorized access.
4. Fix errors and remove duplicates.
5. Monitor the quality and timeliness of your data.

Problems with data in an unstructured database

Symptoms of poor data quality include the following:

Data Duplication

Records in the database appear more than once. For example, the same person might appear multiple times with different sets of data.

Omission of Data

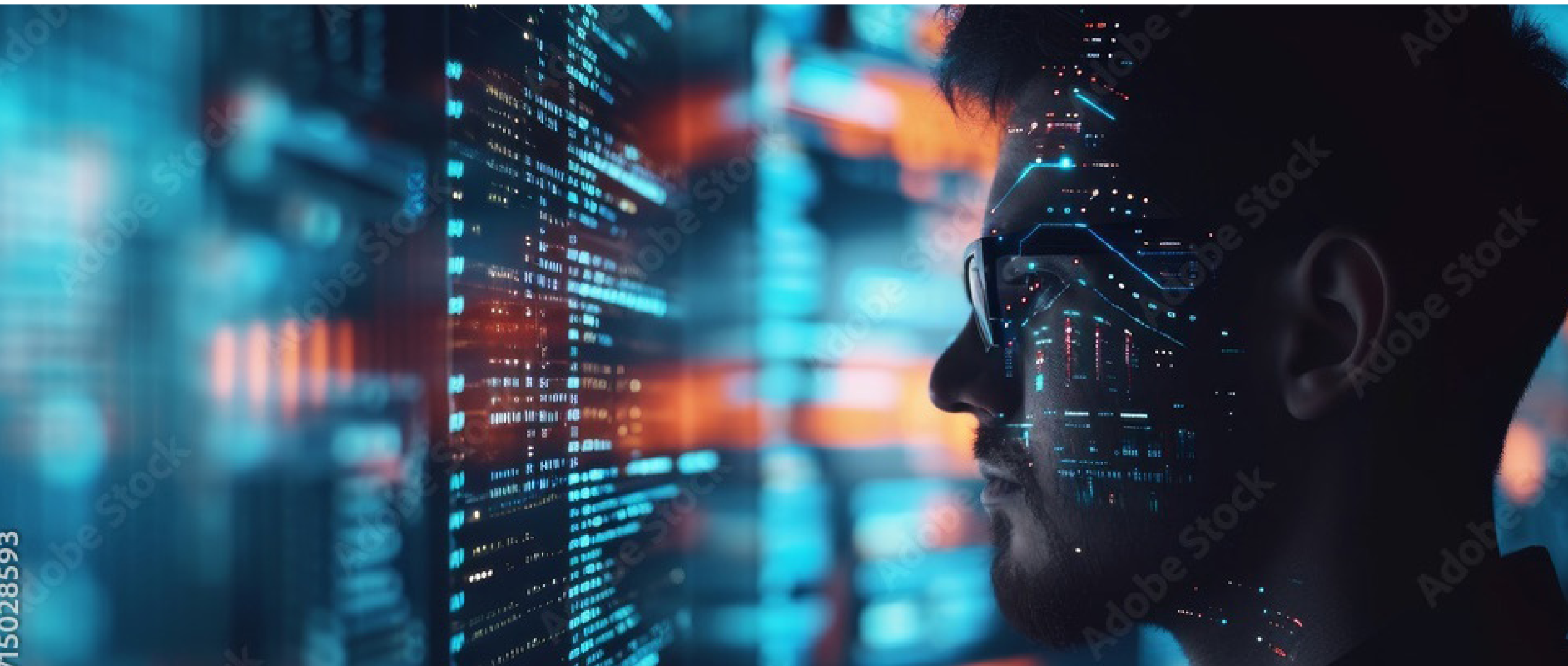
Searches do not return all the required data for a particular situation, preventing you from getting a complete picture.

Data Inconsistency

The same data exists in different formats across multiple tables. This results in multiple files containing different information about the same object or person.

Data Inaccuracy

The database contains incorrect or outdated values, making it difficult to make informed and optimal decisions.



Low-quality data:
typical problems within a company

Sales and marketing

DiscoverOrg conducted a study in 2019 on data quality in companies, revealing that sales and marketing departments lose about 550 hours and up to $32,000 per sales representative due to incorrect data.

According to an MIT Sloan report, data analysts spend 60% of their work time cleaning and organizing data. Other employees waste up to 50% of their time manually sifting through important data and improving its quality.

In marketing, this can lead to overspending. Potential customers may become annoyed when they receive the same content multiple times due to duplicate data. This is a common issue when multiple records in the database have the same name but are stored slightly differently. Such minor mistakes can cause significant reputational damage.

In online sales, poor-quality data can result in customers receiving the wrong products. This risk arises when there is no reliable data about the products and target audience. What happens if there is no automatic record verification in the database? Imagine a customer's VAT number accidentally placed in the phone number field. A courier would be unable to contact the customer with delivery information, leading to further complications.

Low-quality data: typical problems within a company

Finance and banking

In financial reporting, inconsistent data can yield multiple answers to the same question. This inconsistency results in inaccurate reports that may mislead decision-makers. They can give a false sense of security or, conversely, an alarming sense of financial insecurity.

Errors in revenue or cost data can lead to a misallocation of resources or an overly optimistic assessment of a project's profitability. Discrepancies in accounting methods or cost classification further compound the issue, making financial reports inconsistent and challenging to compare. This complexity hampers the assessment of company performance and impedes the ability to make strategic decisions.

Low-quality data: typical problems within a company

Manufacturing

Data quality is crucial in production, as even minor inaccuracies can result in losses and erroneous decisions. For instance, outdated material prices in cost estimates can distort profit margins.

Moreover, poor data quality adversely affects production growth and profits. Research from the Institute of Industrial Management at RWTH Aachen University indicates that supply chains lose between 1% and 3% of productivity due to data quality issues, costing manufacturers an average of 0.5% of their revenue. The quality of data aggregated by companies significantly impacts market success and stable growth.

Production data is often intricate and disorderly, sourced from various outlets such as machines, sensors, and software systems. Integrating data from these disparate sources can be challenging and resource-intensive, complicating data analysis. Although manufacturing companies adopt MES-class systems, these alone do not resolve deeper data analysis issues.

Challenges persist in analyzing the vast amounts of data that stream in daily, necessitating their translation into various facets of manufacturing operations, including employee objectives. Moreover, changing parameters or generating additional reports and visualizations incurs additional costs and prolongs data retrieval times.

Deloitte's "2024 Manufacturing Industry Outlook" report reveals that 45% of manufacturing decision-makers plan to enhance operational efficiency by investing in the Internet of Things (IoT). This technology facilitates the connection between products, end-users, and manufacturers, enabling real-time monitoring. Manufacturers can leverage this data for various purposes, including product design and warranty repairs. However, integrating this data with existing systems remains a challenge.

Low-quality data: typical problems within a company

Supply chain

Inaccurate data can also lead to significant challenges within supply chains. Automating processes becomes exceedingly difficult when decisions are made based on unreliable location information. Outdated, incomplete, or incorrect inventory data can pose challenges in controlling inventory levels and planning orders effectively.

Incomplete or inconsistent product data further exacerbates the issue, making it challenging to identify and track products accurately. This can result in delays in deliveries and difficulties in meeting regulatory requirements. Given the growing emphasis on environmental sustainability, recycling, and circular economy principles, the importance of product tracking is heightened.

Low-quality data:
typical problems within a company

Management

Data quality can significantly impact the ability to achieve long-term goals.
Poor data quality can:

- adversely affect the ability to adapt and respond quickly to new trends and market conditions,
- increase the difficulty of complying with key privacy and data protection regulations, such as GDPR, HIPAA, and CCPA,
- hinder the use of predictive analytics for corporate data, potentially leading to riskier decision-making,
- make it impossible to predict when machines are likely to fail, hindering proactive maintenance efforts to reduce downtime and so enhance productivity.

Challenges of data hygiene

Research indicates widespread issues with data hygiene. The Salesforce portal found that 73% of leaders believe reliable data reduces uncertainty and supports sound decision-making.

However, many companies struggle to maintain the quality of their data. A study published by Harvard Business Review revealed that, on average, 47% of new data records contain at least one critical error impacting jobs. Even a 3% error rate is considered „acceptable" by the lowest standard.

This low data quality poses challenges when integrating or automating data. According to the publication „The Costs of Poor Data Quality," 88% of data integration projects fail or exceed budget expectations due to poor data quality.

Wysoka
jakość danych

Błedy
krytyczne

Nowe
rekordy

What makes data hygiene difficult?

Growing variety of data sources

Companies used to rely solely on data generated by their own business systems, resulting in common data “silos” where sales, production, and marketing data were stored separately.

However, in business today, data is sourced from a wide array of channels, including the Internet, the Internet of Things, scientific publications, and experimental results. The proliferation of data sources makes it increasingly challenging to control the quality of data and ensure they remain unaltered.

Each additional system integrated into the data processing framework heightens the risk of losing the value of these data, as different sources generate diverse data types. This risk is particularly pronounced with unstructured data, which accounts for an estimated 80% of all data worldwide. Processing such data poses inherent risks, as it may inadvertently be truncated or altered.

Growing volumes of data

We’re living and operating in the era of big data, where the volume of data continues to surge. Since 1970, data has been doubling every three years. With this exponential growth, collecting, cleaning, integrating, and ensuring reasonably high data quality has become increasingly challenging.

When dealing with predominantly unstructured data, processing times escalate even further. Unstructured data necessitates transformation into structured or partially structured data, which, although feasible, compromises data processing quality.

Accelerating the speed of data usage

„Real-time data” has become a buzzword over the past five years. The more data you generate, the faster you need to process it. However, increasing processing speed risks system overload.

Data behaves like liquid in a pipe: the faster it flows, the higher the chance of the pipe bursting. The only way to manage the increasing volume is to enlarge the “data pipe,” processing it faster to match its flow rate. This task is challenging and best approached in collaboration with experienced experts.

Nevertheless, real-time data processing is still a relatively new field. Consequently, it’s possible that some important data may escape processing while irrelevant ones are successfully processed. That’s why regular monitoring of data is crucial, along with maintaining their hygiene.

Lack of clear data quality standards

Product quality standards have been in existence since 1987 when the International Organization for Standardization (ISO) published the ISO 9000 standard. However, official data quality standards only emerged in 2011 with the introduction of ISO 8000. These standards are still evolving, and there is no singular universal standard established yet. Each company must develop its own set of rules. It is advisable to seek assistance from experienced consultants and data analysts to navigate this complex terrain effectively.

Best practices in data hygiene

Work on data quality standards is ongoing, but data hygiene best practices already exist. It’s worth adhering to these practices to attain high data quality and consciously implement data hygiene in your company.

Compliance

To comply with data-related regulations, determine the purpose and principles of data collection, especially if the data comes from consumers. Set rules for storing and deleting data. Use retention schedules, i.e., set time limits for storing data in your system.

To uphold data hygiene, it's imperative to:

1. Identify the data being stored,
2. Understand the purpose behind storing the data,
3. Determine the specific criteria and timelines for data deletion.

It's also worth implementing good data compliance practices, e.g. keep a record of all data protection measures and audit procedures — the more detailed, the better.

Data management

Data management encompasses a range of components, including processes, roles, policies, standards, and metrics. These elements work together to harness information effectively in order to achieve business objectives.

In data management, you determine:

1. The responsible parties who take action,
2. The specific data upon which actions are based,
3. The situations or conditions that trigger these actions,
4. The methods or approaches used to carry out these actions.

Management supports the organization's data quality assurance.

Automation of data management

In conclusion, data hygiene entails automating processes to uphold data quality, primarily through automated data updates conducted as frequently as possible. This approach ensures that the data remains current and accurate over time.

Data cleaning systems are designed to sift through large volumes of data efficiently. They utilize algorithms to detect anomalies or outliers that may arise from human error. Additionally, these systems can scan databases to identify and eliminate duplicate records, a particularly crucial function in manufacturing companies where employees frequently utilize interfaces for manual data entry. Automating the deletion of sensitive personal data that is no longer required is a worthwhile step.

How to assess data quality?

Data quality depends on many elements. High-quality data is:

1. **Up-to-date**
   created, managed, and readily available when needed.

2. **Concise**
   avoiding redundancy.

3. **Consistent**
   there are no conflicts in information within or across systems.

4. **Accurate**
   correct, precise, and kept current.

5. **Complete**
   including all necessary and available data.

6. **Compliant**
   stored in a suitable, standardized format.

7. **Correct**
   originating from known, reliable sources, ensuring authenticity.

If your data meets these criteria, you'll have the most reliable information in your systems. This quality data serves as a foundation for enhancing customer service, improving user experience, and driving business performance to new heights.

CASE STUDIES

It’s time to delve into examples from our practice. Each case illustrates how we tackled specific business challenges, outlining how we analyzed the client’s situation and devised a solution. Data stood at the forefront of each project, guiding our approach and enabling successful outcomes.

1. Data scraping in a training company, or how we accelerated the sales process by automating data retrieval

Challenge

A training company has to stay informed about available funds from the National Training Fund, which are published on the websites of District Labor Offices (PUP) throughout the country.

The company faces the demanding task of reviewing hundreds of websites daily. Given the competitive nature of securing funding, time and attentiveness are paramount, as the order of application can significantly impact success.

The individual responsible for manual web scraping must review 340 pages of PUPs daily. Subsequently, they input information regarding application calls, including available budget and document submission deadlines, into an Excel spreadsheet. Only after this data is compiled can the company send out emails to its clients.

Data acquisition is further complicated by the heterogeneous structure of PUP websites. Lacking a unified standard, these sites often present information across different tabs in a non-intuitive manner. This lack of uniformity increases the complexity and time required to gather the necessary data.

Manual work consumes resources and poses significant challenges in situations such as vacations or emergencies. The absence of a skilled individual proficient in navigating PUP websites can detrimentally impact a training company's sales efforts.

Solution

We proposed that the company automate data collection through web scraping. We developed a tool capable of automatically retrieving information on training funding.

In stage one, we developed an advanced application (bot) using Java with Spring Boot. This bot was designed to browse job centers' websites autonomously, eliminating the need for humans to engage in tedious and repetitive tasks.

In stage two, our focus shifted to detecting changes on web pages. We configured the bot to autonomously monitor each page and identify any updates or modifications, ensuring that it promptly captured new information as it became available.

Automated web scraping enabled swift access to current data regarding funds available from the National Training Fund.

Benefits

- Automation of a tedious process
- Efficient utilization of hard-to-reach and dispersed data
- Data standardization facilitating comparison of information from diverse sources
- Liberation of resources within the company
- Acceleration of the sales process
- Potential for expansion to related processes

2. Custom database migration, or how we eliminated a bottleneck in a manufacturing company

Challenge

A company in the metals industry sought to upgrade its software version to leverage new and enticing features. The third-party solution provider tasked us with assisting in custom migration of two databases.

Our analysis revealed that at the client's site, one production team had been working on a commercial Microsoft SQL database, while another had created an independent database in MySQL over several years. Consequently, the data within the CNC machine tool manufacturing documentation management program (CIMCO) was stored in two separate databases. This lack of integration between the databases led to inaccuracies in machine availability status and posed challenges in utilizing the machines effectively for daily production needs.

It became evident that two essential departments in the process were unable to access the data they had generated, despite the frequent need for it in various projects. The specific challenge we faced was to eliminate the need to work with separate databases. Our objective was to eradicate information "silos" within the company and mitigate the expenses associated with upgrading versions of both systems.

Solution

We devised a migration strategy that ensured the client did not lose any of the documentation amassed over the years. Additionally, we successfully preserved the interdependencies between the databases throughout the migration process.

Indeed, both data sources were relational databases, featuring explicit relationships such as indexes and foreign keys. However, they also contained implicit relationships, which were more challenging to capture. Given the lack of detailed documentation from the client, we analyzed the technical aspects of the databases and the business data that required transfer.

We transitioned the MySQL database to MS SQL architecture and executed the migration utilizing the Microsoft SQL Server Migration Assistant software. To accomplish this, we leveraged proprietary software developed by our team explicitly for this project.

Benefits

- Both teams work on a single database.
- All users, projects, and documentation are now in one consistent database.
- The company's accumulated knowledge is still available.
- We have eliminated the need for administrative support of two separate databases.
- There is no longer a need to maintain additional server space and backups.
- The migration has reduced the cost of upgrading to a new version of the software.

The true utilization of data commences once it has undergone cleansing. The intrinsic value of data resides in its organization and accuracy. Upholding data hygiene unlocks the potential to revolutionize your business operations and fuel innovation.

Raise awareness among employees about the importance of the data they introduce. Implement established protocols and utilize cutting-edge tools to safeguard the confidentiality, integrity, and availability of data. Recognize that this forms the bedrock for business expansion in the digital era.

Your datasets hold significant value when they unlock new opportunities for action and enhance the quality of your business operations. Thus, ensuring data hygiene is not merely a technical concern but also a strategic imperative.

If you're aiming to maximize the potential of your data or require assistance from our analysts and implementers for your data management project, we encourage you to reach out to us.

Author:

Jacek Nowak
CEO at CCA Europe

CCA Europe.pl

CCA Europe.pl offers IT solutions tailored for the financial sector and industry. We specialize in managing intricate central banking and payment systems development projects. Additionally, we provide software solutions for banks and IT tools designed specifically for manufacturing companies. Our expertise extends to database integration, enabling businesses to optimize their data utilization. We employ an agile approach to testing, leveraging proprietary tools and solutions. Moreover, we provide independent pre-implementation business and technical analysis services to ensure the success of your projects. Visit us at CCA Europe.pl | Rozwiązania IT dla bankowości i przemysłu