

WHITE PAPER

Data Governance:

Collaborative metadata-centric approach

Data nowadays



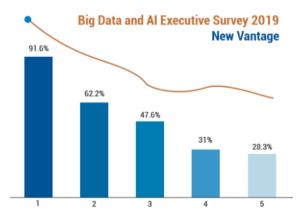
One of the reports from the World Economic Forum 2012¹ says "data is a new class of economic asset, like currency or gold". This fact, that affects all organizations regardless of their size, business nature or purpose, grew data's importance, elevating it status to strategic asset that every organization must manage accordingly.

What do organizations do to manage such an important asset?

It depends heavily on their data strategy, though in most of the cases they try to broaden their capacities inside the data lifecycle: acquisition, storage, processing, transformation, exploitation, and data availability. Almost every organization say that they want to become data-driven, there are less than few organizations which consciously quote: "Let's not use data, our feeling will lead us to make right decisions."

Nevertheless, having a look at the following graphic², we can find some outstanding values in data-related activities inside the organizations:

- **1 91.6%** of the leading organizations are Investing in Big Data and AI.
- 2 Only 62.2% are reaching positive measurable results from their investments.
- **3 47.6%** say that they compete with their data and analytics offerings.
- 4 Barely 31% achieved to create a data-driven company.
- 5 Only 28.3% have managed to create a data culture.



¹ 2012 World Economic Forum. Documentation available at https://es.weforum.org/events/world-economic-forum-annual-meeting-2012.

² New Vantage Partners, "Big Data and AI Executive Survey 2019" available at https://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-Updated-010219-1.pdf



Compliance and talent

In hindsight, the return on investment does not have the expected results for the greater part of the survey. Meanwhile, organizations must face a high load of compliance and regulations around data processing and data use - each regulation with of their share of acronyms, like RDA, LOPD, GDPR, PDPL, PDPA, CCP... depending on their nature and geographical placement. Consequently, enterprises have to special care when developing and deploying new versions due to this legal context.

On the other hand, even as important, these same organizations are caught in a war to attract the scarce best and high cost talent.

The CDO and Data Governance

Why are these initiatives failing and do not have the required impact?

As stated in that same report, taking as a reference the 3 axis of data digital transformation, the motivations are distributed as shown in the next image (image 2):

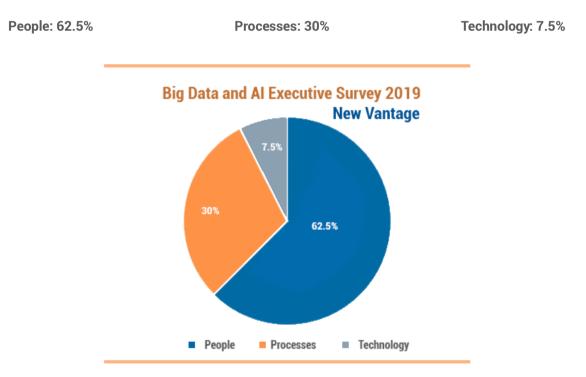


Image 2: Causes of failure in Data Governance initiatives.

Checking this **DATA**, technology does not seem to standout being the culprit, rather showing that there must be a focus, instead, on profiles, on internal processes, and, of course, on taking care of the raw matter (the data). This brings, inevitably, **Data Governance** to the table.



Who is in charge of implanting an efficient and effective Data Governance?

Here is where we start talking about Chief Data Officers and their importance in an organization that wishes to become data driven. For instance, Gartner estimates that 90% of the biggest world corporations will have a CDO at the end of 2019³. However, they also estimate that 50% of them will not reach expectations in the first 18 months after creating this role.

The CDO figure has suffered an evolution inside organizations. New Vantage and Gartner define 4 phases for the role depending on their level of maturity, shown in table 1:

The 4 phases of the CDOs

CDO 1.0	In this phase, they are exclusively dedicated to define the policies and procedures for data management in relation to metadata, quality, lineage, remediation plans, roles and responsibilities, all scoped to cover the current compliance.
CDO 2.0	In this phase, they extend their functions to non-regulation centered use cases and they adapt their role to support advanced business analytic initiatives.
CDO 3.0	Now they lead a great portion of the digital transformation initiatives based on data use and processing (especially Big Data, Cloud and Analytics), guaranteeing data governance as part of these initiatives.
CDO 4.0	In the final phase, they report to the executive board as another strategic figure inside the organization, changing the board's vision towards a more product view, managing profits & losses for data monetization purposes.

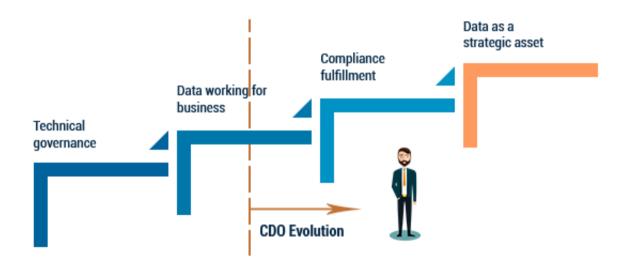
 Table 1. Lines for CDOs depending on their maturity

Those data-driven organizations that make a profit out of their data investments have achieved data governance models based on a well-combined set of policies and procedures, establishing the processes that supervise and manage data in order to transform it into a valuable asset.



Evolution of Data Governance initiatives

The evolution of the CDO commented above is deeply linked with the evolution of how the organizations face and implement Data Governance initiatives. The evolution of these initiatives is, at the same time, also linked back to the promotion of this profile inside the organization.



Technical governance

Data governance was born from the "Data Management" discipline, connected mainly to IT departments, as an add-on to databases and to the database administrator role.

In their first stages, data is becoming organized in several silos, with IT leading roles and technical profiles. They must ensure metadata, quality and integrity of data in any of their systems, facilitate exploitation, and even provide, in the majority of cases, ad-hoc information to users. These stages match with an increase of data usage for decision making inside organizations, and with the arrival of Business Intelligence and reporting tools to help create dashboards and reports.



Data working for business

With the Internet boom, the proliferation of mobile devices, and the appearance of new technologies to capture, storage, process and exploit data, a new set of innovative companies emerge who understand data in a different way and think it has a differential strategic value.

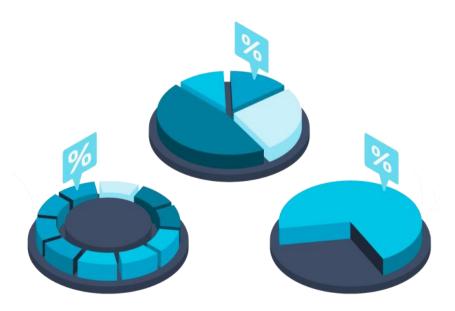
These purely technological enterprises put data first and launch new business lines and products around it. Henceforth, a new era appears where it is necessary to govern differently, to transform data into information and information into knowledge. Now data governance is business too, instead of technology only.

Regulatory compliance

With the appearance of RDA and its 14 principles for an effective aggregation of risk and reporting data, in the non-data-oriented organizations a new figure driven by compliance needs arises, the Chief Data Officer.

The Chief Data Officer role appears in organizations out of the need to manage data, its metadata, traceability, lineage, quality, etc. from a central point and with a board view. Initially, priority is regulatory compliance, but as compliance is no longer purely technical, CDOs must create data governance projects at scale, withstanding a high human workload, defining policies, standards and procedures, albeit frequently neither agile and nor flexible.

Afterwards, regulation hardens, and data use increases exponentially for decision making, and the CDO role becomes more relevant inside the organization, budgeted with more resources and capacities, aiming to create a differential value for business.





Data as a strategic asset

Just without time to take a breath after the investment effort in data governance for regulatory requirements (not all sectors suffer the same impact), organizations start to understand the value of data throughout existing and new business lines, launched for the sake of competitiveness, extending what has been deployed for regulation compliance to the other areas.

Despite this fact, three new problems, that had not been taken into account with the previous approach, derived from this approach, as shown in the following table:

Three new current problems of Data Governance

- Data governance, at this point in time, does not provide the necessary flexibility nor agility that they need in this new stage, where the aim is not exclusively addressing regulation anymore, but generating value for business in a quick and effective way. Thus, providing tools that automate processes and help change management is essential.
- 2 New technologies open a totally new paradigm, evolving from data governance silos to a collaborative governance. Big Data and Cloud bring with them distributed storage, process parallelization, ephemeral platforms and on-demand services, generating a lot more complexity. Furthermore, proliferation of open-source and the progression of standards mandate an agnostic view of technology, avoiding the trap of having to change implementations every few months.
- 3 A big gap between offering and demand arises in terms of capacities, talent, culture and profiles around data, from the technological and business standpoint. It is imperative to count with expert and specialized profiles in each aspect of data, but also business profiles that have the right culture and have a sensibility towards data, and the right knowledge to use it. In the same way, mixed or hybrid profiles (with technological background and business view, or business knowledge and technological expertise) grow in importance taking board and management chairs.

Table 2. Current problems with Data Governance.

Why do Data Governance iniciatives fail?

Data Governance initiatives are not an easy task. They require time, energy, investment, and business and technical experts with complementary skillsets. In addition, change is more complicated when being a culture transition transcending all corporate levels, so obtaining real value from data is not trivial.

In his article "9 biggest mistakes companies make when implementing data governance"⁴, Nicola Askham condenses the motivation behind the failure of this initiatives:

1	The initiative is leaded by IT only from a technology approach.	2	Data Governance is a time limited project instead of a long-term strategic initiative.
3	There is no earlier knowledge of the data panorama inside the organization.	4	Using "big bang" approaches rather than an incremental proposition.
5	Thinking that purchasing a tool is the solution.	6	The organization tries to afront Data Governance without being prepared or having the necessary maturity.
7	The data strategy and data governance are not aligned with the global corporate strat-egy.	8	The defined organization framework does not reach all the necessary levels inside the enterprise.
9	The objective defined consists only in reaching minimums to cover certain regula-tion or compliance.		

⁴ Askham, Nicola, "9 biggest mistakes companies make when implementing data governance" available at https:// static1.squarespace.com/static/52ed2570e4b02079a82e6ff3/t/56111545e4b0890ee92b5901/1443960146356/Nicol a+Askham+-+9+biggest+data+governance+mistakes.pdf



To all the aforementioned mistakes, we could add two more:

Building a Great Wall between businessand IT likely growing isolation between them.

11 Not involving all the stakeholders, from the developers to the business users.

From the start to the end of the data lifecycle many people from a wide type of profiles and areas are involved, who have great variety of technologies available, together creating a big amount of processes of different nature. To make sense of all this, it is imperative to establish a framework, with a set of components that guarantee the desired outcome.

Yet not all is lost. We have walked some part of this road already, after years of trial and error, innovating and trying several approaches, and, luckily enough, new stronger voices appear that vigorously spread this new data culture, trying to provide some keys and best practices that, once correctly landed and applied, will increase the chances of success.



The need of a different approach

As we have seen throughout this article, for the last few years technology has evolved at a breathtaking speed, bringing new possibilities never seen before; data is now a new active asset for organizations, the capacity to use and process data increase exponentially, and the work culture has evolved towards new models of work and management, creating Data Governance challenges that have to be faced with an innovative stance, away from the traditional, where opposing roles promptly have to collaborate.

Keys in the collaborative metadata-centric approach

It is a new vision of Data Governance, disruptive and innovative, based on three fundamental pillars, show in table 3:

Three new current problems of Data Governance

People	This first pillar builds a collaborative environment with roles and responsibilities well defined, where each participant provides their share of collaboration, receiving back the resulting overall value from the community.
Processes	This second pillar builds agile and flexible processes, and thinks about evolving them over time, with a great grade of automation to bridge the gap between business and IT, intending data democratization in a secure and governed environment.
Technology	This third, and last pillar, builds a totally agnostic approach towards data processing technologies in order to support both previous pillars. It must be designed and developed keeping in mind effective and efficient metadata management (business, technical, operational and others) and their relationships in a centralized placement.

Table 3. Fundamental Data Governance pillars.



Roles and responsibilities

Roles and functions definition is the first of the steps that have to be taken so each stakeholder is identified inside the Data Governance model with their rights and responsibilities. In a collaborative environment, it is very important that each of the players add-up for the common plan, retrieving a direct value back from the workgroup.

There are several models or frameworks that propose some figures with their associated functions, though we have taken into account that any model must be brought back to the enterprise's context and adapted to our organization, and can and will be changing over time.

There are, however, a series of roles that will always exist in one way or another. From these, the most significant are those suggested in table 4:

Most relevant roles and outstanding responsibilities

Executive board

- In spite of not being part of the day to day Data Governance activities, they are in charge of weighing in and bolstering the Data Governance culture inside the organization, also making sure that the highest level decisions are aligned with the corporate strategy, and this global strategy with the data strategy.
- They must have an open and innovative mindset over the processing and use of data and push the required changes inside the organization.

Their role is key because they sponsor the Data Governance initiatives providing budget and resources to reach the planned target.



Owners and Stewards

- In fact they may not represent the same figure nor they may have the same functions when going into details (there is a lot of current literature on the topic), but for the sake of simplicity we are going to group them as being "data providers and data managers".
- On the other hand, we can see stewards as a cross figure because of functions more related to technology and operations, while owners are always more on the business side and will be the person in charge of the data under their wing.
- Both owners and stewards have data accuracy and quality responsibilities. They also have to reach data provisioning SLAs for all interested parties. In order to do so, it is mandatory that they define together the controls and KPIs that cover quality and data availability and use, including severity and its associated thresholds.

- Depending on what profiles are available inside the organization, these two roles can be assigned to the same group or even be the same person.
- In general, the owners are in charge of having business terms well and uniquely defined (business metadata) and the stewards are in charge of feeding good data in timely manner (technical and operational metadata).
- It is essential that they establish policies and specific terms of use, treatment and audit of their data and worry about who uses it, where, why, when, and for what reason, creating processes that manage and enrich not only metadata but also lineage and traceability during the whole data lifecycle.



CDO / Data architecture

We group the CDO role and data architecture role as both have cross functions to all business areas despite the fact that the second commonly is under the wing of the first and is a more technical role linked to data modelling.



- In any case, they need to have an holistic view of organization data, from the business point of view as well as from the technology standpoint, being in charge of aligning the data strategy with the corporate strategy and governance model, including the policies and procedures under them.
- Moreover, they will be in charge of selecting, implanting, managing and evolving the solutions and tooling that support all the governance of the data.
- It is paramount they define together an architecture and a data lifecycle for the organization, in sync with the technical architecture and software lifecycle, and share them with the whole enterprise.
- Lastly, yet not less crucial than the above, they will have the responsibility of optimizing the processes and the resources to deliver, maintain and expand the governance.

IT

- > Oftentimes, we will encounter IT working as a cross area to all business areas, totally independent of them, mainly working with external or subcontracted profiles (even more with the preponderance of public clouds). For this reason, it is very important to recognize that IT is a vital enabler, making sure everything happens in a correct and controlled manner.
- IT is in charge of providing, maintaining, operating, administrating and evolving the technical infrastructure necessary (machines, servers, network, databases, equipment, ...) in the whole data lifecycle guaranteeing availability and security of access to the data at the lowest physical level.
- Here is where the biggest effort has to be done to automate processes and reduce time, without leaving aside security and controls.

Legal, compliance and Audit

- > These are the three central areas that take care of the compliance of the particular internal and external rules and regulations that the enterprise is exposed to.
- In order to so, also counting on business and technical profiles, they exercise audit on the use and handling of data and analyze the related processes. Typically, after an exhaustive analysis they emit recommendations for their implementation based on the revised information, to help areas address the misalignments with regulatory requirements.

Developers

Developers, who have gained a lot of weight inside enterprises in recent years due to the fact that technology has gained weight too, are forgotten by the data governance, in spite of being the most flexible of all the aforementioned roles, and the first who interact with data.



- Stemming from business, sometimes completely separate from it, they are always fundamental, preparing all the data reads and writes, becoming frequently consumers and/or producers, regularly wearing both hats.
- Whatever side it may be, developers are a critical piece in metadata enrichment, maintenance and trace from the technical standpoint, and their work must certify data quality in production environments.
- The value that they can extract from data governance depends if they are on the data production or consumption side. On the first side, they can do an impact analysis before the deployment of any changes, if having a trace handy; on the second side, they will be able to better understand the data they have to deal with, if the data is well metadated and governed.



Business Users

- It is the role that will obtain the most benefit out of data governance because, thanks to proper deployment, they will have agile and quick access to data, being able to transform it into information and later on into knowledge.
- In exchange, they will have to respect the data terms and conditions established by the governance bodies and follow the procedures defined in the governance model.

Table 3: Data culture roles and responsibilities description

Committees and workshops

As significant as the roles inside the organization are the committees and the workshops who follow each initiative. They serve as a focal point where participants make decisions, resolve and can comment. Committees and workshops have a recurring nature, yet somewhat spaced in time, since their aim is to monitor and take high level decisions and must not be abused with unbounded workloads.

Committees and workshops should have an official status and exist as a permanent list plus other optional guests, depending on the matter at hand. Also, these governing bodies must gain a series of well-defined hierarchical functions, and carry non-overlapping decision-making capacity, to keep them independent – decisions can be rerouted from one body to another. For these entities to be effective and efficient they need a previously agreed agenda for each meeting and have a periodic follow-up on the approvals or themes to be addressed so they do not exceed available time.

Committees and workshops can be newly created data specific, or new data topics can be introduced in existing entities. Commonly, there is a mixed plan: first starting by defining and implementing committees, making them more relevant, later in time, adapting workshops, generally grouped by functional and/or technical areas. Committees frequently are organized in work-type, executive and board-type committees. They also separate technical issues from other functional or hybrid ones. An interesting proposal is to consider forming specific instruments in governance matters depending on the importance of each matter inside the organization (architecture, quality, metadata, security, ...).



Policies and procedures

Policies define "game rules" and the procedures establish how roles interact. It is important to define their scope, as a policy does not apply in the same way for all data types, or a procedure does not have to be the same depending on the step in the data lifecycle where it is to be applied.

This is why policies and procedures have to be born agile and flexible, adaptable to a disparity of situations, revisable and evolving over time, although striving, for commodity reasons, for the highest level of standardization. Likewise, considering "fast-track" or "waiver" mechanisms for a subset of cases, agreeing thoroughly on the conditions for its application, so exceptional mechanisms are not abused. In the case of role-based procedures, they have to be conceived for automation, guaranteeing their trace, measure and audit.

Examples of "Policies and procedures" are: new and/or modification of business terms, register of new projects or extension of use cases, definition and deployment of quality rules, creation and edition of data structures, data access and data request, information ingest request or data services, security mechanisms application (encryption, anonymization, masking, ...), deployment of new processes in production,...

Change management and communication

Like any culture change initiative in an organization, deploying an effective and efficient communication model is more than necessary. To do so, not only we must have strong communication channels that reach all enterprise levels, but also shared repositories, accurately organized and with good administration, to publish newly generated documentation, of interest to all participants.

Additionally, workshops and communication sessions can be held to show new releases, to spread awareness and capture participants attention, and also for question solving and help requesting. As another resource, bots can be created in the diversity of channels and group mailboxes, to publish information and for question answering.

Lastly, to allow the interaction of the diversity of members, and to favor collaboration, it is helpful to use more akin mechanisms and communication tools, instead of the more common office and mailing tools, deploying workflows, reminders, messages, and automated alerting, all based on roles, to simplify and foster agility for the most operational tasks.



Data lifecycle

The data lifecycle comprehends the definition of each of the information layers for the whole organization and it has to be aligned with the technical architecture and with the full software lifecycle. For each layer, requirements and conditions for modelling and development will be defined, as well as possibly best practices, making sure that they are adopted for the correct use cases.

In every data lifecycle phase, metadata requirements will be defined too, as also independent lineage, traceability and audit, though many can be common and shared, being deeply related to processing characteristics (online, real time, batch, ad-hoc, ...)

Examples data lifecycle layers are: operational, transformational, informational, data lake, cold storage, analytics, exploitation, ...



Technology-agnostic metamodel

The importance of one unified metamodel, centralized and technology-agnostic, resides in the understanding of data through one single common language, from both a business and a technical view. Any system or any technology will handle their own metadata and have their own metamodel, being essential to have a unified view that agglutinates all of them, enriched with new metadata.

A complete metamodel has to be the sum of the object definition (data structures, processes, roles, users, systems, applications, ...) and their characteristics, object relationships, and have to consider the hierarchy between objects of the same type.

As explained previously above, the metamodel has to be flexible, adaptable, and scalable, defining the group of minimum necessary metadata to be informed, depending on the data lifecycle phase and the policies to apply, for each of the data types.

Lastly, it is paramount that inside this flexibility there is a standardization with the aim to ease the management and allow the automation of technical processes based on informed metadata.



Automation of technical processes

One of the direct benefits of standardization and homogenization previously shown is the possibility to automate technical processes, reducing the number of resources dedicated to these tasks, execution time and operational errors derived from the human interaction, while providing traceability and audit. This automation also generates value for the involved parties, fostering user experience when implementing data governance.

We are not only talking about automating procedures but also automating technical processes underlying the data processing thanks to informed metadata, like, for example, definition and execution of ETLs and ingest, development of quality processes, new user creation, setting permissions, data structure creation, data availability as consumption services, security mechanisms application, deployment of new software applications, etc.

Incremental approach

As seen previously, the "Big Bang" approach is one of the most frequent reasons why a data governance initiative fails, and it makes very difficult to measure the results as metrics are not isolated for each scenario.

Therefore, it is key to select scope-limited use cases that cover the breadth of what was defined, allowing the measuring of concrete results, not only to achieve internal visibility, but also to accomplish an improvement through the measure, tracking and analysis of the designated metrics. Step by step, later on, opening to more use cases, understanding that, whatever is defined and deployed for one use case, has to extrapolate and adapt for the use cases to come.

An extended practice that works quite well is separating stock from new production, selecting use cases that run upon stock, applying this incremental approach, whilst trying to add data governance for all new production initiatives from the very start.





Proactive and preventive management

In the best of cases, governance starts in parallel like any other part of a data initiative, though experience shows that governance lands in a context where already multitude of data and processes are to be overseen, so an evolution of data governance maturity needs to be followed.

As the governance implementation goes forward, it is essential to turn to a proactive and preventive management as a differential factor, anticipating problems and, at the same time, raising the speed at which the incremental approach is applied.

Thanks to the commitment of several proactive governance techniques, like platform monitoring analysis of theoretical vs. real metadata, non-governed data and processes can be detected, and gaps between what is believed to be happening and what is exactly happening highlighted.

Another extended and powerful technique is detecting bottlenecks in our governance model to address them, evolving the less efficient processes consequently enhancing user experience.

Finally, due to the great amount of metadata gathered, advanced analytics can be employed to further complete the data governance. For instance, it can be used for data structure and processes optimization, recreation of audit from partial data, duplicate and inefficiency detection, or the anticipation of errors in data quality.





Solutions and tooling

They are not essential from scratch. However, it is essential to know when to include them in the Data Governance model, so they help achieve our targets and, above all, helping to early onboard all parties.

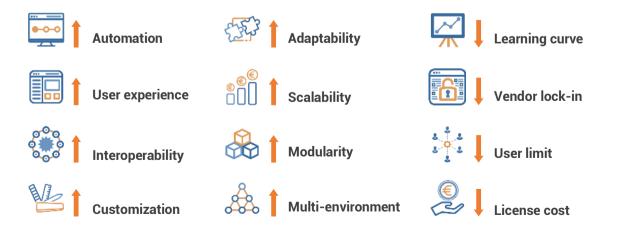
To accomplish onboarding, a high level of automation of manual tasks is fundamental. Also, an amenable and intuitive user learning curve experience, so no new wall to new joiners is created.

In such way, technological solutions have to be understood as change accelerators and change facilitators, but not as a panacea, since they will have to integrate all existing ecosystem and governance. This is why the choice has to consider local enterprise context and will need a thorough evaluation of all available possibilities.

Configuration and adaptation capacities, scalability and interoperability are crucial variables in a data governance solution in the current era. Therefore, it is critical to opt for solutions technology agnostic from data processing tools to be able to readjust to governance needs as these mature.

Hence, the search is narrowed down to API-first solutions developed with a services-oriented modular architecture, that adopt consolidated open-source technologies providing high scalability and adaptability to hybrid architectures, certain Big Data capacities, and can be easily adapted to our governance model and our enterprise, and, on the long run, do not create an endless vendor lock-in.

Based on this scene, it is reasonable to consider that the best strategy is whether to opt for a specific market solution offered by a provider, whether to do a few internal developments with internal or external profiles, or whether to reuse tooling already available inside the enterprise with this new aim. Typically, picking one of these options does not exclude the others, creating an environment where a large group of diverse pieces fit together like a puzzle, and where this puzzle integrates with our governance model, that, on the long run, is what matters.



Conclusions

The organizations with the vision and the capacities to convert raw data into information and create a new knowledge set have now a perfect opportunity to do so. When the boards acknowledge the value of data, they will advance from a cost center to a profit center. An adequate data governance deployment has a high impact on the organization performance and can create a competitive advantage. Albeit, it is a big challenge to deliver the right mix of people, processes and technologies to design a successful initiative.

To achieve this challenge, we must create an effective data governance strategy, directed by business objectives, providing parties with the best decision-making capacities, helping the enterprise reach their aspirations. A capable plan must assure that the corporate mission, the business strategy, investment and systems are all aligned.

Additionally, it is important to highlight that a data governance model must be built agile and adaptive, being thought of live being that invariably evolves to reach organization objectives. Likewise, a good change management must be in place, taking care of communicating what is to be deployed and how, and when users will see the expected results.

To do so, it is vital to start with policies, general guides and high-level diagrams. As the ecosystem matures, so will the formal documentation and level of detail for each of the identified scope. Initial use cases selection and new initiatives proposition, that are later applicable to the rest of the enterprise, will be also key factors for success.

Finally, there must be an effort to push the data governance strategy as an integral part of corporate vision, iterating and obtaining ever more details with each cycle. As business needs change, there has to be a plan to evaluate and reinvent continuously, keeping in mind current and future technology trends to build a successful data governance strategy.

Here we are proposing an innovative and disruptive model, away from a traditional model, adapted to the new technological era, that, together with the correct solutions, can deliver a robust implementation of our data governance initiatives.

Contact:

Mario de Francisco Ruiz mario.defrancisco@anjanadata.com

This technical document presents "Data Governance according to Anjana Data: collaborative metadata-centric approach". Even so, we will openly consider all your suggestions or comments to complement this view. For further information you can visit **www.anjanadata.com** or drop us an email at **info@anjanadata.com**.



www.anjanadata.com

St. Poeta Joan Maragall, 23. 28020 - Madrid, Spain

