Paper SAS3247-2019

# Data Governance: Harder, Better, Faster, Stronger

Vincent Rejany, SAS Institute Inc

## ABSTRACT

All the major trends call for advanced control and accountability toward the use of data: from the migration to cloud applications and data warehouses, to the deployment of big data environments, the democratization of analytics and artificial intelligence or the increasing requirements related to data privacy and data protection.

Data Governance turned from being a nice to have to a must have, with an ever-expanding scope to address; gone are the days of marketing databases, some ERP processes, or specific regulations such as Solvency 2 or BCBS239 being the limits. Most of the organizations came through strong challenges aligning people and processes, trying to sustain the governance effort the time, and progressively this dream of "Enterprise Data Governance" is fading.

Organizations are now looking at more surgical initiatives to take control of their data lakes and at ensuring that their analytical processes are fed with reliable information and their data privacy policies enforced, and they want results immediately.

In this paper we will look at why and how Data Governance can be smarter and inspire trust, and how it can be automated, by relying on analytics and artificial intelligence.

## INTRODUCTION

Harder, Better, Faster, Stronger! Even after 15 years, this song from electro music group Daft Punk keeps on being avant-garde. Not only do the French electronic music duo, formed in 1993, always present themselves as robots, but the tracks from their album Discovery, released in 2001, sound present and could easily rank in the top 10 of the charts. Why an analogy with Data Governance? Because Data Governance has never been so that "Harder" to execute considering the advent of regulations requiring data management excellence and data protection assurance, the explosion of data volumes, applications and security breaches, and moreover the movement to the cloud. Essentially, about data, organizations don't know what they know, and if they do, they don't know where to find it. Therefore, for organizations to face this increasing challenge, they need to move Better, Faster, Stronger. But how?

First, we need to demystify the concept of Data Governance and the challenges it faces today. Next, we will look at how Data Governance can be executed in a more efficient way with little help from artificial intelligence, so that it can inspire trust and become a real strong awareness, the concern of everyone and not only IT people.

## DEMISTTIFYING DATA GOVERNANCE

Most companies would agree that today, data is the very lifeblood of their business, that digital transformation is holy grail for not being disrupted. We often hear then "data" is a corporate asset like money, employees, buildings, and machines. However, "Accounting", "Human Resources" and "Procurement" do not require any definition. They speak by themselves, and so should the poor ugly duckling "Data Governance". For years, it has been a strategy struggling for recognition and acceptance. Most of the time Data Governance has been treated as a technology project rather than a business transformation imperative. At the end of the day, business users just don't understand the value of data governance and what it really involves: a cultural change.

Wikipedia gives a very formal definition of Data Governance as "*a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data. The key focus areas of data governance include availability, usability, consistency, data integrity, and data security and includes establishing processes to ensure effective data management throughout the enterprise such as accountability for the adverse effects of poor data quality and ensuring that the data that an enterprise has can be used by the entire organization."*

This definition makes data governance resonating like an esoteric concept. However, this is probably one first mistake and a reason why it has been generating little adoption from business people. Quoting Albert Einstein, "if you can't explain it simply you don't understand it well enough". So, let's try to simplify for bringing clarity in the confusion.

## PEOPLE, PROCESS, AND TECHNOLOGY?

In many articles and papers around Data Governance you will read about the famous data governance recipe: People, Process, and Technology.

| **People** | **Process** | **Technology** |
|---|---|---|
| A team responsible for the data assets in the organization, with data owners accountable for the quality of the data and the support of data quality activities and initiatives company-wide. | Processes include data processes, meaning how and where data will be stored, moved, changed, accessed, and secured. But also encompass control, audit, monitoring, and escalation processes. | Technology is useless without People and Processes but is essential for moving from reactive mode to a governed mode and allowing to gain efficiency, minimize risk and increase revenue. |

**Table 1. Data Governance Triumvirate**

It is true; Data Governance does rely on people, processes, and technologies. However, are these elements not today "the Golden Triangle" of almost all business activities and operations in organizations, which pretend to be "data driven"? It won't be a shortcut to see here a nice syllogism or transitive logic:

> *"If Data Governance encompasses the people, process, and technology that are required to ensure that data is fit for its intended purpose, and, business activities are driven by data, then Data Governance drives business activities."*

This definition allows to put data governance where it should reside, within business and to highlight that is now a critical component as people and money are.

## TRUST AND DATA DEMOCRATIZATION

Therefore, if we follow the line of reasoning, "Data Governance" is about managing data efficiently for driving business activities. But what does "managing data efficiently" mean? When can we say that one data management activity is achieved in a quick and organized way and delivers the expected results? When it allows to increase revenue, reduce costs or minimize risk? Yes, if metrics and targets have been defined? It helps but depends who does define them and how they are measured.

And by challenging this former question, we can already feel the concern here. Before all Data Governance is about infusing "Trust" for creating the conditions for efficiency and generating value. Trust is essential in the definition of Data Governance, and it does not only concern internal stakeholders, but also the individuals from whom personal data is being processed. We will come back to this notion later.

> *"Data Governance aims at showing that data and data management processes are trustworthy and credible."*

Other key business activities like Accounting or Human Resources do inspire trust, because of the regulations, the policies, processes and certifications they do rely on. They do involve frameworks, laws, binding corporate rules, and methodologies. Moreover, managing a budget or resources is not only owned by one single team, it is shared, and each manager is accountable for his team, expenses or the potential P&L he oversees.

The same transformation should apply to data, and data governance. For several decades now, business users have lived under the misapprehension that their IT department owns the data of their organization. Data Governance can no longer be measured only from an IT perspective and not from a business one. IT owns and is responsible for the infrastructure. The business is responsible of what that data is, how and why it is held. The demand and consumption of data is increasing, everybody wants to access, and analyze data without requiring outside help. Business people and Data Scientists ask for transparency and do want "Data Democratization" instead of IT dictatorship or data scientist's aristocracy or even in most cases complete anarchy. This call for democracy requires changes in the way data governance is being done.

## DATA GOVERNANCE IN PRACTICE

In practice, how data governance is executed apply. Well, there are for sure multiple approaches and strategies, from the most complex to the most pragmatic, combining the definition of an organization, roles, and processes. From a methodology perspective, we could summarize it in four steps:

1. Define how data should be organized
2. Assess how data is disorganized
3. Enforce data governance policies and rules
4. Adjust and Improve the data governance framework

In fact, it is like a classic PDCA (Plan, Do, Check, Act) for the control and continuous improvement of processes, that is also often use in data quality management. On a technology perspective, data governance must rely on a variety of products, which support these macro steps. Using excel spreadsheets could work for small projects, but it can't be sustainable at an enterprise level.

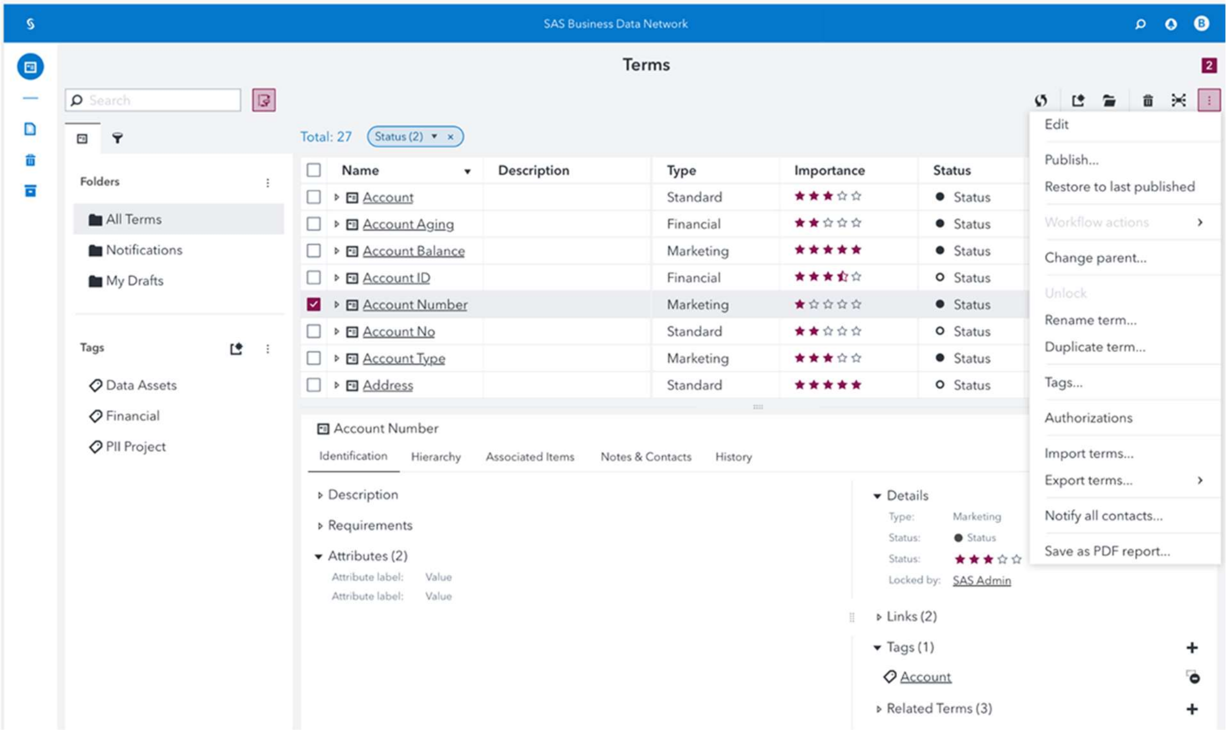
### Collect data assets (metadata) into a Data Catalog

A repository of metadata centralizing information about data sources, schemas, tables, columns extended. A data catalog includes technical attributes (name, description, format, length …) and generated knowledge such as data profiling metrics, and privacy information (descriptive measures, frequency and pattern distributions, content identification …)

### Describe business assets perspective in Business Glossaries

Business Glossaries help organizations to reach agreement between all stakeholders on their Business Assets (for example, terms) and how they relate to data assets (for example, database tables) and technology assets (for example, ETL mappings), known as technical assets. A business glossary can be used as a single-entry point for all data consumers to better understand and govern their data asset through the definition and the maintenance of business terms. Business terms can be organized through hierarchies and relationships and can be linked to different roles such as data or business owner or data steward. Different types of terms can be defined according to the information that needs to be documented. Therefore, it contains the Language of the Business, independent of technology used to:

- Define authoritative meaning
- Increase and share understanding throughout the enterprise
- Establish responsibility, accountability, and traceability
- Represent business hierarchies
- Document business descriptions, examples, requirements, valid values,
- Find relevant information assets



| ☐ | Name ▾ | Description | Type | Importance | Status |
|---|---|---|---|---|---|
| ☐ | Account | | Standard | ★★★☆☆ | ● Status |
| ☐ | Account Aging | | Financial | ★★☆☆☆ | ● Status |
| ☐ | Account Balance | | Marketing | ★★★★★ | ● Status |
| ☐ | Account ID | | Financial | ★★★⯪☆ | ○ Status |
| ☑ | Account Number | | Marketing | ★☆☆☆☆ | ● Status |
| ☐ | Account No | | Standard | ★★☆☆☆ | ○ Status |
| ☐ | Account Type | | Marketing | ★★★☆☆ | ● Status |
| ☐ | Address | | Standard | ★★★★★ | ○ Status |

Account Number

Identification | Hierarchy | Associated Items | Notes & Contacts | History

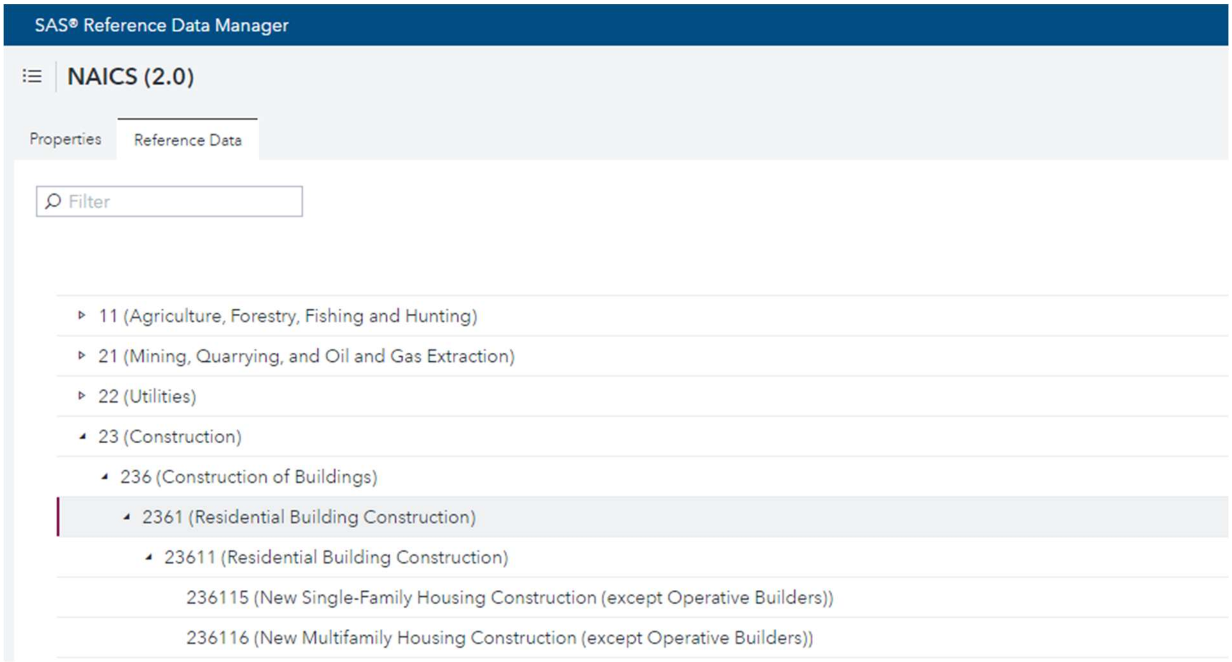
- Description
- Requirements
- Attributes (2)
  - Attribute label: Value
  - Attribute label: Value

- Details
  - Type: Marketing
  - Status: ● Status
  - Status: ★★★☆☆
  - Locked by: SAS Admin
- Links (2)
- Tags (1)
  - Account
- Related Terms (3)

**Display 1: SAS® Business Data Network Main View**

## Centralize Reference Data

Every organization has some common set of data that are used by many different business processes to provide a standard "library" of terms within various applications. This reference data usually comes from outside the organization (though this is not always the case) and changes infrequently. A few good examples of reference data might be, a list of all countries and their ISO country codes, the "official" list of sellable products, organizational hierarchies, store locations by city and state, approved abbreviations for medical terms. Reference data do describe the acceptable values for business terms.



SAS® Reference Data Manager

NAICS (2.0)

Properties | Reference Data

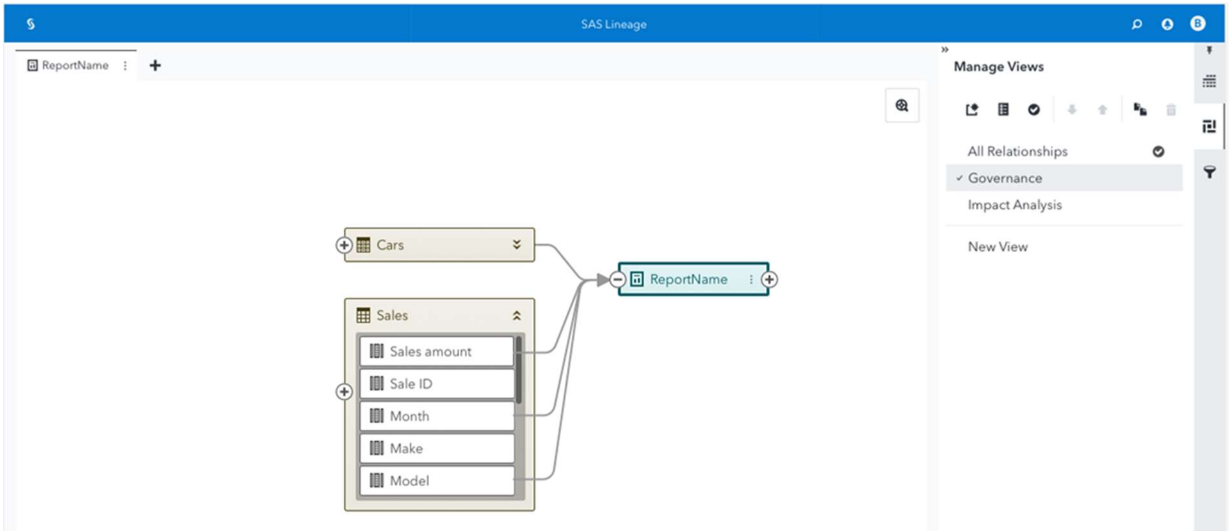
Filter

- 11 (Agriculture, Forestry, Fishing and Hunting)
- 21 (Mining, Quarrying, and Oil and Gas Extraction)
- 22 (Utilities)
- 23 (Construction)
  - 236 (Construction of Buildings)
    - 2361 (Residential Building Construction)
      - 23611 (Residential Building Construction)
        - 236115 (New Single-Family Housing Construction (except Operative Builders))
        - 236116 (New Multifamily Housing Construction (except Operative Builders))

**Display 2: SAS® Reference Data Manager**

## Search Asset Catalog and Browse Lineage

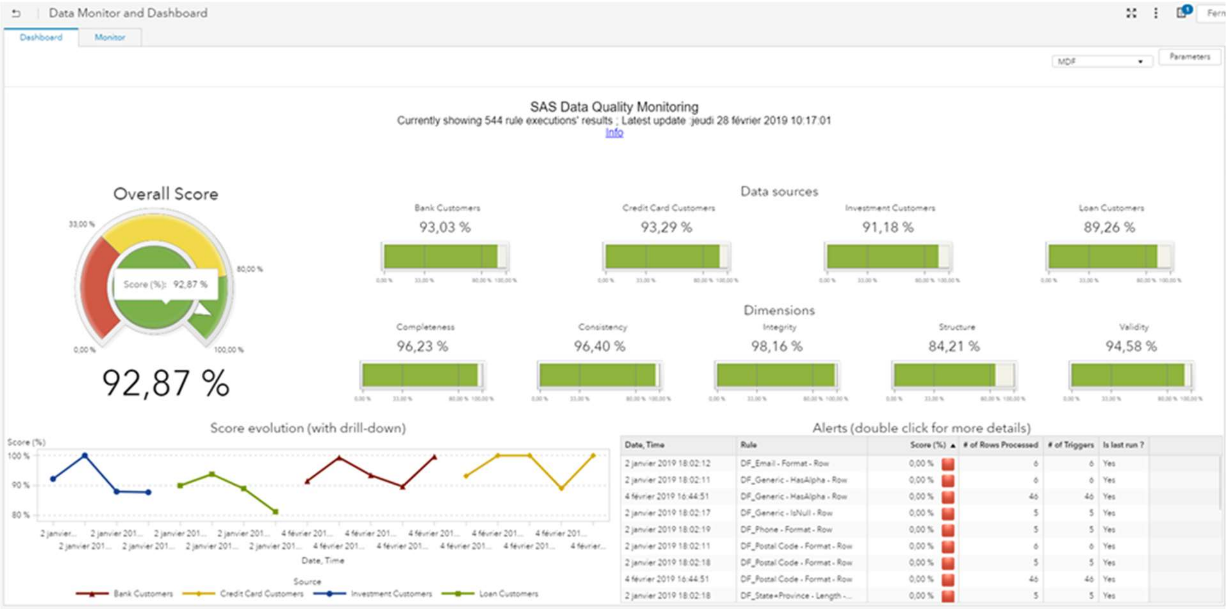
An Asset Catalog embeds the data catalog, business glossaries, reference data and adds all the other metadata, such as BI reports, ETL processes, analytical models, data preparation jobs as well as metadata from third-party vendor platforms. Lineage supports the management and analysis of object and metadata relationships, including dependencies and life cycle. This management and analysis process reveal where data comes from, how it is transformed, and where it is going, and all the steps in between.



**Display 3: SAS® Lineage**

## Enforce Data Quality and Business Rules and Remediate issues

Data quality and business rules need to be design and applied on data assets. Issues identified by these rules are to send to a remediation process that provides means to identify, review, and correct the problem data before it reaches the downstream systems. Data monitoring results are presented into data governance dashboard to present system by system and dimensions by dimensions how much data quality and governance are improving. They can also be added to the data catalog layer as additional knowledge on data.



SAS Data Quality Monitoring

Currently showing 544 rule executions' results ; Latest update :jeudi 28 février 2019 10:17:01

Info

Alerts (double click for more details)

| Date, Time | Rule | Score (%) ▲ | # of Rows Processed | # of Triggers | Is last run ? |
|---|---|---|---|---|---|
| 2 janvier 2019 18:02:12 | DF_Email - Format - Row | 0,00 % 🟥 | 6 | 6 | Yes |
| 2 janvier 2019 18:02:11 | DF_Generic - HasAlpha - Row | 0,00 % 🟥 | 6 | 6 | Yes |
| 4 février 2019 16:44:51 | DF_Generic - HasAlpha - Row | 0,00 % 🟥 | 46 | 46 | Yes |
| 2 janvier 2019 18:02:17 | DF_Generic - IsNull - Row | 0,00 % 🟥 | 5 | 5 | Yes |
| 2 janvier 2019 18:02:19 | DF_Phone - Format - Row | 0,00 % 🟥 | 5 | 5 | Yes |
| 2 janvier 2019 18:02:11 | DF_Postal Code - Format - Row | 0,00 % 🟥 | 6 | 6 | Yes |
| 2 janvier 2019 18:02:18 | DF_Postal Code - Format - Row | 0,00 % 🟥 | 5 | 5 | Yes |
| 4 février 2019 16:44:51 | DF_Postal Code - Format - Row | 0,00 % 🟥 | 46 | 46 | Yes |
| 2 janvier 2019 18:02:18 | DF_State+Province - Length -... | 0,00 % 🟥 | 5 | 5 | Yes |

**Display 4: SAS® Monitor and Dashboard Governance Report**

## HARDER DATA GOVERNANCE

Data Governance is at a cross road: high demand of trustworthy data, data privacy regulations and individuals asking for ethical, secured, transparent, and lawfulness data processes, combined with increasing volumes and use. Data is big, demand is big, opportunities, good or bad, are big and so are risks. Sounds like a very good cocktail for change, a revolution, or a crisis. Let’s summarize these challenges in two main groups, which are chasing after each other:

- The rise of the data democracy
- The pressure from regulatory compliance

### THE RISE OF THE DATA DEMOCRACY

Data does not stop to get bigger. Big data today is small data yesterday, according to Forbes, 2.5 quintillion bytes of data are created each day at our current pace. It is there and everywhere, immediately accessible. Companies have quickly accrued massive amounts of data, adopted big data environments to store it and now looking at how to drive their digital transformation. However, the emerging and explosion of Hadoop and cloud platforms and new processing engines: in database, in-Hadoop, in-motion, in-containers combined with the power of analytics solutions are hiding the complexity and the disorder. This widespread data disorder is the most significant obstacle preventing organizations from realizing the full potential of their data assets today. Digital transformation cannot be successful without an emphasis on how data is collected, processed, controlled, and secured. While insights might be buried within all that raw data; if no one knows where it came from, how to find it, what it means or if they can trust it, it will remain untapped and untouched.

Margaret Rouse from Techtarget.com gives the following definition: "Data democratization is the ability for information in a digital format to be accessible to the average end user. The goal of data democratization is to allow non-specialists to be able to gather and analyze data without requiring outside help." True that computer sciences are no longer IT people only skills. New generations of workers, raised with computers, are entering the market with good and even advanced BI and analytics knowledge, and expectations as they cannot imagine being successful at their job without the access and the consumption of data they need. According to Laura Hellis in her article, "Building Data Democracy", "therefore data has become the new language of business management […] to truly excel at your job you need to dig into the "Why" behind the "What" questions.".

Unfortunately, most organizations do data governance in an ad hoc or firefighting manner across different parts of the business and most of the time only within IT. In the worst cases there is no governance program in place or only at a project-based level. In the best scenarios' governance has been enforced thanks to compliance with regulations involving data management best practices, but again not all business processes, departments and geographies will be equally supported. Enterprise Data Governance is still a sweet dream for many organizations.

What do business users ask? To get the power, access to trustworthy data and to have more control over their analytics work. The challenge is about how to set up the infrastructure, the architecture and the functional organization that will facilitate and serve data analysis, through simple tools, simple presentation of data, high and verifiable quality data, in other words "Data Governance". However, quoting a famous commercial, "Power is nothing without control" and the democratization of data also comes with watchdogs and standards.

## REGULATORY COMPLIANCE PRESSURE

Over last years, many regulations like Solvency 2, SOX, CIA, BCBS239, MiFID, CCAR, Transparency Act, IDMP, HPAA, EU GDPR and recently CCPA called for better data management, either by the production of specific reports, or by directly requiring dedicated actions. Regulators require organizations to control what data they use to make business decisions, to pro-actively prevent and detect data breaches or fraud, and to manage financial risks. For example, BCBS 239, Basel Committee on Banking Supervision's standard number 239, was one of the first regulation including principles describing how a bank's risk data aggregation capabilities and risk reporting practices should be subject to strong governance.

| | |
|---|---|
| Principle 3<br>Accuracy and Integrity | A bank should be able to generate accurate and reliable risk data to meet normal and stress/crisis reporting accuracy requirements. Data should be aggregated on a largely automated basis so as to minimize the probability of errors. |
| Principle 4<br>Completeness | A bank should be able to capture and aggregate all material risk data across the banking group. Data should be available by business line, legal entity, asset type, industry, region and other groupings, as relevant for the risk in question, that permit identifying and reporting risk exposures, concentrations and emerging risks. |
| Principle 5<br>Timeliness | A bank should be able to generate aggregate and up-to-date risk data in a timely manner while also meeting the principles relating to accuracy and integrity, completeness and adaptability. The precise timing will depend upon the nature and potential volatility of the risk being measured as well as its criticality to the overall risk profile of the bank. |

|  |  |
|---|---|
|  | The precise timing will also depend on the bank-specific frequency requirements for risk management reporting, under both normal and stress/crisis situations, set based on the characteristics and overall risk profile of the bank. |

**Table 2. BCBS 239 Data Principles**

More recently, the EU General Data Protection Regulation has been a massive game changer in Europe and other countries impacted as well. For the first time, one regulation is impacting all industries as 99% of them do process personal data.

EU GDPR requires businesses to protect the personal data and privacy of EU citizens for transactions that occur within EU member states. And non-compliance could cost companies dearly. The GDPR allows for steep penalties of up to €20 million or 4 percent of global annual turnover, whichever is higher, for non-compliance. It avoids building a use case for launching a Data Governance program.

The aim of data protection regulations such as EU GDPR is to change behaviors and mindsets. Taking that perspective, the accountability principle (in Article 5 of the EU GDPR) makes the data controller to be the one responsible for demonstrating compliance with these EU GDPR principles:

- Lawfulness, fairness, and transparency must exist in processes that manage personal data.
- Limitation of purpose. Personal data must be collected for specified, explicit, and legitimate purposes.
- Data minimization. There should be no reason to use more data than necessary for the defined purpose.
- Accuracy. Data quality must be ensured, and personal data be kept up-to-date.
- Storage limitation. Personal data must be processed for no longer than is necessary.
- Integrity and confidentiality. Appropriate security measures must be taken.

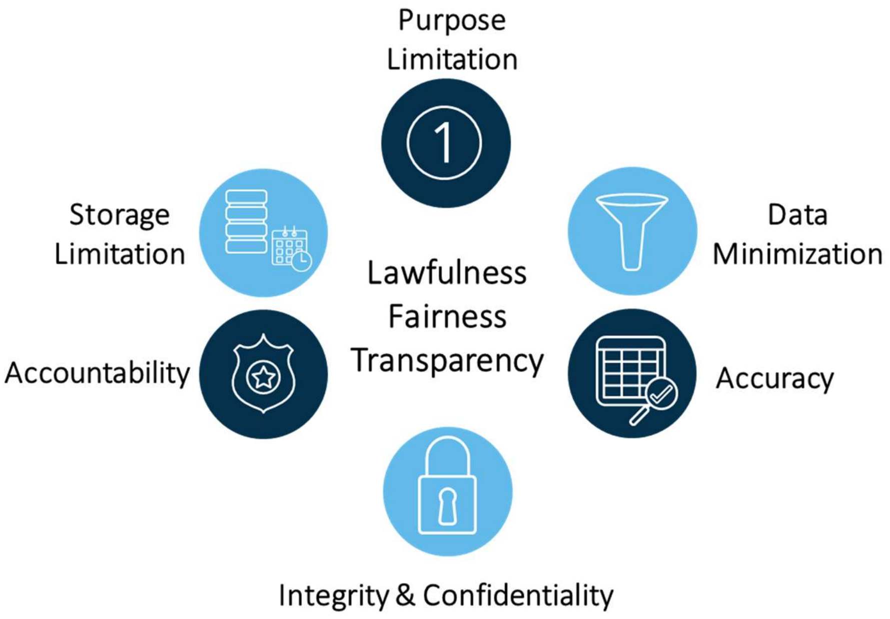


**Figure 1. EU GDPR Principles**

The main action to be taken for demonstrating “Accountability” is to document internally all your processing activities, and to make this documentation available to supervisory authorities upon request. This “record of processing activities” is required by EU GDPR Article 30 and will facilitate the compliance with the other principles. It implies to assess which type of data elements are used, what for, where there are stored and for how long.

Data retention is a big and complex issue for companies. Determining retention strategy and policies, how to implement those policies, and how to create automated processes with rules that depend on what country you're doing business with is a tedious work. The rules are different depending on what country you're doing business in and the type of data being process and the purpose of the processing. It's a very complex area of data management and governance.

Companies could have also to carry out data protection impact assessments (DPIAs) when data processes could represent a high risk to individuals’ rights and freedoms, particularly when new technologies are involved. The DPIA is required by Article 35 of the EU GDPR and contains information about how a new or modified application might affect the privacy of personal information processed by or stored within the application.

Considering the first challenge related to data volumes and data democratization we described earlier, organization having hundreds of systems, data assets, and processing activities, and thousands of personal data types to review daily, weekly, or monthly, describing these items is a significant effort but maintaining an up-to-date view of them is even more time-consuming and is prone to errors. In terms of Data Governance, the typical manual or semi-automatic steps no longer stand a chance when facing EU GDPR requirements.

## BETTER DATA GOVERNANCE

Considering the challenges mentioned previously and IT budgets growing slowly, we can ask ourselves how organizations can:

- Allocate their resources in the most efficient way to support their existing businesses,
- Investigate new data opportunities before being disrupted,
- Comply with regulations efficiently and economically

Classic data management approach cannot just scale. Considering the volume data, it is like crossing the universe, it is just too big, it will go to slow, it will be to error prone, it will cost too much, and it will be probably useless. Not all the answer can be brought by technology, it must be mixed of cultural changes supported by technology as an enabler. Collaboration and automation are the fundamental for unleashing data governance.

As it is not that difficult for machines to do better and quicker than human, automation is the only option for scaling and facing the variety, the volume and the velocity, boosting productivity, supporting data governance principles, detecting new opportunities. With automation business self-service enablement increases through driven and sustainable actions.

Collaboration is also critical as data is no longer the concern of few people in the organization. It must become a shared responsibility for generating trust and acceptance. Automation does not mean that there is no longer anything to do, it means that data stewards, data scientists, data analysts can focus on more value-added activities and collaborate on making data democracy real and sustainable.

## THE PHYSIOLOGICAL NEED OF TRUST

We have seen earlier that the notion of trust is essential for creating the conditions for efficiency and generating value. And it is obvious that we usually trust:

- what we can understand, when knowledge is shared and not esoteric,
- what relies on a robust process and when we have elements for checking the credibility,
- when roles and rules are clearly communicated,
- when communication is done frequently, and vulnerabilities ae not hidden,
- when excellence is recognized, and feedback is considered

In a recent Harvard Business Research journal, Paul J. Zak, Harvard researcher, Founding Director of the Center for Neuroeconomics Studies and Professor of Economics, Psychology and Management at Claremont Graduate University, and author of "The Trust Factor: The Science of Creating High Performing Companies," shared that there is a direct correlation between the amount of oxytocin a person's brain produces and the level of trust they feel in any given situation. The higher the oxytocin, the higher the empathy. The higher the empathy, the deeper the connection. Yes, you could suppose that this link is far-fetched but for creating a culture of trust in data we need to work on increasing end user's oxytocin. And Zak summarizes height strategies and some of them can be applied to data governance:

1. **Share information broadly**: If you've seen the film "The Big Short", about the 2007 housing market crash and sub primes crisis. it opens with a fake Mark Twain quotation mark "It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so." Zak mentioned that "only 40% of employees report that they are well-informed about their company's goals, strategies, and tactics [..] which leads to chronic stress (a fear-based response), wish and exhibits the release of oxytocin and undermines teamwork." From a data governance perspective, it is about rising data awareness, training data users about the data ecosystem of the organization, the major internal sources of information, as well internal data policies and data privacy requirements. Data users must know and understand why they have or don't have access to certain data.
2. **Recognize excellence and Intentionally build relationships**: Some data users can be good at preparing data and building nice dashboards and reports. Through this process they can also discover inconsistencies and data quality issues and become real data governance guards and advocates. Recognizing immediately and publicly their contribution, or even through gamification activities. Comments, ratings, challenges, contests, surveys integrated in the data platform are often a great source of motivation and productivity. Users get to practice giving and receiving feedback in a way that is meaningful and timely.
3. **Facilitate whole-person growth**: Training has a well-known effect on engagement and retention of employees. So, let's unleash the data knowledge. There are so many ways today for facilitating self-training through virtual learning and MOOCs. Gamification of data governance can also help here with users leveling up in their data management awareness through a clear learning path and getting access to more data preparation capabilities.
4. **Give people greater control over how they work with data and enable job crafting**: We can suppose that giving more autonomy to data users could be tough for organizations, which try to standardize their processes and their software for managing and processing data. However, once data users have access to data they should be able

to act as "citizen data scientists" and build solutions that meet ever-changing needs. At the same time, through the creation of innovation labs or crowd sourced projects, new experimentation and approaches for manipulating and controlling data can be identified. They will bring their specific business knowledge and insights to bear on defining analytics needs.

The driving idea is that enabling business users to define their own information requirements, answer their own questions, and create their own tools will energize business processes and generate trust by engaging these users' innovation and business knowledge. This is foundation for data democracy.

## DEMOCRACY IN DATA GOVERNANCE

Data democracy has already started by bringing analytics closer to decision makers and business. This was partially addressed by self-service BI and got recently extended to data preparation and manipulation. The next step is Data Governance for giving access to business definitions, the underlying technical metadata, and to answer to data user when they ask the following questions:

| | | |
|---|---|---|
| Where do I find the information? | Who owns it? Who can I call to ask questions about it? | When I see it, how do I know what it means? |
| How does the information relate to information that exists in other systems? | When was the last time it was updated? | Where do I find out more about this data element? |
| Who changed it last? | Does the data conform to a corporate standard? | Is the data fit for internal or external reporting, including to regulatory bodies? |

**Table 3. Data Users' classic questions**

Looking at what has been done for BI and data preparation, infusing democracy into data governance requires a new generation of solutions made for business users combining three critical themes: Simplicity, Quality, and Collaboration.

**Figure 2. Democratized Data Governance**

1. **Simplicity**: It could sound a bit obvious but proposing simple and easy to use solutions is the most critical. It means, the ability to support zero-coding features, for example, click through or drag and drop design processes. Far is the time when users had to master SQL code for doing data management. Simplicity means also the ability to make easy as possible the integration with third-party solutions and business applications.

Moreover, data governance tools should be adaptable whatever is the size of the organization, the maturity, the industry and the volume of information being governed.

2. **Quality**: Quality is indissociable from the notion of governance as data governance aims at creating quality data systems that users can trust. The focus is here on the ability to support "Just in Time" requirements, like the famous 5 zeros from the Toyota Production System:

- 0 delay: Data must be easy to find and immediately available. The last thing you need is people running around trying to figure out which data assets can be trusted.
- 0 stock: Data redundancy should be constantly checked, and retention policies defined and apply. The less rogue or shadow data sets are maintained the more trustworthy is the data ecosystem.
- 0 paper: All the meaningful information about data should be centralized within data catalogs and business glossaries. Comments and feedback from users as well as the ability to raise alert, change, or data request is critical in terms of traceability.
- 0 default: Getting the best data quality is the top priority when on boarding new data sets or preparing data, to minimize risk.
- 0 weakening: Thanks to a regular, rigorous maintenance and review of business definitions and data quality controls, data governance processes are aligned with business expectations and priorities.

3. **Collaboration**: Equally important, Data Governance products you need to create the conditions for a supportive environment and to put the effort into fostering collaboration and creating enablement avenues. You don't want users to feel stuck and alone when they hit a roadblock. If they do, the adoption of data governance will suffer. Creating a supportive environment is part of the cultural adaptation that needs to happen. Supporting the ability to record business user feedback and expectations through discussions over data assets, comments, or ratings toward the quality of certain tables or metrics is a must have. Data governance must become fun and gamification principles could help in fostering the adoption and the change of behavior.

These three themes are essential. However, as for BI and data preparation, the democratization of data governance must be supported by one fundamental capability: "Automation". The automation of data management activities is the key for masking the underlying complexity and density of data environments, and for allowing to surface and prioritize the best actions to be taken.

## FASTER AND STRONGER DATA GOVERNANCE

How to execute data governance faster and stronger when there is so much work to do? One key element to help enforce data governance is a set of data services built on the technology platform to automate the data governance processes and to enforce data governance throughout the enterprise. The first rule for automation is of course to build once and use repeatedly. Data preparation jobs, ETL processes, data quality controls, should be always designed in a generic way, so they can be reused. The second rule is to embed and anticipate data governance as earlier as possible in all projects, meaning to do "Data Governance by design".

However, there should be also a set of built-in capabilities, which would facilitate the identification of data and the completion of classic data management tasks. A lot of knowledge is already there and only calling for being consolidated and used.

“Automation” would encompass all the disciplines allowing to generate insights on top of metadata, content, user activities and feedback, and data governance policies. It is not restricted to the use of artificial intelligence, but can also relies on classic integration and analytics, reporting or even deterministic, rule-based approaches (If then). “Automation” can also facilitate a personalization and smoothness of the user experience in the solution, as it streamlines actions that user does often, wants to do or has never been thinking about.

Let’s look at the types of information that could be analyzed and processed for supporting “automation”.

- Tables, text files, documents, and other data sets, with their respective metadata
- Metrics generated over data sets through data profiling and data discovery activities.
- Referenced data and controlled vocabularies such as list, hierarchies, look up tables.
- User's actions, behaviors from application logs
- User's feedback, comments, and rating
- Internal or external policies

This is list is not exhaustive and other set of information could also be used. There are as well multiple use cases for using and crossing these different sources for generating insight about which data management or governance actions to be executed. These use cases can be summarized into four categories: data discovery, suggestion and recommendation, anomaly detection, development, and administration.

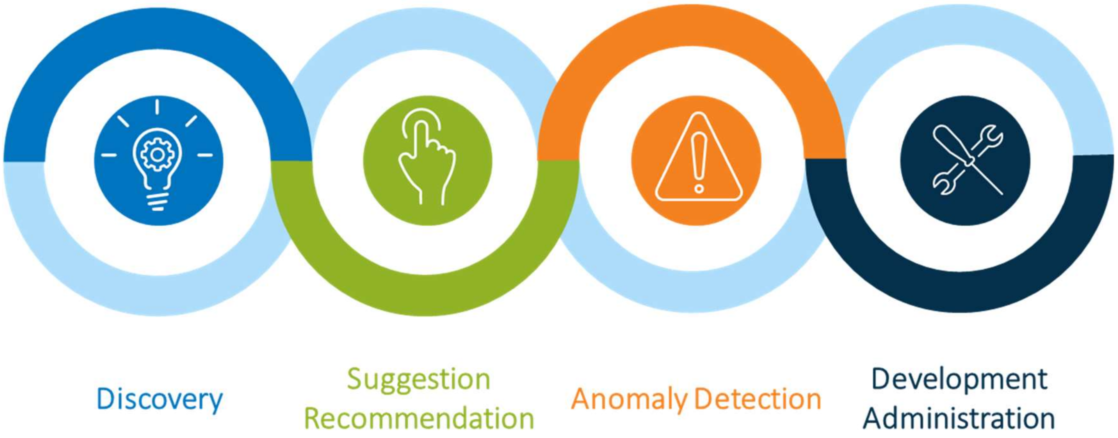
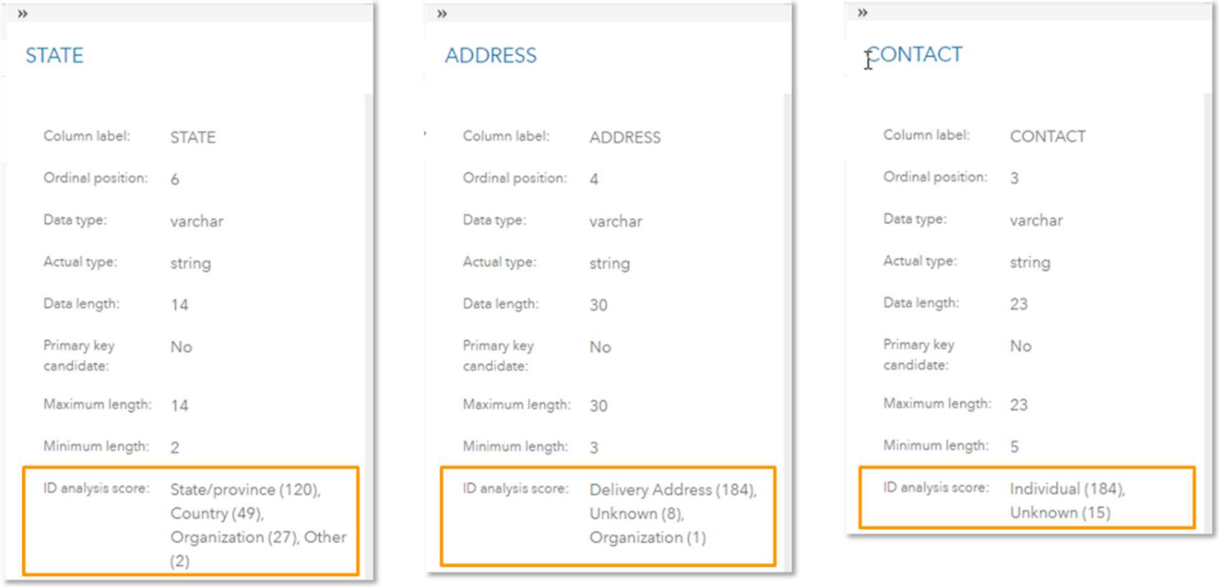


**Figure 3. Automated Data Governance Use cases**

## DATA DISCOVERY

Data Discovery covers a wide number of features. It aims at reviling either what is obvious from a human perspective but not documented, or what could be cumbersome to find out. For example, as a user I know that one data set contains one column "Email" as I can easily identify the column label and the pattern. However, if I had to review a whole data lake for assessing where I could find emails, and this information is neither in my data catalog, nor documented in my glossary then it becomes clearly a heavy and error prone exercise. It would be far more efficient if the machine could do it automatically and analyze all the data sources with a consistent logic.



STATE

Column label: STATE
Ordinal position: 6
Data type: varchar
Actual type: string
Data length: 14
Primary key candidate: No
Maximum length: 14
Minimum length: 2
ID analysis score: State/province (120), Country (49), Organization (27), Other (2)

ADDRESS

Column label: ADDRESS
Ordinal position: 4
Data type: varchar
Actual type: string
Data length: 30
Primary key candidate: No
Maximum length: 30
Minimum length: 3
ID analysis score: Delivery Address (184), Unknown (8), Organization (1)

CONTACT

Column label: CONTACT
Ordinal position: 3
Data type: varchar
Actual type: string
Data length: 23
Primary key candidate: No
Maximum length: 23
Minimum length: 5
ID analysis score: Individual (184), Unknown (15)

**Display 5: SAS® Data Preparation - Data Explorer - Data Profiling & Identification Analysis**

SAS® Data Preparation - Data Explorer supports already such functionality when profiling CAS tables. This functionality is supported by an identification analysis definition name "Field Content" available in the SAS® Quality Knowledge Base. Depending on the country and the language, different types of variables can be identified, such as: Individual, Organization, Delivery Address, Payment Card Number, Email, Postal Code, City, State/Province, Phone, Social Security number. This definition can be enhanced to support additional variables. It is a helpful exercise when running compliance programs with data privacy regulations like EU GDPR. Extensions of identification capabilities are currently under investigations through the use machine learning techniques, especially for extending the types covered and allowing users to build their own identification models. The identification of languages being used or sensitive data (such as data related to race and ethnic origin, religious or philosophical beliefs, political opinions, trade union memberships, health biometric or genetic information, sexual life, or preferences) or medical records information could benefit from this approach.

The classification of data sets is an interesting area too. From the identification of the columns it could be interesting to deduce and assign a domain name to data sets, i.e. "Contacts Records", "Invoice Data". Such tagging capability at the data set level facilitates the search, the access and the security of the data ecosystem.

Identification and classification capabilities are key features because they are the mandatory step in terms of data cataloging and data privacy principles enforcement. The automation of data masking processes like pseudonymization or the securing of sensitive data sets depend on the ability to identify content.

From an analytical perspective, there is also a wide number of use cases to identify whether the information quality of a data set is suitable for analysis. To support an accurate use and analysis of data, we need to ensure that all data needed for analysis are complete, or if not to propose the calculation of imputation values for missing values. The computation of analytical metrics such as "Skewness" and "Kurtosis" for interval variables combines with completeness rate helps in assessing if certain variables should be excluded or not. The extension of data profiling activities with such analysis are a foundation for the generation of insights, such as:

- Assessing the overall quality of one data set
- Scoring the readiness of one data set for analytics
- Clustering data sets with similar structure and data
- Identifying redundant data sets
- Applying data masking automatically

Through data profiling, identification of content and computation of advanced statistical metrics, data discovery is the cornerstone of automated data management and a pre-requisite for making suggestions and recommendations.

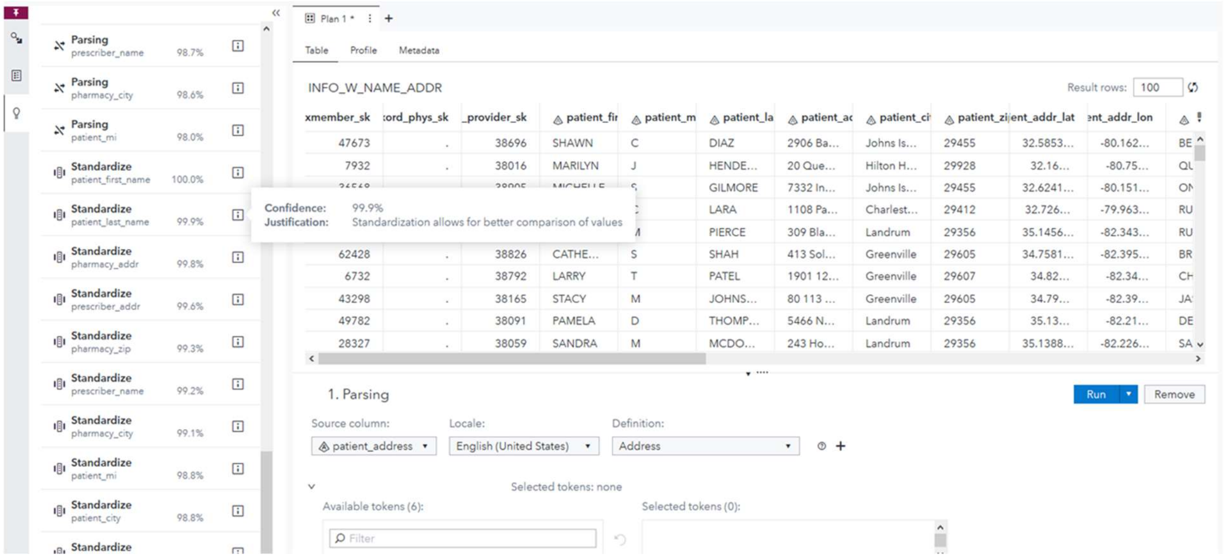
## SUGGESTION AND RECOMMENDATION

It could be quite complex to make a distinction between suggestions and recommendations. "Recommendation" could be considered as a benevolent information, an alert, a signal, that is not obvious or an outcome that could be predicted, in other words "This data set would need to be pseudonymized". Recommendations are usually proposed based on the analysis of past actions for recommending new ones. A recommendation is not always an action and it should inspire trust and confidence. One good illustration is the recommendation engines available in most retailer websites such as Amazon. According to research by McKinsey, a mind-boggling 35% of Amazon's sales and 75 percent of what users watch on Netflix come from product recommendations based on such algorithms. These statistics were reported in 2013 and it might be higher today.

Classic use cases would typically propose to combine one main data set with complementary or alternative data sets either because it makes sense or because many other users did it too, such as:

- Recommend "corporate" sponsored data sets or actions
- Recommend a well rated prepared version of the same data
- Recommend another table containing the same type of columns/records to use/substitute or union.
- Recommend a table for enriching the data with additional variables
- Include in the user interface a "Did you know?" widget

We would consider that "Suggestions" aim at proposing an action, like suggesting a next best action, that is: "Apply standardization on ZIP code". For example, SAS® Data Preparation will soon embed suggestions of data preparation steps.



Parsing prescriber_name 98.7%

Parsing pharmacy_city 98.6%

Parsing patient_mi 98.0%

Standardize patient_first_name 100.0%

Standardize patient_last_name 99.9%

Standardize pharmacy_addr 99.8%

Standardize prescriber_addr 99.6%

Standardize pharmacy_zip 99.3%

Standardize prescriber_name 99.2%

Standardize pharmacy_city 99.1%

Standardize patient_mi 98.8%

Standardize patient_city 98.8%

Standardize

Confidence: 99.9%
Justification: Standardization allows for better comparison of values

INFO_W_NAME_ADDR

Result rows: 100

| xmember_sk | :ord_phys_sk | _provider_sk | patient_fir | patient_m | patient_la | patient_ac | patient_ci | patient_zi | ent_addr_lat | ent_addr_lon | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 47673 | . | 38696 | SHAWN | C | DIAZ | 2906 Ba... | Johns Is... | 29455 | 32.5853... | -80.162... | BE |
| 7932 | . | 38016 | MARILYN | J | HENDE... | 20 Que... | Hilton H... | 29928 | 32.16... | -80.75... | QU |
| [illegible] | [illegible] | [illegible] | MICHELLE | S | GILMORE | 7332 In... | Johns Is... | 29455 | 32.6241... | -80.151... | ON |
| [illegible] | [illegible] | [illegible] | [illegible] | C | LARA | 1108 Pa... | Charlest... | 29412 | 32.726... | -79.963... | RU |
| [illegible] | [illegible] | [illegible] | [illegible] | M | PIERCE | 309 Bla... | Landrum | 29356 | 35.1456... | -82.343... | RU |
| 62428 | . | 38826 | CATHE... | S | SHAH | 413 Sol... | Greenville | 29605 | 34.7581... | -82.395... | BR |
| 6732 | . | 38792 | LARRY | T | PATEL | 1901 12... | Greenville | 29607 | 34.82... | -82.34... | CH |
| 43298 | . | 38165 | STACY | M | JOHNS... | 80 113 ... | Greenville | 29605 | 34.79... | -82.39... | JA |
| 49782 | . | 38091 | PAMELA | D | THOMP... | 5466 N... | Landrum | 29356 | 35.13... | -82.21... | DE |
| 28327 | . | 38059 | SANDRA | M | MCDO... | 243 Ho... | Landrum | 29356 | 35.1388... | -82.226... | SA |

1. Parsing

Source column: patient_address

Locale: English (United States)

Definition: Address

Selected tokens: none

Available tokens (6):

Selected tokens (0):

**Display 6: SAS® Data Preparation– Suggestions**

Examples of automated data preparation suggestions that would be supported within SAS® Data Preparation:

- Apply SAS® Quality Knowledge Base data quality functions such as gender analysis, parsing, standardization
- Enrich data through address verification and geolocation
- Rename, remove, convert, obfuscate columns
- Impute missing values
- Fix outliers
- Transforms for normality: exp, 1/x, x2; ln
- Normalize values, so they are all in the same range

From the data discovery metrics, it could be also interesting to propose business and data quality controls through the analysis of frequency distributions of values and patterns as well as the combination of variables.

From a data governance perspective, "Suggestions" could also help in getting critical terms to be created relying on the analysis of variables across reports and data sets, in prioritizing issues to be remediated or recommending data retention period based on the sensitivity of data.

## ANOMALY DETECTION

The anomaly detection use case aims at identifying potential risks in data, or in data management operations. There are multiples opportunities in that domain as data discovery metrics provide several measures allowing to measure how spread variables are, as well as outliers and frequency distributions of values and patterns. Such measures help in identifying values that are out of range or in detecting inconsistencies in columns that have been identified (for example, inconsistent emails, URLs, ZIP codes, codes with specific patterns or referring to defined reference data).

Record-level analysis can also allow to identify potential duplicates and combine with suggestions, entity resolution processes can be created. In case of presence of personal data, the analysis of the risk of reidentification of individuals is also an excellent use case for fulfilling data privacy principles.

Another area of automation is the analysis of data quality metrics trends such as completeness, consistency, accuracy. Uncommon variations of these metrics are typical signs of data quality issues, for example: number of records processed on day 2 differs by 50% versus day 1, or completeness rate for one variable did drop by 10% over one period.

Platform logs analysis can also reveal incompliant use of data or data breaches. The only caveat here is the generation of too many anomalies, therefore there is a need to assess their importance.

### DEVELOPMENT AND ADMINISTRATION

This final category focuses on facilitating and streamlining Data Management and Governance activities for ETL developers, data stewards and platform administrators. The intention is to automate repetitive manual tasks i.e.:

- Auto complete of ETL or data preparation steps with the most likely configuration.
- Auto map variables when building ETL jobs
- Propose integration/preparation templates according to specific use like de-duplicate one data set, match, and merge two data sets, enrich one master data set, build SCD type processes …
- Select the most appropriate compute engine depending on the operations, the databases used, and the volume of date.
- Support performance self-tuning within data integration jobs, queries according to data volumes, storage type
- Alert on scheduled data management processes duration taking more and more time

Possibilities are unlimited until speed is increased.

## CONCLUSION

Considering the increasing volume, variety, and velocity of data to analyze, added to the metadata, users' feedback/rating and actions created as part of the platform, automation is the only way to perform effective Data Governance and build the necessary trust in your data and from stakeholders. It must be combined with a constant focus on democratizing data governance and to take it out from an IT only perspective. Data Governance is not a centralized activity but rather part of all managers activities and mission objectives.

Data Governance can be smarter, it can be automated, by relying on analytics and artificial intelligence so personal and sensitive data can be detected, business rules or quality controls can be suggested, and remediation actions can be proposed. Without minimal human interaction. Tremendous times are coming for building such services, which will empower data governance products to make suggestions and recommendations of actions to performed to users. Past this, it could even surface the invisible business rules or relationships between columns or data sets and facilitate the subsequent remediation of issues through prioritization.

## REFERENCES

Gregory S. Nelson ThotWave Technologies. 2018. "Data Management Meets Machine Learning" *Proceedings of the SAS Global 2019*, 1683-2018, Denver CO.

Margaret Rouse, “Definition of Data Democracy”, Techtarget.com, February 2017, Available at https://whatis.techtarget.com/definition/data-democratization

Laura Ellis, “Building a Data Democracy”, Data Science Central, May 1st 2018, Available at https://www.datasciencecentral.com/profiles/blogs/building-a-data-democracy

Paul Zak, “The Neuroscience of Trust”, Harvard Business Review, February 2017, Available at https://hbr.org/2017/01/the-neuroscience-of-trust

Ian MacKenzie, Chris Meyer, and Steve Noble, “How retailers can keep up with consumers”, Mackinsey, October 2013, Available at https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

## ACKNOWLEDGMENTS

I would like to express my special thanks to the SAS Data Management community and especially to the Product Management and Research and Development teams. I would also like to thank Amy Peters and Amy Wolf of SAS Institute Inc. for thoroughly reviewing this paper.

## RECOMMENDED READING

*Data Quality for Analytics Using SAS, Gerhard Svolba, May 2015*

*Data Preparation for Analytics, Gerhard Svolba, April 2015*

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Vincent Rejany
Domaine de Grégy
Grégy-sur-Yerres
77257 Brie Comte Robert Cedex
SAS Institute, Inc.
+33 (0)6 40 54 17 99
vincent.rejany@sas.com
http://www.sas.com