TOWARDS ANALYTICS DATA GOVERNANCE
IN FAMILY COMPENSATION FUNDS IN COLOMBIA - CASE STUDY

SUSANA LONDOÑO MUÑOZ

Thesis

Advisor
Marta Silvia del Socorro Tabares Betancur

UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA
MEDELLÍN
2024

TABLE OF CONTENTS

FIGURES

TABLES

ABSTRACT

Family compensation funds are private social organizations, characterized by delivering various services to workers and their families. With a wide diversity of services and public, their data will be equally heterogeneous. The nature of family compensation funds functions means that data decision-making directly affects the delivery of services associated with the quality of life of workers and their families. The impact of data and analytics on the lives of these individuals is profound, making the role of Analytics Data Governance even more crucial. This article proposes a model of Analytics Data Governance that can be applied in the family compensation funds of Colombia, emphasizing on analytics data by adjusting the elements of a transactional data governance model to the phases of analytical life cycle proposed in CRISP-DM, this will allow control over the different data sources, formats, and participants in the various phases of the analytics data lifecycle that are part of decision-making in these types of organizations. Only two articles were found addressing data governance in family compensation funds in Colombia, and their emphasis was purely on transactional data management. Crucial elements of data governance such as roles, responsibilities, policies, standards, and procedures are proposed to address these organizations' challenges in analytics data governance. In the context of Comfama, where the implementation of a data governance is in progress, this work addresses how the elements proposed are applied to extend data governance to analytics data.

Keywords: Analytics Data Governance, Family compensation funds, Data management.

INTRODUCTION

Currently, the importance of data and analytics in strategic decision-making is undeniable. However, it remains a significant challenge for today's organizations to understand the implications of data management for analytics, artificial intelligence (AI), and proper decision-making. The ability of analytics to provide objective and accurate information makes data essential tools for optimizing resources and mitigating risks, giving companies that base their decisions on data a significant competitive and strategic advantage. Data analytics facilitate market segmentation (Fantini, 2023), help predict and take actions to reduce absenteeism (Lawrance et al*.,* 2021), and even allow prediction of life events such as moving or having a child (De Caigny et al*.,* 2020). The challenges in data management for analytics and AI are pressing, and the proposed analytics data governance (ADG) model could mitigate these challenges, providing a structured and practical approach to governance in family compensation funds (Cajas de Compensación Familiar, CCF).

Implementing a solid ADG program facilitates the effectiveness and quality of decision-making through analytics, ensuring the appropriate control over analytics lifecycle. Data governance (DG) is the exercise of authority, control, and decision-making over the management of data assets. It aims to enable an organization to think, manage, and use data as an asset, similar to how auditing manages financial assets. DG exerts structured control over data in an organization, defining, approving, communicating, implementing, and monitoring principles, policies, procedures, and responsibilities. The ultimate goal is to ensure that data is of high quality, accessible, secure, and used ethically to generate value (DAMA International, 2017; Eryurek et al*.,* 2021; Ladley, 2019). The challenge of managing and governing a high volume of information from different sources is particularly relevant in CCFs, which collect information from companies, their workers, and their families. This challenge has extended from transactional data to analytical data.

This article proposes an ADG model for CCFs in Colombia to facilitate the control of analytics data used throughout its lifecycle and its corresponding impact on the organization. The development of this work will begin with a conceptual investigation of DG, frameworks, and working models for DG. Subsequently, it will review how these concepts have been applied in CCFs in Colombia, providing an overview of the topic in the proposed context. Based on this conceptual understanding, it will proceed to understand and characterize the different mechanisms and components the ADG model should have in CCFs in Colombia and propose it specifically to address these organizations' challenges and needs. Finally, a section will present how the elements proposed can be applied in a comprehensive case study.

THEORETICAL FRAMEWORK AND STATE OF THE ART

DATA MANAGEMENT AND GOVERNANCE

According to DAMA International (2017), data management is the business function responsible for developing and executing plans, policies, practices, and projects that control, protect, and enhance the value of data and information throughout their lifecycle, which includes the creation and collection of data under business rules, followed by processing it that will allow them to be stored in a standardized and controlled manner in systems set up for their consultation, distribution, and use; finally, when the data completes its temporal cycle according to regulations, it is archived or properly disposed.

DAMA International (2017) defines data governance as the exercise of authority and control (planning, monitoring, and control) over data management, defining what should be done (policies), how it should be done (standards and procedures), and who should do it (roles and responsibilities) in each phase of the cycle. Data management and governance enable efficiency, innovation, and resource savings. More efficient data use has proven to be a determining factor for the success of companies that base their decisions on data (Grover et al., 2018). Various data management and governance frameworks, such as the two editions of DAMA-DMBOK, have been proposed and widely adopted globally. These frameworks offer a functional model for implementing data management practices, establish a common vocabulary, and serve as a fundamental reference guide for those dedicated to data management (DAMA International, 2017; DAMA International, 2009).

ANALYTICS

Analytics is a set of techniques that examines data sets to discover patterns, identify trends, and predict phenomena, thus facilitating decision-making. In the organizational context, it improves the efficiency and effectiveness of business operations and optimizes internal processes, resources, and marketing orientation by improving the understanding of user behavior. This field is constantly evolving, driven by exponential growth in data availability and the development of new technologies and techniques for processing and analyzing these data, such as machine learning (ML) and artificial intelligence (AI). With the development of these data analysis techniques, new risks, roles, and responsibilities also appear, requiring DG to extend to data analysis and ensure that all analytics projects in an organization comply with specific standards and protocols (Fadler et al., 2021; Grover et al., 2018).

Typically, data initiatives begin with descriptive and diagnostic analytics, which seeks to answer questions such as: What happened? What is happening? Why did it happen? or why is it happening? This type of analytics allows analytics teams to show results, whether successes or failures, and thus take concrete action. Predictive analytics allows, with a particular error level, to give the probability of possible outcomes over time, guiding companies to make strategic decisions (Barton et al., 2023). This type of analysis involves the development of probabilistic models based on variables, including historical data, related to possible events (DAMA International, 2017). Techniques used in this type of analytics include prediction models, regression analysis, forecasting, multivariate statistics, pattern recognition, and machine learning (ML) (Gartner, Inc. 2024). Suggesting how to take advantage of an opportunity or avoid risk and recommending specific actions that will affect the outcomes gives way to prescriptive analytics (Barton et al., 2023). This type of analysis can continuously incorporate new data to refine its predictions and prescriptions, improving accuracy (DAMA International, 2017).

Various proposed methodologies offer a structured approach for data science projects, including CRISP-DM (Cross Industry Standard Process for Data Mining) and KDD (Knowledge Discovery in Databases). CRISP-DM is particularly relevant in corporate contexts and is structured in six phases, which can be understood as the analytics lifecycle.

1. Business Understanding involves identifying and understanding the business need, evaluating the context, identifying resources, requirements, and constraints, defining project objectives, scope, and success criteria, and building an execution plan.
2. Data Understanding: Identifying data sources, collecting, and initial exploration to deepen the understanding of available data and detect quality issues.
3. Data Preparation: Different processes are performed to obtain a high-quality data set for modeling. These consist of cleaning and transforming the data, handling outliers and missing data, normalizing or standardizing it, and selecting relevant variables.
4. Modeling: In this phase, various statistical and machine learning models are selected, built, and applied to the prepared data, performing fit and validation tests based on the defined success criteria, with several iterations possibly needed for the best results.
5. Evaluation: Present a comparison of the different models to determine the most suitable based on the business's objectives, needs, and success criteria. Another element is analyzing model results to obtain relevant information that may have been extracted.
6. Deployment: The best model is deployed to production, and project results are communicated to stakeholders. This phase also includes process documentation and recommendations for monitoring, maintenance, and identifying improvement opportunities.

Although theoretically, these phases are presented in a specific order, in practice, many tasks can be performed in a different order, in parallel, or even revisiting previous phases as necessary. These models provide a workflow that helps organizations manage their internal tasks effectively, define roles and responsibilities, and establish guidelines for activities.

FAMILY COMPENSATION FUNDS (CCFs) IN COLOMBIA

The family subsidy system was officially established in 1957 as a measure to redistribute the economic contributions of companies among the working class (Comfama, 2022). In Colombia, there are 42 CCFs, and as of September 2023, 21 million people benefited from the family subsidy system, with 790,151 companies contributing around one billion Colombian pesos (COP) per month, redistributed in-kind subsidies (food, clothing, study scholarships, school kits, medicines, among others), social credit, housing subsidies, unemployment, child protection, and family subsidy (Superintendencia de Subsidio Familiar, 2022). The CCFs are subject to state control and supervision by various entities, such as the Superintendence of Family Subsidy (Superintendencia de Subsidio Familiar, SSF) and the Departmental Comptroller, depending on the services offered. They also have a control unit: the Internal Audit and Fiscal Review Unit (Comfama, 2020).

Regarding data, in Colombia, the protection of personal data arises with two regulatory systems that apply to the CCFs: Law 1266 of 2008, later modified by Law 2157 of 2012, related to financial, credit, and commercial habeas data defines classes of personal data, such as private, semi-private, and public data, and Law 1581 of 2012, which refers to any personal data in databases of public and private entities. Law 1581 of 2012 establishes special categories of personal data, such as sensitive data and personal data of children and adolescents. More recently, Law 2300 of 2023, known as *dejen de fregar*, seeks to set limits on contacting customers and consumers of products and services (Cabezas-Azuero, 2023).

DATA GOVERNANCE IN FAMILY COMPENSATION FUNDS IN COLOMBIA

In Colombia, Supersubsidio and the *Unidad de Gestión Pensional y Parafisales* (UGPP) have implemented several DG initiatives to establish basic guidelines for CCFs. Some guidelines are issuing normative circulars standardizing information reports using a standard and unified language, developing information reporting manuals applying information collection standards, and policies for reports on 4% contributions. These initiatives facilitate the oversight and control work of these entities over CCFs by standardizing procedures, language, and concepts around data.

The literature highlights two proposals for DG reference frameworks for CCFs in Colombia (Tobar-Ruiz, 2021; Balanta-Sinisterra, 2022). Both works take a family compensation fund from some department in Colombia as a case study. In (Tobar-Ruiz, 2021) the DAMA-DMBOK is adopted as the fundamental pillar of its framework, proposing three sequential phases: planning, execution, and monitoring and control. In planning, the mission, vision, objectives, scope, drivers, capabilities, and program policies are defined, in addition to an operational model based on the organizational structure and data management, including the structure of a DG committee. It also presents the value chain and data lifecycle within the organization.

Execution is divided into disciplines to govern, principal, and support disciplines. The disciplines comprise general guidelines such as mission, vision, objectives, governance policies, roles and responsibilities, and a business glossary. The main disciplines include data quality, master and reference data, and metadata definition, while the support disciplines include security and privacy, data modeling and design, and architecture. The last two disciplines include a schedule for deploying the necessary activities. Tobar Ruiz proposes the objective, roadmap, and required actions for cultural change management within the organization and monitoring and control indicators.

On the other hand, Balanta-Sinisterra (2022) uses the Data Governance Institute (DGI) type SAS methodology for its flexibility and ease of molding to any approach. The focus centers on data quality and master data, starting with a maturity diagnosis and developing DG implementation in three phases: organizational structure, processes and decisions, and operations plan. The organizational structure establishes authority over data; processes and decisions define governance policies, change management, problem identification, definitions, use, and data security; the operations model drives teams to develop process descriptions and additional policies and perform monitoring and control.

Both approaches recognize the profound need for robust DG in CCFs, highlighting the importance of defining roles, responsibilities, processes, and policies covering almost the entire data lifecycle and its undeniable connection to data analytics. However, none of these frameworks present the elements for an ADG encompassing management and advanced analytics.

PROPOSED ANALYTICS DATA GOVERNANCE MODEL

In the context of CCFs, the emergence of regulations that consider not only transactional data management but also its analysis and use is imminent since decisions are made based on this directly impact the delivery of services associated with the quality of life of workers and their families. When a CCF begins to explore the possibility of using data not only for management and reporting metrics, it decides to enable capabilities that facilitate data use for decision-making, thus entering the analytics lifecycle, where they will be collected, transformed, shared, and enabled for decision-making. Understanding that analytics (descriptive, diagnostic, predictive, and prescriptive) transforms data leads us to think they must be governed under new usage schemes, proposing policies, standards, procedures, and new roles and responsibilities.

The proposed ADG model takes four essential areas of knowledge from DAMA International (2017). These are data governance, quality management, metadata management, and data warehousing and business intelligence management, as a basis for designing data management governance for analytics data, ensuring control through every proposed phase by CRISP-DM lifecycle. As the conductor of data management, data governance for analytics must prepare everything to bring these elements to life within the organization’s operation: quality, metadata, data storage from strategic alignment, and definition of roles, responsibilities, policies, procedures, and standards. Based on these, supervision, problem management, and ensuring internal and external regulatory compliance are exercised.

Data quality must be understood from the conception of the strategy and organization’s goals; identifying business data needs and key indicators builds quality because existence is guaranteed with these definitions clear. This element accompanies us from the business understanding phase, through data understanding when a quality check is performed, to data preparation, where cleaning is done based on understanding. While this element cannot be managed on models or evaluation, its impact on results is massive.

Metadata is a fundamental element in data governance and management, not only at the transactional level but also at the analytical level. Metadata provides context to data, improves understanding between technical professionals and users, increases operational efficiency by reducing search and identification time, and avoids using outdated or incorrect data. These characteristics make them relevant in the initial phases of analytics, such as business and data understanding and preparation. They also participate in modeling, evaluation, and deployment, providing comprehensive documentation of the data’s context, history, origin, and, in this case, the developed models and their implications, assumptions, and functioning. This way, metadata reduces training costs and mitigates the impact of staff turnover.

Data Warehouse (DW) and Business Intelligence (BI) tools gain importance in deployment as they are where results will be housed and enabled for decision-makers, stakeholders, and data users. Storage tools will also be the primary data source for analytics, as the organization's master and reference data are there. It is essential to clarify that data stored on DWs must comply with internal and external regulations. The organization's Data and analytics ecosystem is governed by the elements presented below:

DATA AND ANALYTICS STRATEGY

As a fundamental element of any initiative to be carried out in a company, it is essential to align it with the organizational strategy, understand the drivers and scope, and propose objectives that drive the strategy. For this, it is necessary to identify the critical prioritized initiatives for the organization's purpose and analyze how data and analytics are essential to provide value to the organization, from which the scope of governance should be defined and options for implementing an ADG program should be evaluated.

These two elements should be used to define a data and analytics governance strategy. The strategy should determine outcomes based on the different ADG elements and show their impact on the critical prioritized initiatives for the company's purpose. An activity and role matrix with the respective responsibilities is proposed.

In CCFs, regulatory compliance and positively impacting the quality of life of affiliates and their families with provided services are highly relevant. It is essential to declare the importance of data, analytics, and their proper governance and management to mitigate regulatory risks, optimize resources, and impact the most significant number of users with the most relevant.

ROLES AND RESPONSIBILITIES

Integrating analytics into the core of data governance is fundamental to creating a robust, analytics-oriented data ecosystem. This integration implies including professionals from these areas in existing processes and defining roles around analytics.

Traditional data governance roles have focused on transactional data management, compliance, and quality. However, this orientation has limited the capacity of roles to capture the actual value of data, mainly realized through its analytical use. For this reason, roles at three organizational levels (strategic, tactical, and operational) are proposed. The strategic level is proposed as an integration between organizational, business, and technology strategies around data and analytics:

- **Chief of Data Governance (CDG):** Strategic leader of analytics data governance (ADG) within the organization. They are pivotal in guiding the data governance strategy for analytics, operating models, and implementation efforts. The CDG's influence extends to fostering the adoption of the data governance and analytics culture through training sessions, promoting best practices, and encouraging team collaboration. Their strategic vision of the capabilities needed to deliver new value through analytics is communicated throughout the organization, underscoring the importance of their role.

- **Chief Information Officer (CIO):** Proactive role in guiding business technology. In constant communication with the strategic team, the CIO proactively manages the adoption of new technologies and supports and manages tools to achieve ADG objectives. Their proactive approach ensures that the organization is always at the forefront of technological advancements, providing a sense of control and direction.
- **Business Manager (BM):** As data owners and business decision-makers, BMs are crucial in creating an ADG program plan. Their responsibility in establishing data governance for adaptive analytics, which involves creating policies, setting up governance structures, and ensuring that the organization's data is used effectively and responsibly, is particularly significant.
-

On the tactical level, strategy is brought down to paper:

- **Business Operation Leader (BOpL):** This role collaborates with business leaders to define transactional and analytical data management policies and procedures. It oversees processes ensuring data quality and accuracy within the business and manages data governance and management issues. This is a critical role, as the BOpLs ensure the data's quality, security, and utilization in their domain.
- **Information Security Leader (ISL) and Data Protection and Compliance Leader (IPL):** Engages with data and analytics leaders to ensure privacy and information security implications are understood and addressed in the governance initiative; guides planning activity for data asset management and governance and compliance activities for data privacy laws and regulations and implements technical security measures to protect data.
- **Data Analytics and Curatorship Leader (DACL):** Establishes the foundation and framework for ADG processes and practices; assesses maturity and cultural barriers; coordinates and enables governance activity through engagement between key business and IT people to achieve success.

Finally, the respective teams operationalize ADG in their domain. One team will focus mainly on transactional data, and the other will focus on analytics and the resulting data.

- **Business Operation Team (BOpT):** This team performs data management tasks, including data cleansing, data transformation and loading, metadata creation and management, rules and standards documentation, data quality profiling and improvement, among others. In other words, they provide technical expertise to materialize policies through standardized procedures. They constantly improve their skills to meet the organization's changing needs and ensure the smooth operation of the ADG in their domain.
- **Data Analytics and Curatorship Team (DACT):**
  - Data Engineer (DE): This role is responsible for developing and maintaining the data preparation and transformation processes. They collaborate with data scientists to optimize the performance and efficiency of analytical processes. Their technical expertise ensures that the ADG operates smoothly and efficiently.
  - Data Scientist (DS): transforms data from analytical models to obtain results that, when presented clearly and concisely, enable the understanding and use of data for decision-making.
  - Curator (DC): Implements and manages policies and procedures for creating, using, and disposing of analytical models. This role approves models that meet quality, ethical, and regulatory compliance standards. All these roles provide feedback on the quality and usefulness of the transactional data used and identify business needs for data and analytics.

With roles defined, responsibilities can be assigned. An activity matrix will correctly identify critical points, such as where policies should be implemented and who is accountable. In the case study, an example of a roles and activities matrix will be presented for clarity. Team (BOpT), Data Analytics and Curatorship Team (DACT), Information Security, Data Protection and Compliance Team) operationalize ADG, following the procedures and standards to effectively manage transactional and analytical data. Table 1 is an example of how a role, activities, and responsibilities matrix is built, assigning the respective accountability for each role in the defined activities.

PRINCIPLES, POLICIES, PROCEDURES, AND STANDARDS

Principles are fundamental ideas that guide behavior. They are the basis for building theories, belief systems, and behaviors. They act as guides for action, providing a framework for making decisions and evaluating situations coherently and substantively. Data principles in organizations are fundamental to ensuring that data management and use are effective, ethical, and aligned with the organization’s strategic objectives.

Policies are global, fundamental rules that codify management principles and support expected behaviors in data management and use. These describe what to and should not do for each element applied to each data and analytics lifecycle phase. There should be relatively few policies that should be expressed briefly and directly. Standards and procedures describe "how" activities should be done, ensuring that tasks are performed correctly and creating a basis for ensuring quality, efficiency, and compliance, thus facilitating effective management and continuous improvement of organizational processes.

Activities / Roles	Chief of Data Governance (CDG)	Business Manager (BM)	Business Op. Leader (BOpL)	Analytics and Curatorship Leader (DACL)	Business Op. Team (BOpT)	Analytics and Curatorship Team (DACT)		
						Data Engineer (DE)	Data Scientist (DS)	Data Curator (DC)
Accountability: Responsible (R), Executor (E), Informed (I), Technical Advisor (A), Validator (V)								
Strategic Alignment								
Identify the prioritized initiatives critical to the organization's purpose and analyze how data and analytics are essential to delivering value.								
Redefine the success factors, scope and drivers of the DG program.								
Establish the expected deliverables for each phase of the GDA program implementation.								
Define management and control indicators for the GDA program.								
Communicate results of key initiatives and the generated value by the GDA.								

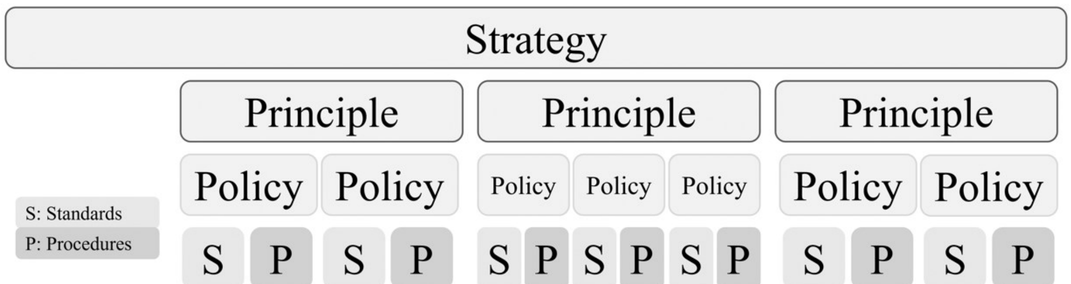
Table 1. Roles, activities and accountability Matrix. Example of roles and responsibilities matrix for strategic alignment with the proposed roles. Made by author.

We can think of these three elements as a tree branching out. The main trunk is the strategy, divided into main branches (principles), each with secondary branches (policies); finally, the leaves are the standards and procedures Figure 1.

From this understanding, it is necessary to emphasize the importance of including sections referring to analytical data, those produced from preparation and transformation through analytical models, in the procedures and standards within the organization, e.g., including within the data quality policy, data quality verification procedures at the time of their preparation for analytical models; it is crucial to define within the data use policies which are suitable for what type of analytical models, that is, if a recommendation model for employment service is to be carried out, the gender variable should not be used, as it may bias the recommendation, or conversely, which variables are indispensable according to government regulations.

Additionally, it is suggested that transparency, ethics, and responsibility in data use be incorporated as foundational organizational principles when applying analytics techniques. This approach would help manage risks related to analytical models, including potential sanctions the organization might face for failing to comply with data protection regulations or ethical expectations from users or executives.

Figure 1. Governance Tree. Minimum elements and how to define principles, policies, standards, and procedures from organizational strategy.



Source: Made by author.

Some proposed policies and procedures within the scope of the proposed ADG model:

Model Metadata Policy

- Purpose: Maintain a detailed and controlled record of all versions of analytical models to ensure traceability and reproducibility of analyses.
- Procedures:
  - Use version control systems to manage changes in models.
  - Document each change, including author, date, and reason.
  - A committee reviews and approves each new version before deployment.

Model Quality Validation Policy

- Procedures: Ensure that analytical models meet quality and performance standards, guaranteeing their accuracy and effectiveness before using them for decision-making.
- Procedures:
  - Define appropriate performance metrics for each model according to its specific application.
  - Perform cross-validation tests or use independent test datasets to evaluate model performance.
  - Document validation results and obtain formal approval before deployment.

  - The model must be periodically revalidated after any significant updates.

Model Deployment Policy

- Purpose: Ensure that models are rigorously evaluated regarding accuracy, effectiveness, and ethics before use and that their implementation is appropriate and secure.
- Procedures:
  - Develop a formal review and approval process for new models or significant changes in existing models. This process should include evaluation criteria such as accuracy, effectiveness, ethical compliance, risk, regulatory compliance, alignment with the strategy, and forming an evaluation committee.
  - Maintain records of all approvals and the basis for those decisions.
  - Conduct periodic reviews of models to ensure they remain adequate and effective.

Model and Associated Data Storage, Archiving, and Disposal Policy

- Purpose:
  - Define how and where models, their versions, and the prepared and created data should be stored, ensuring their protection and accessibility.
  - Determine guidelines on the retention period of models and related data.
  - Establish clear guidelines for disposing of no longer used models and associated data to avoid using obsolete models and ensure information security.
- Procedures:
  - Use document and archive management systems to store models and their associated documentation.
  - Define clear criteria for archiving and subsequent disposal of models, considering obsolescence, performance, incompatibility with current standards, and security.
  - Implement security measures to destroy any sensitive data associated with discontinued models.
  - Properly document and archive retired models along with the justification for their elimination.

ORGANIZATIONAL DATA CULTURE

Cultural change is a fundamental element for adopting an ADG model in organizations. Promoting a data culture within an organization is a process that requires the participation of all levels of the company. Implementing a series of mechanisms, from continuous training to actively promoting ADG principles, is essential. Below are some proposed mechanisms:

Committed Leadership. Top management must commit to the data culture and demonstrate support for ADG. The data and analytics strategy should be aligned with the broader organizational objectives and be part of high-level strategic discussions, so there must be visible leadership roles in data. In leadership spaces, data-driven decision-making should be promoted based on clear and robust data analysis. When data-based decisions are made, success stories must be reviewed, evaluated, and communicated internally to learn what was done well and keep doing it.

Communication. Plays an essential role in promoting cultural change, so continuous internal communication campaigns should be implemented, using different media to communicate success stories. Periodic bulletins should also be issued that include the importance of data governance, sharing policy updates and best practices, and reminding ADG policies within the organization and how to implement them in daily activities.

Training and Development. To promote the data culture in the organization, training at all levels of the organization is very relevant. Training in data skills should be offered to all employees, including leaders, according to their function and interaction with data. This should include modules on data protection, security, and ethics, teaching employees about the importance of handling data securely and responsibly. It is proposed that a manual on the ethical use of data for the organization be included, including the consequences of not handling data with due responsibility.

The introductory learning path to the organization should include a general understanding of how data is collected, managed, and used. ADG policies, ethical principles, security, and data protection of the organization should be made known, with an evaluation according to their role after completing the training to ensure they are effectively applied.

Advanced skills development is also essential. Interested employees should be offered opportunities to enhance their technical skills at an advanced level. Access to professional certifications in the technologies used by the organization, which can be promoted as recognition for their work, could lead to better practices and the development of advanced competencies.

CASE STUDY

In the context of CCFs, the emergence of regulations that consider not only transactional data management, but also analytical is imminent since decisions are made based on this directly impact the delivery of services associated with the quality of life of workers and their families.

When a CCF begins using data not only as management and reporting metrics, enables capabilities that facilitate data use for decision-making, thus entering the analytics lifecycle, where they will be collected, transformed, shared, and enabled for decision-making. Understanding that analytics (descriptive, diagnostic, predictive, and prescriptive) transforms data leads us to think analytics data must be governed under new usage schemes, proposing policies, standards, procedures, and new roles and responsibilities.

COMFAMA

Comfama is the third biggest CCF in Colombia. It has approximately 2.8 million affiliates, and more than 200 products of 8 different services. Comfama's purpose is to consolidate and expand the working class, providing solutions that improve life quality to workers and their families.

At Comfama, a program of data governance and care (GCD) has been developed that aligns with the organizational culture and the data and analytics strategy. This strategy emphasizes the centralization of data and the decentralization of analytical capabilities. The program has successfully defined information domains, roles and responsibilities, a business glossary, and an information catalog.

Nonetheless, the program hadn't consider the analytical data and how it should be governed, this case study proposes a route to include analytics data to the governance program, including a comparison of roles and activities based on CRISP-DM analytics lifecycle.

Structure

The GCD program at Comfama is structured as a centralized data governance scheme, with a decision-making entity known as the Data Governance and Care Forum. This forum includes the GCD manager, technology GCD manager, data custodian, and executive business data stewards for each macro domain.

Information Domains. Seven macro domains have been identified at Comfama: family compensation, transversal services, subsidiaries, families, health, companies and regions; these are further subdivided into 24 domains and 36 subdomains. These domains align with organizational structure and data architecture.

Roles and Responsibilities.

- Data Governance Manager: Responsible for deciding on data governance strategy, defining policies, and ensuring compliance with regulations.
- Data Steward: They are the first level of supervision of their domain, ensuring data quality and governance.
- Data Administrator: Responsible for the operationalization of data governance, including policy implementation and resource management.
- Security Custodian: Provides cross-cutting support in information security archetypes, ensuring data protection.
- Privacy Custodian: Provides cross-cutting support in privacy archetypes, ensuring personal data is handled in accordance with regulations.
- Analytical Steward: Responsible for data processing regarding business intelligence and analysis.
- Transactional Steward: Support and knowledge of transactional systems, ensuring proper capture and flow of operational data.
- Data Influencer: Key individuals who influence the Data Driven culture within the organization.

STRATEGIC ALIGNMENT

Some activities are proposed to adapt the program to the ADG proposed model in Table 2. It is essential to identify the critical organizational goals that should be leveraged from data and analytics to align the DG program strategically; for example, CCF's services coverage is a monthly reported indicator.

Some fundamental definitions of governance programs must be evaluated and, if necessary, redefined, such as objectives, drivers, scope, and activities to be carried out in each data and analytics cycle. This evaluation sets the foundations to include the analytics lifecycle within the scope of the DG committee.

For Comfama, fundamental definitions like objectives, drivers and scope are being evaluated in the light of analytics data, to ensure that not only data governance program is applied to transactional data.

Proper roles, responsibilities, policies, procedures, and standards must be defined here. However, a list of activities comes to life when the appropriate role and responsibility are identified to execute, advise, validate, or supervise each activity.

ROLES AND RESPONSIBILITIES

Therefore, the next step is to evaluate and adjust roles and responsibilities, including activities for each phase of the analytics lifecycle. Comparing the roles described in Comfama's DG program and those proposed in the ADG model (fig 2), some adjustments are proposed considering the organizational structure.

The IT manager must participate and communicate with the organization at the strategic level and is responsible for managing the adoption of new technologies and tools to achieve DG objectives. IT managers should know and drive the organization's goals through emerging technologies.

Although Comfama already has a role related to analytics, there is no clear leadership at the tactical level, so there may not be communication between the strategic and operational levels. The lack of lead leaves the operational role under the business operational administrator’s guidance, therefore performing activities related more to management than to analytics or even acting autonomously without prioritizing business needs, reflected in the poor value perceived by leaders from the data.

Based on this definition, Table 3 can be used as an example to build a matrix of roles, responsibilities, and activities, which will identify what each role must do and its responsibility in that activity (the complete matrix can be found in Annex A).

However, at this point, each Comfama employee will carry out their activities in their own way; hence, governance must define what and how the activities will be performed.

Activities
Strategic Alignment
Identify the prioritized initiatives critical to the organization's purpose and analyze how data and analytics are essential to delivering value.
Redefine the success factors, scope and drivers of the DG program.
Identify stages in the lifecycle of transactional data and the analytics process to identify potential adjustments to roles, responsibilities, policies, procedures, and standards or define new ones.
Evaluate and adjust roles and structure considering the analytics process.
Evaluate and adjust policies, epphasizing quality and security, ensuring the specifications of analytic-related processes.
Define the operating model for managing transactional data and the analytics process.
Establish the expected deliverables for each phase of the GDA program implementation.
Define management and control indicators for the GDA program.
Communicate results of key initiatives and the generated value by the GDA.

Table 2. Strategic alignment activities. Activities proposed for the implementation of ADG model proposed for CCFs. Made by author.

PRINCIPLES, POLICIES, PROCEDURES, AND STANDARDS

As the purpose of ADG is to ensure that data within an organization is accurate, consistent, secure, and accessible, the fundamentals of behavior (principles), what can be done (policies), and how each activity related to data management, both transactional and analytical, should be done (procedures and standards) must be defined.

When reviewing the existing documentation of the fund's DG program, policies of compliance guidelines, information access, responsibility over information in processes, database treatment, information security management system, information treatment assurance, information generation, supply and visualization, data warehouse, business intelligence, information purging and retention, data domain, employee data updating, human talent information access and modification, and information governance are found.

Figure 2. Organization and proposed roles compared.

Comfama Role	Proposed Role
Data Governance Manager	Chief of Data Governance
Data Steward	Business Manager
Data Administrator	Chief Information Officer
Security Custodian	Business Operation Leader (BOpL)
Privacy Custodian	Information Security Leader, Data Protection and Compliande Leader

Comfama Role	Proposed Role
Analytical Steward	Data Analytics and Curatorship Leader
Transactional Steward	Business Operation Team (BOpT)
Data influencer	Data Analytics and Curatorship Team (Data engineer, data scientist, data analyst)

Source: Made by author.

The organization has not defined data management principles, procedures, or standards. Therefore, given the scope of the article, it is proposed that policies for each of the information management elements previously described be defined. It is important to consider not only the lifecycle of transactional data but also of analytical data and each phase of analytics. For example, in the data quality policy, standardization procedures (validation and verification, updating, and maintenance) should be included, and they should be valid for transactional data and those created from preparation and modeling. In the previous section, some examples of policies related to specific phases of the analytics cycle are presented. However, it is essential to define:

- Which data should be used for descriptive, predictive, and prescriptive analytical models?
- Which analytical models should be implemented in each data set? Can black box models be used for advanced analytics, or should they be transparent?
- Which temporal range of data is acceptable to use?
- Is there a data quality threshold for using them in descriptive, predictive, and prescriptive analytics processes?

Activity / Role	CDG	ITC	BM	BOpL	DACL	BOpT	DE	DS	DC
Business Understanding									
Determine business objectives (context, goals, scope, success criteria).	I	I	V	R	R	A	E	E	E
Evaluate the context (resources, requirements, assumptions, and constraints), risks, costs, and benefits.	I	I	A	A	R	E	E	E	E
Frame the business need within a data project, considering the context, resources, and scope. I	I	V	V	R	V	V	E	E	E
Define the objectives and success criteria of the data project.	I	I	V	V	V	A	E	E	E
Build a project execution plan (evaluate tools, resources and techniques availability).	I	I	I	A/V	R/V	A	E	E	E
Data Preparation									
Clean the data, ensuring adherence to standards for handling duplicate data, outliers, and record errors.	I			I	R	V	A	E	V
Select the most relevant variables for model construction.	I			I	R	A	A	E	V
Enrich the data with new variables from other information sources or by combining existing ones.	I			I	R	A	A	E	V
Integrate different data sources to create a single dataset for modeling.	I				R		E	A	
Apply data reformatting techniques (standardization, normalization, encoding) for modeling.					R		A	E	V
Make the prepared dataset available in the appropriate location.	I	I			R		E		V

Accountability: Responsible (R), Executor (E), Informed (I), Technical Advisor (A), Validator (V). Roles: Chief of Data Governance (CDG), Information and Tech. Chief (ITC), Business Manager (BM), Business Operation Leader (BOpL), Data Analytics and Curatorship Leader (DACL), Business Operation Team (BOpT), Data Analytics and Curatorship Team (DACT); composed by: Data Engineer (DE), Data Scientist (DS) and Data Curator (DC).

Table 3. Role and accountability matrix. Describes how roles interact with responsibilities based on each phase's general tasks. Made by the author.

COMMUNICATION AND CULTURE

A large part of the success of DG programs is due to the organization's cultural adoption. This change must occur at all levels since data transits through every area and hierarchical level. An important team for this purpose is DACT, with its leader. As primary users of data, this team is called to materialize value through data and make it visible in the organization.

Therefore, the DACT must propose and develop communication, training and development plans; some suggestions are presented below:

Committed Leadership. The most effective way to involve a leader is by presenting results directly impacting business objectives. In this context, DACL must demonstrate the value derived from data and analytics processes to engage top organizational leadership as stakeholders of ADG and data culture. Furthermore, it is crucial that success stories resulting from data-based decisions are not only reviewed and evaluated but also communicated internally. This practice fosters a culture of learning from past successes and continuous improvement.

Communication. The organization must consistently communicate success stories, leadership commitment, and updates on data governance policies. This has to include sharing best practices and reminding everyone of the importance of adhering to the ADG policies in their day-to-day work.

Training and Development. The ideal is that every person that works in Comfama at least knows the importance of data. A course presented on entry level for everyone that shows their importance and value for the organization could be the first step to make aware everyone of data culture.

Enabling and promoting cultural change in the organization to achieve an understanding of the value of data must start from an understanding of the benefits that data management and governance bring to the organization, its employees, and its affiliates and their families.

CONCLUSION

The design and implementation of a sustainable and effective data governance program focusing on analytics in a CCF represent a key proposal in the current context, characterized by exponential data growth. Privacy and security are priorities, requiring efficient information management for data-driven decision-making.

This kind of initiative will allow the company to fulfill its purpose of positively impacting society's well-being. In an increasingly data-informed business environment, where accurate, data-driven decision-making is prioritized, the role of an ADG model becomes highly relevant. It provides usability, quality, and integrity of the primary material for analytics: data. This facilitates accurate decision-making, as the quality of inferences obtained from data processing highly depends on the quality of the data itself. Analytics provides valuable tools for decision-making. It can identify patterns, trends, and opportunities in data, allowing more effective resource allocation and personalized and efficient service offerings for people benefiting from the organization's services. By designing and implementing these programs in a CCF (organizations playing a crucial role in the social welfare and quality of life of workers and their families), using analytics and data for resource distribution is optimized and encouraged.

Additionally, it facilitates the precise identification of beneficiaries' needs, enabling the design of programs and services that positively impact society. The implementation of solid data governance responds to the need to adapt to an ever-evolving environment, where data becomes increasingly relevant and must be secured, accessed, and ensured of quality to use responsibly.

The case study allowed Comfama to evaluate the data governance that was being implemented and consider extending the governance to analytics data, which will allow the organization to ensure quality, security and responsibility on decision-making processes within the organization. This approach positively impacts the organization and the millions of Comfama affiliates, improving operational efficiency, strengthening decision-making, and contributing to improving people's quality of life.

Acknowledgments To my family, who had patience with me for about 30 years, to MP and my friends, who supported and encouraged me. To my editor, S.C., an international journalism award winner who endured a thousand versions. I thank N and T, great companions during long writing nights. And to the professor, who finally got me to do the thesis.

REFERENCES

Balanta Sinisterra, L. A. (2022). Modelo conceptual gobierno de datos para Caja de Compensación Familiar del Valle del Cauca (Master's thesis, Universidad Nacional Abierta y a Distancia, Cali). Maestría en Gestión de Tecnología de Información.

Barton, T., & Müller, C. (Eds.). (2023). Apply data science: Introduction, applications and projects (1a ed.). Springer Vieweg.

Cabezas Azuero, J. S. (2023). Tratamiento de datos personales y compliance en Colombia. Revista de la Facultad de Derecho y Ciencias Políticas, 53(138), 1–25. https://doi.org/10.18566/rfdcp.v53n138.a2

Comfama. (2020). Entidades y mecanismos de control que regulan a las Cajas de Compensación Familiar. Recuperado de https://www.comfama.com/conoce-comfama/entidadesmecanismos-control-cajas-compensacion-familiar/

Comfama. (2022). ¿Qué es y para qué sirve una caja de compensación? Recuperado de https://www.comfama.com/conoce-comfama/que-es-una-caja-de-compensacion/

DAMA International. (2009). The DAMA Guide to the Data Management Body of Knowledge (M. Mosley, M. Brackett, & S. Earley, Eds.; 1.a ed.). Technics Publications LLC.

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge (D. Henderson, S. Earley, & L. Sebastian-Coleman, Eds.; 2.a ed.). Technics Publications LLC.

De Caigny, A., Coussement, K., & De Bock, K. W. (2020). Leveraging fine-grained transaction data for customer life event predictions. Decision Support Systems, 130. https://doi.org/10.1016/j.dss.2019.113232

Eryurek, E., Gilad, U., Lakshmanan, V., Kibunguchy, A., Kibunguchy-Grant, A., & Ashdown, J. (2021). Data Governance: The Definitive Guide: People, Processes, and Tools to Operationalize Data Trustworthiness. O'Reilly Media, Incorporated.

Fadler, M., & Legner, C. (2021). Toward big data and analytics governance: Redefining structural governance mechanisms. Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS). https://doi.org/10.24251/HICSS.2021.691

Gartner, Inc. (2024). What is data analytics? Recuperado de https://www.gartner.com/en/topics/data-and-analytics

Grover, V., Chiang, R., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. Journal of Management Information Systems, 35(2), 388-423. https://doi.org/10.1080/07421222.2018.1451951

Ladley, J. (2019). Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program. Elsevier Science.

Lawrance, N., Petrides, G., & Guerry, M. A. (2021). Predicting employee absenteeism for cost effective interventions. Decision Support Systems, 147. https://doi.org/10.1016/j.dss.2021.113539

Fantini, F. (2023, 10 abril). Analytics for Marketers. Harvard Business Review. https://hbr.org/2023/05/analytics-for-marketersSuperintendencia de Subsidio Familiar. (2022). Preguntas y respuestas frecuentes. Recuperado de http://tinyurl.com/bde4kccd

Tobar Ruiz, P. A. (2021). Marco de referencia para un modelo de gobierno de datos: Caso de uso en una Caja de Compensación Familiar (Master's thesis, Escuela Colombiana de Ingeniería Julio Garavito, Decanatura de Ingeniería de Sistemas, Bogotá). Maestría en Gestión de Información.

ANNEX A

Complete roles, activity and accountabilities matrix based on data analytics lifecycle. Can be used as example to develop a ADG program for CCFs in Colombia, buy must be adjusted to the identified cycle and processes in the organization. Made by author.

Activities / Roles	Chief of Data Governance (CDG)	Business Manager (BM)	Business Op. Leader (BOpL)	Analytics and Curatorship Leader (DACL)	Business Op. Team (BOpT)	Analytics and Curatorship Team (DACT)		
						Data Engineer (DE)	Data Scientist (DS)	Data Curator (DC)
Accountability: Responsible (R), Executor (E), Informed (I), Technical Advisor (A), Validator (V)								
Strategic Alignment								
Identify the prioritized initiatives critical to the organization's purpose and analyze how data and analytics are essential to delivering value.								
Redefine the success factors, scope and drivers of the DG program.								
Establish the expected deliverables for each phase of the GDA program implementation.								
Define management and control indicators for the GDA program.								
Communicate results of key initiatives and the generated value by the GDA.								

Activities / Roles	Chief of Data Governance (CDG)	Business Manager (BM)	Business Op. Leader (BOpL)	Analytics and Curatorship Leader (DACL)	Business Op. Team (BOpT)	Analytics and Curatorship Team (DACT)		
						Data Engineer (DE)	Data Scientist (DS)	Data Curator (DC)
Accountability: Responsible (R), Executor (E), Informed (I), Technical Advisor (A), Validator (V)								
Business Understanding								
Determine business objectives (context, objectives, scope, success criteria).								
Assess context (inventory of resources, requirements, assumptions, and constraints), risks, costs, and benefits.								
Frame the business need in a data project, considering context, resources, and scope.								
Define data project objectives and success criteria.								
Build a project execution plan (evaluate tools and techniques).								

Activities / Roles	Chief of Data Governance (CDG)	Business Manager (BM)	Business Op. Leader (BOpL)	Analytics and Curatorship Leader (DACL)	Business Op. Team (BOpT)	Analytics and Curatorship Team (DACT)		
						Data Engineer (DE)	Data Scientist (DS)	Data Curator (DC)
Accountability: Responsible (R), Executor (E), Informed (I), Technical Advisor (A), Validator (V)								
Data Understanding								
Define the data to be used based on the need and verify its availability.								
Collect the defined data (keeping in mind that privacy must be taken into account, how this data can be used, and who can access it).								
Understand what the data contains from the metadata, the catalog, and the insights of business experts, ensuring that the interpretation of the data is accurate and as it should be.								
Perform an exploratory analysis of the data to understand in depth its content and the relationship between the different variables.								
Verify the quality of the data, including checking for duplicity of records, atypical data, and possible errors in the records, to ensure the reliability of the analysis and results.								

Activities / Roles	Chief of Data Governance (CDG)	Business Manager (BM)	Business Op. Leader (BOpL)	Analytics and Curatorship Leader (DACL)	Business Op. Team (BOpT)	Analytics and Curatorship Team (DACT)		
						Data Engineer (DE)	Data Scientist (DS)	Data Curator (DC)
Accountability: Responsible (R), Executor (E), Informed (I), Technical Advisor (A), Validator (V)								
Data Preparation								
Perform data cleaning, verifying the standards for handling duplicate and outlier data and considering recording errors.								
Select the most relevant variables for the construction of the model.								
Enrich the data with new variables from other sources of information or a combination of existing ones.								
Integrate the different data sources to have a single data set for modeling.								
Apply data reformatting techniques (standardization, normalization, codification) according to modeling needs.								
Make the prepared dataset available in the appropriate place.								

Activities / Roles	Chief of Data Governance (CDG)	Business Manager (BM)	Business Op. Leader (BOpL)	Analytics and Curatorship Leader (DACL)	Business Op. Team (BOpT)	Analytics and Curatorship Team (DACT)		
						Data Engineer (DE)	Data Scientist (DS)	Data Curator (DC)
Accountability: Responsible (R), Executor (E), Informed (I), Technical Advisor (A), Validator (V)								
Modeling								
Define the most appropriate modeling technique for the stated objective under the standards defined for the data used.								
Develop multiple test models.								
Evaluate the different models developed from fit and validation tests.								
Refine the model parameters to obtain better results.								
Model Evaluation								
Evaluate the models in terms of success criteria and tests of fit and efficiency.								
Analyze the results of the different models to find relevant results and present possible actions for decision-makers.								
Approve the most suitable model that meets the project's success criteria and responds to the business needs. It must also comply with quality, ethics, and compliance standards.								

Activities / Roles	Chief of Data Governance (CDG)	Business Manager (BM)	Business Op. Leader (BOpL)	Analytics and Curatorship Leader (DACL)	Business Op. Team (BOpT)	Analytics and Curatorship Team (DACT)		
						Data Engineer (DE)	Data Scientist (DS)	Data Curator (DC)
Accountability: Responsible (R), Executor (E), Informed (I), Technical Advisor (A), Validator (V)								
Deployment								
Plan and execute the deployment and publication of data with data architects and engineers, following standards and policies.								
Build a plan and execute the monitoring and maintenance of the model and deployment.								
Perform reporting and socialization of the project and results.								
Properly document the model, the data created, and the result obtained. The metadata for the model and the data must be filled out completely.								
Review project experience and documentation								