Marco de Gobierno de Datos para la Plataforma Digital Nacional

Versión 1.0 - DRAFT

Resumen Ejecutivo

Este documento establece un marco integral de gobierno de datos para la Plataforma Digital Nacional, enfocándose en los sistemas S1 (Declaraciones), S2 (Servidores públicos en contrataciones), S3 (Servidores públicos sancionados) y S6 (Contrataciones públicas). El marco propuesto se basa en las mejores prácticas de DAMA International y define una arquitectura MDM (Master Data Management) que incluye un repositorio central, reglas de calidad de datos, y un framework de detección de corrupción.

La arquitectura propuesta aborda los desafíos actuales de inconsistencia y duplicación de datos a través de un modelo de capas que incluye una entidad central PERSONA_MAESTRA, dimensiones estandarizadas y tablas de hechos. El sistema implementa validaciones específicas anticorrupción y establece métricas de calidad para asegurar la integridad, precisión y oportunidad de los datos, con especial énfasis en la detección de patrones sospechosos y el análisis de redes de relaciones entre servidores públicos y contratistas.

0. Contexto Relevante de DAMA y Gobierno de Datos

DAMA (Data Management Association) International es la autoridad global líder en el establecimiento de estándares y mejores prácticas para el gobierno y gestión de datos. A través de su marco DAMA-DMBOK (Data Management Body of Knowledge), proporciona una guía estructurada para desarrollar e implementar prácticas efectivas de gestión de datos en organizaciones.

El gobierno de datos, bajo el marco DAMA, abarca múltiples componentes esenciales. Tras un análisis de los datos encontrados en los Sistemas, identificamos que los siguientes componentes requieren una atención prioritaria:

- **Políticas y Estándares**: Establece las reglas fundamentales para la gestión de datos, incluyendo su creación, almacenamiento, uso y eliminación.
- **Calidad de Datos**: Define los criterios y métricas para asegurar la precisión, completitud y confiabilidad de los datos.
- **Seguridad y Privacidad**: Determina los protocolos para proteger la información sensible y cumplir con regulaciones.

- **Gestión de Datos Maestros**: Establece las prácticas para mantener una única fuente de verdad para datos críticos.
- **Arquitectura de Datos**: Define la estructura y los modelos para organizar y almacenar datos eficientemente.

En el contexto específico de plataformas gubernamentales anticorrupción, el gobierno de datos adquiere una dimensión crítica por varias razones:

1. **Integridad y Transparencia**: La información debe ser precisa y verificable para sustentar investigaciones y decisiones.
2. **Cumplimiento Regulatorio**: Debe adherirse a leyes y regulaciones específicas sobre transparencia y protección de datos.
3. **Interoperabilidad**: Los datos deben poder compartirse e integrarse efectivamente entre diferentes sistemas y dependencias.
4. **Trazabilidad**: Cada modificación o acceso a los datos debe ser registrado y auditable.

La implementación de un marco de gobierno de datos basado en DAMA proporciona:

- Una estructura organizacional clara para la gestión de datos
- Procesos estandarizados para el manejo de información
- Métricas y controles para evaluar la efectividad de la gestión
- Mecanismos para la mejora continua de la calidad de datos
- Herramientas para la detección y prevención de irregularidades

La adopción del marco DAMA para la Plataforma Digital Nacional es particularmente relevante debido a varios factores críticos:

1. **Complejidad de los Sistemas**: La plataforma integra múltiples sistemas (S1, S2, S3, S6) con estructuras de datos complejas y diversas que requieren una gestión estandarizada y robusta.
2. **Sensibilidad de la Información**: Los datos manejados incluyen información personal, financiera y legal de funcionarios públicos que requiere los más altos estándares de gestión y seguridad.
3. **Necesidad de Detección de Patrones**: La identificación de posibles casos de corrupción requiere un análisis sofisticado de datos que solo es posible con información bien gobernada y de alta calidad.
4. **Requisitos de Transparencia**: Como herramienta anticorrupción, la plataforma debe mantener los más altos estándares de transparencia y trazabilidad en la gestión de sus datos.
5. **Escalabilidad**: El marco DAMA proporciona prácticas probadas que pueden escalar conforme la plataforma crece en alcance y complejidad.

Este contexto es fundamental para entender la importancia y el alcance del marco de gobierno de datos propuesto para la Plataforma Digital Nacional, ya que establece las bases para una

gestión efectiva y transparente de la información en la lucha contra la corrupción, apoyándose en estándares internacionalmente reconocidos y probados.

1. Análisis de Fuentes de Verdad Actuales

1.1 Diccionarios y Especificaciones Existentes

Sistema S1 - Declaraciones

- Estructura jerárquica compleja con múltiples niveles de anidación
- Datos personales y patrimoniales con requisitos específicos de privacidad
- Campos estructurados para información financiera y patrimonial
- Validaciones complejas entre diferentes secciones

Sistema S2 - Servidores en Contrataciones

- Estructura más simple y lineal
- Enfoque en datos de identificación y roles
- Referencias a catálogos estandarizados
- Validaciones temporales y de consistencia de roles

Sistema S3 - Servidores Sancionados

- Estructura orientada a procesos y sanciones
- Campos con requisitos legales específicos
- Referencias a documentos y expedientes
- Validaciones de fechas y periodos

Sistema S6 - Contrataciones Públicas

- Estructura jerárquica enfocada en identificación y seguimiento de actores en contrataciones.
- Sistema unificado de identificación personal y empresarial con soporte multimodal.
- Manejo integral de personalidades jurídicas y denominaciones legales estandarizadas.
- Arquitectura de referencias cruzadas mediante identificadores principales y adicionales.
- Sistema de validación y trazabilidad de ubicaciones e información de contacto.

1.2 Diagnóstico del Estado Actual

1.2.1 Problemas de Independencia de Entidades

- Duplicación de información personal entre sistemas

- Falta de llaves únicas consistentes entre sistemas
- Ausencia de un identificador maestro para servidores públicos
- Información redundante y potencialmente inconsistente

1.2.2 Inconsistencias Descriptivas

- Variaciones en la forma de capturar nombres y datos personales
- Diferentes niveles de detalle para la misma información
- Falta de estandarización en catálogos de valores
- Inconsistencias en formatos de fechas y valores monetarios

1.2.3 Análisis de Integridad Referencial

- Los sistemas implementan validaciones internas pero no entre sistemas
- Falta de mecanismos para detectar referencias circulares
- Ausencia de validación cruzada de información
- Necesidad de fortalecer la trazabilidad de cambios

1.2.4 Ejemplo de Inconsistencia Potencial: Datos Financieros y Personales en Declaraciones Patrimoniales

Contexto:
Supongamos que el sistema original crudo registra datos de declaraciones patrimoniales anuales de funcionarios públicos. Al analizar los datos transformados en un esquema referencial, encontramos una inconsistencia relacionada con los ingresos netos declarados por un individuo a lo largo de varios años.

Caso Detectado: Declaraciones Patrimoniales Inconsistentes

1. **Descripción del Problema:**
    - En el sistema referencial, se asoció un único identificador (`persona_id`) a un funcionario, consolidando datos de diferentes años de declaraciones patrimoniales.
    - Al revisar la tabla de hechos (`fact_declaraciones`), se encontró que los ingresos netos anuales declarados por este individuo fluctuaban drásticamente sin una justificación lógica en las demás dimensiones.
2. **Datos Observados:**
    - **Año 2020:**
      Ingresos netos anuales: $200,000 MXN

La declaración incluye propiedades y bienes muebles por un valor consistente con los ingresos.

- **Año 2021:**
  Ingresos netos anuales: $50,000 MXN
  La declaración sigue reportando los mismos bienes y propiedades, pero no hay registro de ventas ni cambios en la tabla `dim_bienes`.
- **Año 2022:**
  Ingresos netos anuales: $1,200,000 MXN
  No se registra un aumento de salario ni se menciona una actividad adicional en `dim_empleos`.

3. **Diagnóstico:**
   - Se identificó que en las declaraciones de 2021, el campo `totalIngresosConclusionNetos` estaba ausente en los datos crudos, por lo que el sistema original asignó un valor predeterminado de $50,000 MXN, sin validación cruzada.
   - En la declaración de 2022, se introdujeron datos sin consistencia con los empleos o bienes registrados.
4. **Impacto del Problema:**
   - Esta inconsistencia podría levantar sospechas de corrupción o error en la captura de información. El sistema referencial permitió identificar fácilmente que los ingresos declarados no correspondían a los bienes ni a las fuentes de ingreso registradas.
5. **Resolución:**
   - Se notificó al equipo encargado de validar los datos, quienes corrigieron el error de 2021 y ajustaron el registro de 2022 para reflejar ingresos adicionales provenientes de una nueva actividad empresarial.

Beneficio del Sistema Referencial

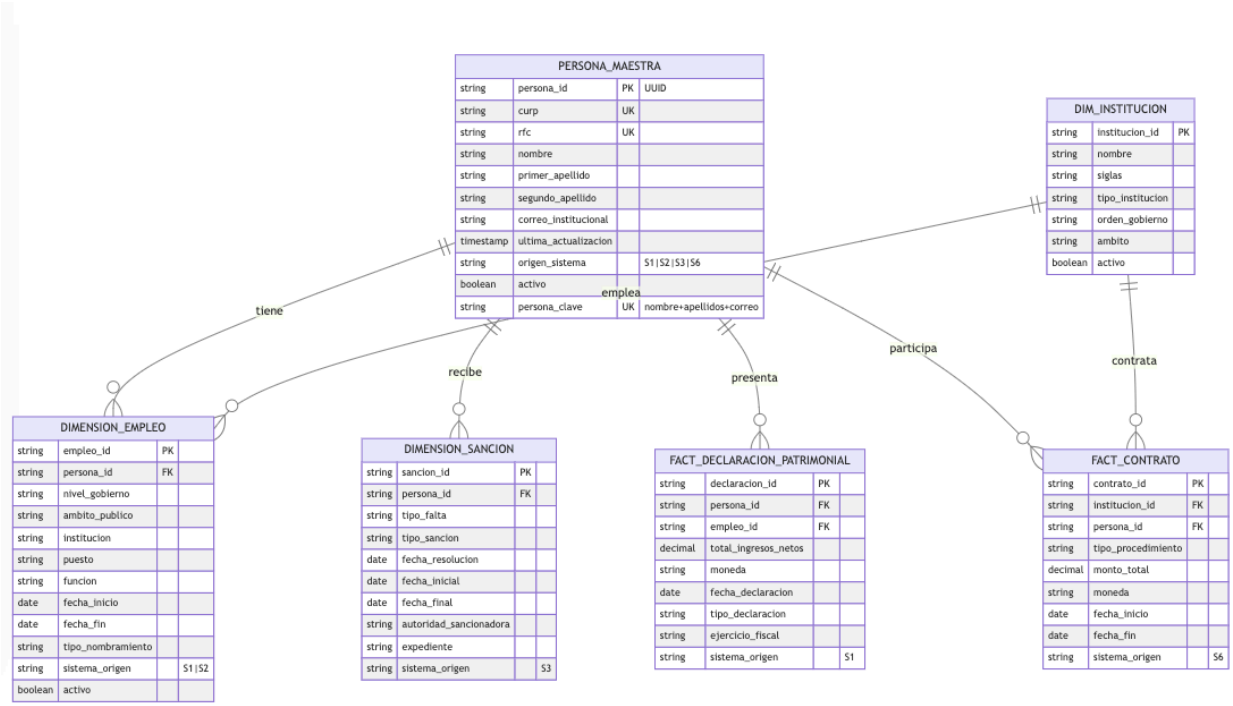
- **Integridad:** El sistema referencial consolidó datos de varios años bajo un identificador único, permitiendo comparar información de manera precisa.
- **Trazabilidad:** La relación entre las tablas de hechos y dimensiones ayudó a identificar las inconsistencias rápidamente.
- **Acción Correctiva:** Con los datos organizados y claros, fue posible corregir los errores y mejorar la calidad de la información para futuros análisis.

Este ejemplo muestra cómo el sistema referencial no solo detecta problemas de consistencia en los datos, sino que también habilita la trazabilidad y la capacidad de resolver problemas de forma efectiva.

2. Arquitectura MDM Propuesta

2.1 Modelo Conceptual

- Capa de Fuentes de Datos: Sistemas S1 (Declaraciones), S2 (Contrataciones), S3 (Sanciones) y S6 (Contrataciones públicas)
- Capa de Integración: Pipeline ETL con transformación a esquema referencial
- Capa de Datos Maestros: Repositorio central con modelos dimensionales
- Capa de Servicios: APIs y servicios de acceso estandarizado
- Capa de Gobierno: Políticas, reglas y procesos de calidad



El diagrama en formato mermaid puede ser consultado en este vínculo.

1. Capa Maestra
    - Entidad central PERSONA_MAESTRA que consolida la información de identidad de todos los sistemas
    - Identificadores únicos estandarizados (UUID, RFC)
    - Mecanismo de vinculación mediante persona_clave compuesta
    - Control de origen y vigencia de los datos

2. Capa de Dimensiones

- DIMENSION_EMPLEO: Historial laboral desde S1 y S2
- DIMENSION_SANCION: Registro de sanciones desde S3
- DIM_INSTITUCION: Catálogo unificado de instituciones
- Cada dimensión mantiene trazabilidad al sistema origen
- Control temporal mediante fechas de inicio y fin

3. Capa de Hechos
    - FACT_DECLARACION_PATRIMONIAL: Declaraciones desde S1
    - FACT_CONTRATO: Información de contratos desde S6
    - Vinculación con dimensiones mediante llaves foráneas
    - Métricas y montos con tipos de datos estandarizados

2.2 Componentes Clave

1. Repositorio MDM Central
    - Almacén centralizado de datos maestros
    - Gestión de identificadores únicos
    - Reglas de gobierno y calidad
    - Historial y auditoría de cambios
2. Sistema de Integración de Datos
    - Proceso ETL para normalización
    - Validación de datos en ingesta
    - Resolución de duplicados
    - Mapeo entre sistemas fuente
3. Motor de Calidad de Datos
    - Validación de reglas de negocio
    - Detección de anomalías
    - Estandarización de valores
    - Monitoreo de métricas
4. Servicios de Acceso
    - APIs RESTful estandarizadas
    - Control de acceso granular
    - Trazabilidad de consultas
    - Caché de datos frecuentes
5. Herramientas de Gobierno
    - Catálogo de datos unificado
    - Registro de linaje de datos
    - Métricas de calidad
    - Gestión de reglas de negocio

2.2.1 Repositorio MDM Central

- Almacén centralizado de datos maestros
- Gestión de identificadores únicos
- Reglas de gobierno y calidad
- Historial y auditoría de cambios

3. Matriz de Reglas de Calidad y Validación

3.1 Reglas de Calidad de Datos

3.1.1 Completitud

- Campos obligatorios deben estar presentes
- Validación de estructuras completas
- Verificación de referencias
- Integridad de vínculos documentos adjuntos

3.1.2 Precisión

- Formatos estandarizados
- Rangos válidos
- Consistencia de unidades
- Validación de cálculos

3.1.3 Consistencia

- Coherencia entre sistemas
- Validación temporal
- Consistencia de referencias
- Integridad de relaciones

3.1.4 Unicidad

- Establecer un identificador maestro (persona_id)
- Detectar nombres similares
- Normalizar campos para evitar variaciones

3.1.5 Oportunidad

- Asegurar la sincronización diaria entre las tablas referenciales

- Identificar registros cuya actualización sea mayor a un año y priorizar su revisión
- Generar alertas automáticas para datos crítico que no han sido ingresados

3.1.6 Integridad

- Registrar un historial de cambios en los datos para mantener una trazabilidad
- Validar que las remuneraciones estén documentadas y mantengan un rango salarial por actividad

3.2 Validaciones Específicas Anticorrupción

3.2.1 Validaciones de Identidad

- Unicidad de identificadores
- Consistencia de nombres
- Validación de documentos oficiales
- Verificación de roles y atribuciones

3.2.2 Validaciones Financieras

- Coherencia de montos
- Balances y totales
- Conversión de monedas
- Umbrales y límites

3.2.3 Validaciones Temporales

- Secuencia de eventos
- Períodos válidos
- Vigencia de documentos
- Histórico de cambios

3.3 Matriz Perfilamiento

Dimensión	Indicador	Métrica
Completitud	% Registros sin valores en blanco o nulos en datos críticos	100%
Precisión	% Registros de nombres con formato estándar	98%

Dimensión	Indicador	Métrica
	% Registros de remuneraciones dentro de los rangos razonables	100%
Consistencia	% Registros alineados con listado de instituciones	100%
Unicidad	%Registros duplicados corregidos	98%
Oportunidad	%Registros actualizados en los últimos 12 meses	98%
Integridad	%Registros con relaciones válidas entre actividad y remuneraciones	93%

4. Framework de Detección de Corrupción

El Framework de Detección de Corrupción propuesto integra un conjunto de herramientas y metodologías diseñadas para identificar y analizar patrones de comportamiento que podrían indicar posibles actos de corrupción en el sistema de contrataciones públicas. Este framework se basa en la arquitectura MDM implementada y aprovecha técnicas avanzadas de análisis de datos e inteligencia artificial para procesar la información consolidada de los sistemas S1, S2, S3 y S6 de la Plataforma Digital Nacional. A través de la combinación de indicadores de riesgo específicos, análisis de redes de relaciones y sistemas automatizados de detección, el framework está diseñado para proporcionar alertas tempranas y evidencia sustantiva que apoye la investigación y prevención de actos de corrupción, adaptándose continuamente a nuevos patrones y esquemas de comportamiento irregular.

4.1 Indicadores de Riesgo

4.1.1 Patrones Sospechosos

- Cambios patrimoniales injustificados
- Patrones inusuales en declaraciones
- Conflictos de interés potenciales
- Relaciones no declaradas

4.1.2 Análisis de Red

- Vínculos entre servidores públicos
- Relaciones con contratistas
- Patrones de contratación
- Conexiones indirectas

4.2 Sistemas de Detección Propuestos

4.2.1 Sistemas Basados en Análisis de Texto y Documentos

1. **Detección de Inconsistencias en Declaraciones**

- *Tipo de Fraude*: Utilización de información falsa, cohecho
- *Sistema Propuesto*: Análisis semántico con NLP y detección de anomalías basado en DBSCAN/TF-IDF (como se muestra en el código ejemplo)
- *Datos MDM*: FACT_DECLARACION_PATRIMONIAL, DIMENSION_EMPLEO
- *Capacidades*:
    - Identificación de patrones anómalos en descripciones de funciones y cargos
    - Detección de información inconsistente o incompleta
    - Validación cruzada de información curricular

2. **Análisis de Documentación en Contrataciones**

- *Tipo de Fraude*: Colusión, utilización de información falsa
- *Sistema Propuesto*: Procesamiento de Lenguaje Natural (NLP) con modelos transformer
- *Datos MDM*: FACT_CONTRATO, documentos adjuntos
- *Capacidades*:
    - Detección de similitudes sospechosas en propuestas técnicas
    - Identificación de patrones en documentación presentada
    - Análisis de consistencia en términos y condiciones

4.2.2 Sistemas de Detección de Patrones Financieros

1. **Monitor de Variaciones Patrimoniales**
    - *Tipo de Fraude*: Cohecho, desvío de recursos públicos
    - *Sistema Propuesto*: Modelo de series temporales con LSTM/Prophet
    - *Datos MDM*: FACT_DECLARACION_PATRIMONIAL histórico
    - *Capacidades*:
        - Detección de incrementos patrimoniales no justificados
        - Análisis de tendencias en ingresos y activos
        - Identificación de patrones estacionales anómalos
2. **Análisis de Precios en Contrataciones**
    - *Tipo de Fraude*: Soborno, ejercicio abusivo de funciones
    - *Sistema Propuesto*: Modelos de detección de anomalías basados en isolation forest
    - *Datos MDM*: FACT_CONTRATO

- *Capacidades*:
    - Identificación de precios fuera de rango de mercado
    - Detección de patrones de sobrecosto
    - Análisis de variaciones significativas en precios unitarios

4.2.3 Sistemas de Análisis de Redes

1. **Detector de Conflictos de Interés**
    - *Tipo de Fraude*: Actuación bajo conflicto de interés, tráfico de influencias
    - *Sistema Propuesto*: Análisis de grafos con Graph Neural Networks (GNN)
    - *Datos MDM*: PERSONA_MAESTRA, DIMENSION_EMPLEO, FACT_CONTRATO
    - *Capacidades*:
        - Identificación de conexiones entre funcionarios y contratistas
        - Detección de relaciones indirectas o encubiertas
        - Análisis de patrones de contratación recurrente
2. **Sistema de Análisis de Colusión**
    - *Tipo de Fraude*: Colusión, participación ilícita
    - *Sistema Propuesto*: Modelos de clustering jerárquico y análisis de comunidades
    - *Datos MDM*: FACT_CONTRATO, datos de empresas
    - *Capacidades*:
        - Identificación de patrones de participación conjunta
        - Detección de rotación de contratos
        - Análisis de comportamiento coordinado en licitaciones

4.2.4 Sistemas de Validación Temporal

1. **Monitor de Secuencias de Eventos**
    - *Tipo de Fraude*: Uso ilícito de atribuciones, contratación indebida
    - *Sistema Propuesto*: Modelos de secuencia temporal con attention mechanisms
    - *Datos MDM*: FACT_CONTRATO, DIMENSION_SANCION
    - *Capacidades*:
        - Detección de irregularidades en tiempos de procesos
        - Identificación de patrones temporales sospechosos
        - Análisis de secuencias de eventos anómalas
2. **Sistema de Alertas Tempranas**
    - *Tipo de Fraude*: Múltiples tipos
    - *Sistema Propuesto*: Ensemble de modelos con sistema de scoring
    - *Datos MDM*: Todas las fuentes
    - *Capacidades*:
        - Integración de señales de diferentes sistemas
        - Generación de alertas priorizadas
        - Dashboard de monitoreo en tiempo real